Università degli Studi di Napoli Federico II

Ph.D. Thesis

in

Information and Communication Technology for Health

Knowledge Graphs for Next-Generation Health Science Applications

Marco Postiglione

Tutor: Prof. Vincenzo Moscato

Coordinator: Prof. Daniele Riccio

XXXVI Ciclo

Scuola Politecnica e delle Scienze di Base
Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione

The important thing is
not to stop questioning

Albert Einstein

Knowledge graphs for next-generation health science applications

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Information and Communication Technology for Health
by

Marco Postiglione

October 2023

Approved as to style and content by

Prof. Vincenzo Moscato, Advisor

Università degli Studi di Napoli Federico II
Ph.D. Program in Information and Communication Technology for Health
XXXVI cycle - Chairman: Prof. Daniele Riccio

http://icth.dieti.unina.it

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Information and Communication Technology for Health is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, December 11, 2023

Marco Postiglione

Marco Postiglione

Abstract

The ongoing digitization of medical records, clinical charts, and health archives has markedly enhanced the accessibility of these critical documents, thereby ushering in a new era in the field of medicine. This data serves as a foundational cornerstone for the field of "precision medicine", whose primary objective is to elevate the standards of personalized diagnoses and therapies by harnessing the individualized attributes of patients, including but not limited to lifestyle factors, medical histories and genomic information. However, a notable challenge is that about 80% of healthcare data is unstructured, comprising textual elements like clinical notes and discharge summaries, and remains largely unexplored.

Traditional Natural Language Processing (NLP) algorithms, when applied to clinical scenarios, have largely depended on shallow matching techniques, template-based approaches, and non-contextualized word embeddings. These approaches exhibit limitations in capturing nuanced contextual semantics. Although there have been significant advancements in the broader NLP domain through language models able to effectively leverage contextual information, many of these general-purpose NLP algorithms face challenges when applied to specific clinical NLP tasks that necessitate specialized biomedical knowledge, especially in low-resource languages where there is a lack of annotated datasets.

This thesis delves into the multifaceted domain of few-shot learning techniques aimed at extracting information from clinical textual data. A pivotal focus is placed on data augmentation strategies and the amalgamation of multiple datasets into a unified model to enhance the learning efficacy. In this work, a novel representation of patients' medical histories is proposed through the introduction of Temporal Knowledge Graphs, which provide a structured framework for encapsulating chronological clinical information. Furthermore, a specialized model is developed, which utilizes recurrent units of Graph Convolutional Neural Networks (GCNs) to effectively harness the temporal dependencies within the data. This approach aims to anticipate potential health disorders a patient may encounter in the future, presenting a significant stride towards proactive healthcare management.

The validation of the proposed framework is carried out on two distinct datasets: the publicly accessible MIMIC-III database and a private dataset furnished by the Department of Advanced Biomedical Sciences at the University of Naples Federico II. The latter offers a unique lens into clinical narratives compiled in Italian, thereby broadening the evaluation spectrum and demonstrating the capability of the framework in handling multilingual clinical text. Through rigorous evaluation, this thesis underscores the potential of harmonizing clinical notes with structured temporal data representation in advancing predictive healthcare analytics.

Keywords: knowledge graphs, few-shot learning, data augmentation, natural language processing, electronic health records.

Sintesi in lingua Italiana

La crescente digitalizzazione dei referti, delle cartelle cliniche e dei fascicoli sanitari ha notevolmente migliorato l'accessibilità a questi documenti, aprendo la porta a nuove opportunità in campo medico. Questo rappresenta una prospettiva completamente nuova per la medicina, poiché la vasta quantità di dati disponibili in questo vasto archivio è ancora in gran parte inesplorata. Questa abbondanza di dati rappresenta un pilastro fondamentale per la "medicina di precisione", il cui obiettivo principale è migliorare le diagnosi e le terapie personalizzate sfruttando le caratteristiche individuali dei pazienti, tra cui lo stile di vita, la storia clinica e le informazioni genomiche, e così via. Tuttavia, una sfida rilevante è dato dal fatto che circa l'80% dei dati sanitari è in forma non strutturata, sotto forma di note cliniche e sintesi di dimissioni, e rimane in gran parte inesplorato.

Gli algoritmi tradizionali di Natural Language Processing (NLP), quando applicati a scenari clinici, hanno in gran parte dipeso da tecniche di "shallow matching", approcci basati su template predefiniti e word embeddings privi di informazioni di contesto. Questi approcci mostrano limiti nel catturare la semantica di elementi dipendenti dal contesto. Nonostante ci siano stati progressi significativi nel campo dell'NLP attraverso modelli linguistici in grado di sfruttare efficacemente le informazioni contestuali, molti di questi algoritmi perdono di efficacia quando applicati a specifici task in ambito clinico che richiedono conoscenze biomediche specializzate, specialmente in lingue con risorse limitate a causa della scarsità di dataset annotati.

Questa tesi si addentra nell'ampio dominio delle tecniche di "few-shot learning", mirate all'estrazione di informazioni dai testi clinici in scenari con scarsità di dati annotati. Particolare attenzione è posta sulle strategie di "data augmentation" e sull'unione di insiemi di dati multipli in un modello unificato per migliorare l'efficacia dell'apprendimento. Inoltre, in questa tesi viene proposta una rappresentazione delle storie mediche dei pazienti attraverso "Temporal Knowledge Graphs", che rappresentano una struttura dati in grado di racchiudere e integrare le informazioni cliniche dei pazienti tenendo anche conto dell'andamento dinamico dello stato di salute. La tesi propone un modello specializzato che impiega unità ricorrenti di Graph Convolutional Neural Networks (GCNs) per sfruttare le dipendenze temporali nei dati, con l'obiettivo di anticipare potenziali disturbi

di salute che un paziente potrebbe incontrare in futuro.

La validazione del framework proposto è stata effettuata su due set di dati distinti: il database pubblicamente accessibile MIMIC-III e un set di dati privato fornito dal Dipartimento di Scienze Biomediche Avanzate dell'Università di Napoli Federico II. Quest'ultimo offre una prospettiva sulle cartelle cliniche raccolte in italiano, ampliando così lo spettro di valutazione e dimostrando la capacità del framework di gestire testi clinici multilingue. Attraverso una rigorosa valutazione, questa tesi sottolinea il potenziale di armonizzare le note cliniche con una rappresentazione strutturata dei dati temporali nell'avanzamento delle analisi di salute predittive.

Parole chiave: knowledge graphs, few-shot learning, data augmentation, natural language processing, electronic health records.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Aims and scope

1.1 Motivation and Objectives

Scientific studies reveal that among the most popular medications in the United States, only 25% of patients get some benefit in the best-case scenario, and a mere 4% in the worst-case scenario [214]. This happens because most clinical research relies on randomized trials where groups of patients are administered medications, and their outcomes are subsequently examined. The reliability of these studies depends on how many people are in the study and their characteristics. However, the ever-increasing availability of digitalized medical records, clinical charts, and health archives is enhancing the accessibility to information on clinical practice, thereby ushering in a new era in the field of medicine.

Throughout the vast span of their evolution, humans have continually harnessed the power of natural languages for communication and information storage, employing the resources at hand, be it inscriptions on stone, handwritten manuscripts, or today's electronic storage media. These natural languages emerge as a highly effective and near-optimal medium for the conveyance of ideas, propelling the dissemination of knowledge, augmenting productivity, providing cognitive scaffolding, and easing decision-making processes among other vital functions.

In the contemporary era of digital transformation, natural languages are systematically transcribed into digital formats encompassing text and audio representations, exhibiting an unprecedented volume of linguistic data that undergoes computational scrutiny via specialized algorithms, specifically, Natural Language Processing (NLP). NLP, predicated on statistical and machine learning techniques, has ushered in a paradigm shift in our daily lives, catalyzed by its widespread applicability across domains such as search engines, chatbots, and recommendation systems. This technological revolution has redefined our interactions with information and communication.

This trend of digitization is having a significant impact within the clinical domain, wherein the realm of digital healthcare data accumulation has reached unparalleled proportions through the pervasive adoption of Electronic Health Records (EHRs). These repositories archive healthcare information pertaining to billions of patients, encompassing medical histories, diagnostic records, measurements, therapeutic interventions, pharmacological regimens, and so on. The information stored in EHRs assumes paramount significance for healthcare practitioners and scientific researchers, as it serves as the cornerstone for patient profile delineation, precision treatment strategy formulation, unraveling the intricacies of complex disorders such as cancer, pioneering advancements in life sciences research, and fundamentally, enhancing patient outcomes.

The availability of EHR data represents a significant step towards the realization of "precision medicine," a transformative approach to healthcare that tailors medical interventions and treatments to the individual characteristics of each patient. In precision medicine, healthcare decisions are not solely based on generalized guidelines or population averages but are instead informed by a comprehensive understanding of a patient's unique genetic makeup, medical history, lifestyle factors, and other relevant data. This approach aims to optimize treatment efficacy, minimize adverse effects, and enhance overall patient well-being by ensuring that medical interventions are precisely matched to the specific needs and attributes of each individual.

However, it is noteworthy that a substantial portion (up to 80%) of the data stored within EHRs exists in an unstructured format [110]. This predominantly entails textual data, necessitating considerable time from clinicians who must engage in manual reading and writing tasks as part of their routine practice. Nevertheless, this reliance on manual endeavors is inherently susceptible to error, thereby instigating consequential rami-

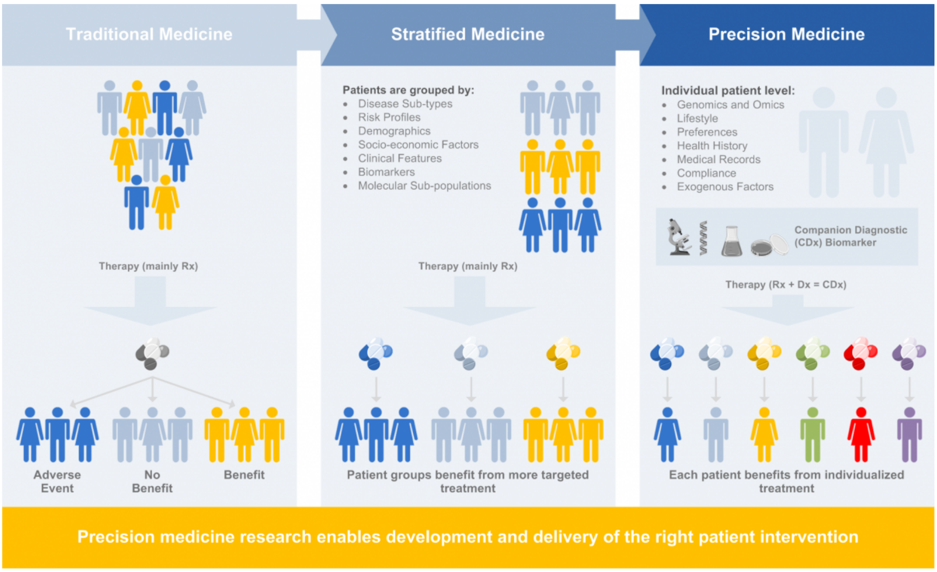


Figure 1.1. Transitioning from one-size-fits-all to precision medicine. Source: Drug Industry Bets Big On Precision Medicine: Five Trends Shaping Care Delivery, *Forbes*

fications, including elevated operational costs, diminished operational efficiency within healthcare service providers, and suboptimal patient outcomes, particularly in cases where individuals at elevated risk of severe diseases suffer from diagnostic inaccuracies or oversights.

Furthermore, the substantial prevalence of unstructured data within EHRs poses considerable challenges when analysing medical histories or using them for predictive downstream tasks (e.g. adverse event prediction). The initial step of extracting vital information from this unstructured wealth of data necessitates advanced NLP techniques, including Named Entity Recognition (NER) and entity linking. However, the deployment of these techniques crucially depends on the availability of annotated datasets, a resource that demands extensive efforts from healthcare professionals and domain experts. This requirement presents a significant hurdle, as it entails not only significant time and labor investments but also relies heavily on domain expertise to ensure high quality. This challenge is particularly pronounced in the context of low-resource languages, where publicly accessible annotated datasets are often scarce or entirely absent. Consequently, the development and application of NLP-based solutions for structured information extraction in EHRs encounter substantial obstacles, hindering progress in this vital field of healthcare data analytics.

1.2 Contributions

This thesis examines the extraction and analysis of data from unstructured clinical texts. Our contributions can be categorized and summarized as follows:

Few-shot learning In specialized sectors such as healthcare, particularly in underrepresented languages, there exists a notable scarcity of data. Addressing this challenge, this thesis offers a comprehensive examination of few-shot NER methods in Chapter 4. Innovative strategies rooted in data augmentation are presented in Chapters 6 and 7, while multi-task learning approaches are discussed in Chapters 8 and 9. Transformer-based models are also applied to Spanish and Italian clinical datasets in Chapters 5, 11, and 12, highlighting the disparities in data availability compared to English.

Data augmentation In addressing the challenge of training with limited data, this thesis highlights data augmentation as a pivotal solution. Perturbation techniques for enhancing Named Entity Recognition (NER) data are examined, with traditional methods often introducing inconsistencies in both syntax and semantic. Two novel strategies are proposed: first, as detailed in Chapter 6, a method is proposed to replace entity mentions using a vocabulary, assessing their congruence through cosine similarities between embeddings; second, as outlined in Chapter 7, a methodology akin to Policy-based Active Learning is employed to select optimal examples from augmented data.

Multi-task learning In this thesis, multi-task learning (MTL) is presented as a method that allows a single model to address multiple related tasks by harnessing their shared attributes. Chapter 8 introduces an innovative MTL approach for Named Entity Recognition (NER), deploying a unified model that integrates knowledge from entity-specific models. Chapter 9 explores the utilization of multi-task deep neural networks (MT-DNNs) in relation extraction, emphasizing their efficacy in limited-sample contexts and assessing potential performance trade-offs.

Future disorders prediction This thesis seeks to develop a system that predicts potential future medical issues for a patient, based on their digitally recorded medical history. In Chapter 10, a novel approach is introduced utilizing temporal knowledge graphs (TKGs). This approach takes into account both the dynamic data from patient histories and the fixed relationships found in medical ontologies. The methodology is tested on the public MIMIC-III dataset in Chapter 10. Further evaluation is conducted using a private dataset from the University of Naples Federico II's Department of Advanced Biomedical Sciences in Chapter 12, with results validated by domain experts.

Italian case study In Chapters 11 and 12 of this thesis, a comprehensive pipeline for Information Extraction and the analysis of medical histories from Italian clinical data is presented. Chapter 11 provides an in-depth look at the raw dataset from the *Campania Salute (CS)* network [229] and describes the approach utilized to extract mentions of medical concepts from clinical notes, subsequently linking them to the Unified Medical Language System (UMLS) ontology. This methodology involved annotation for Named Entity Recognition, Assertion Classification, and Entity Linking tasks. We plan to share these datasets in future research. Chapter 12 delves into the analysis of the derived medical histories, employing the TKG-based methodology outlined in Chapter 10. Validation of the results has been conducted with the aid of domain experts.

1.3 Thesis outline & Contributions

The structure of this thesis is delineated as follows:

Chapter 2 presents the context within this thesis is situated. We underscore the pivotal role of Artificial Intelligence in propelling the field of "precision medicine". This approach seeks to harness comprehensive patient data — ranging from genetics to lifestyle and health history — to tailor individualized treatments. Following this, we introduce Knowledge Graphs as a potent tool to facilitate this endeavor. We further elucidate the challenges posed by the sparse data availability in specialized domains like healthcare and in languages with limited resources, such as Italian.

Chapter 3 delves into the foundational background necessary to grasp the nuances of this thesis. It elucidates the concept of Knowledge Graphs and expounds upon the pivotal tasks associated with their construction and analysis, including Named Entity Recognition, Entity Linking, Relation Extraction, and Adverse Events Prediction.

Chapter 4 presents the few-shot learning challenge, with a particular emphasis of the Named Entity Recognition task. NER identifies predefined entity mentions within unstructured text and plays a crucial role in various subsequent tasks, notably the creation of Knowledge Graphs and Question Answering. The urgency for NER systems that can be trained with a minimal number of annotated examples is especially felt in areas where annotation demands time, expertise, and domain knowledge (like healthcare, finance, and legal) and in languages with limited resources.

Chapter 5 delves into the application of pre-trained transformer models for Named Entity Recognition and Entity Linking in Spanish data. The efficacy of these architectures is emphasized by showcasing the award-winning solution to the BioASQ Disease Text Mining (DisTEMIST) challenge. This challenge focuses on identifying disorder mentions within Spanish clinical notes and associating them with the SNOMED-CT ontology.

Chapter 6 addresses the enhancement of Named Entity Recognition through data augmentation. By harnessing similarity metrics, it introduces a method to produce augmented samples, ensuring the creation of sentences that remain plausible in real-world scenarios.

Chapter 7 delves into the enhancement of data augmentation systems, focusing on the challenge of minimizing noise introduced during the generation of augmented samples, such as syntactically or semantically incorrect sentences. Specifically, we will employ a deep Q-network to learn a heuristic for data selection, utilizing a policy-based active learning framework.

Chapter 8 tackles the intricacies of integrating multiple biomedical Named Entity Recognition datasets in a unique model. Particularly, many of the publicly available datasets specialize in annotating only one entity type, thus limiting their utility for the recognition of multiple entity types. To address this limitation, this chapter introduces TaughtNet, a knowledge distillation-based framework that learns a multi-entity model, i.e. the student, by leveraging multiple single-entity models, i.e. the teachers.

Chapter 9 addresses the scarcity of annotated datasets for biomedical Relation Extraction. Specifically, the chapter introduces a multi-task learning paradigm that uses various publicly accessible biomedical datasets to bolster relation extraction outcomes. This multi-task model amalgamates shared encoding layers across various datasets while dedicating specific layers for task-specific classifications, thereby producing specialized representations.

Chapter 10 explores the use of knowledge graphs to help predict potential future disorders in patients. By analyzing patients' medical histories, we extract key concepts from clinical notes made during their hospital visits. Our prediction framework combines this dynamic medical history data with static information from medical ontologies. We then process these graphs using recurrent Graph Convolutional Network (GCN) units to estimate the probability of future disorders.

Chapter 11 initiates our case study on an Italian dataset supplied by the Department of Biomedical Sciences at the University of Naples Federico II. This dataset is part of the CampaniaSalute network, comprising data from hypertensive clinics located in different community hospitals in the Naples metropolitan area. It includes data of 57,147 individuals, with records extending from 1980 onwards. In this chapter, we outline the data and describe our process of information extraction from clinical notes.

Chapter 12 analyzes the Italian dataset using the information extraction techniques from Chapter 11 to obtain medical histories. These histories constitute the training set for the methodology in Chapter 10, which aims to predict possible future disorders in patients. Results, validated with the help of domain experts, will be discussed in detail.

Chapter 2

Towards next-generation health science applications

2.1 The Artificial Intelligence Era

The term "Artificial Intelligence" (AI) refers to the ability of a computer system to mimic cognitive functions similar to those in the human brain. While it might seem a recent phenomenon, the quest for creating machines with human-like cognitive abilities began in the early 1940s. From its inception, AI has been a focal point of discussions in the scientific arena. Researchers have tirelessly worked on refining its theoretical underpinnings and methodologies, seeking to empower machines with intelligent cognition and actions. Meanwhile, philosophers ponder the implications of attributing intelligence to a man-made entity. Marvin Minsky, a pivotal figure in AI's early days alongside icons like Alan Turing and Frank Rosenblatt, succinctly described the core ambition of AI: *"to create machines that can perform tasks that, if done by humans, would be considered intelligent"* [165].

However, the AI landscape is riddled with challenges, one of the foremost being the absence of a universally agreed-upon definition of *intelligence*. The widely referenced benchmark today originates from Alan Turing's "imitation game," presented in his groundbreaking 1950 paper, "Computing Machine and Intelligence" [245]. Here, a machine's intelligence is assessed based on its ability to deceive a human evaluator through hidden interaction, such that its responses are indistinguishable from those of a human. While some AI models have displayed impressive capabilities, even surpassing humans in certain tasks [237], creating an AI that consistently deceives a human evaluator remains an elusive objective.

A notable characteristic of today's most advanced AI systems is their specialization. They excel in the tasks they are designed for but often falter when faced with diverse inputs or when expected to generalize across domains. Yet, the integration of AI into various sectors is rapidly expanding, from niche technology industries to everyday applications, indicating its significant global economic influence.

The surge in accessible data, coupled with the advent of cutting-edge techniques such as deep neural networks and transformers, has propelled AI's evolution. A testament to this progress is the emergence of Large Language Models. These models, capable of sifting through enormous textual data to discern intricate patterns, have ushered in revolutions in areas like translation and text generation. Their influence marks a pivotal moment in AI-driven language processing, promising a future with even more advanced human-machine interactions.

2.2 Precision Medicine

Precision medicine refers to the customization of medical treatment tailored to the individual characteristics of each patient. This approach considers factors like genetic makeup, environmental influences, and lifestyle to devise treatments, deviating from the conventional "one-size-fits-all" method. Central to precision medicine is the utilization of extensive data sources, most notably Electronic Health Records (EHRs), which encapsulate a patient's comprehensive medical history, from genetic information to lifestyle indicators [176].

EHRs, digital versions of a patient's medical history, have been revolutionary in streamlining patient information management. When processed efficiently, EHRs can offer profound insights into patient health, aiding in accurate diagnoses, treatment suggestions, and future health predictions. However, the heterogeneous and expansive nature of EHRs presents a challenge in deriving coherent and actionable insights.

This is where the power of Artificial Intelligence (AI) becomes indispensable. As detailed in the prior section on the AI Era, AI models, particularly advanced systems like deep learning, have an innate capacity to sift through vast datasets and discern intricate patterns [196]. In the context of healthcare, this ability translates to AI's potential to analyze multifaceted data from EHRs to predict potential health risks, even before they manifest.

In this thesis, the primary focus is on harnessing artificial intelligence to predict future adverse health events in patients using historical medical data stored in Electronic Health Records (EHRs). Through the application of AI algorithms, it is possible to analyze in-depth a patient's medical history, ranging from prior ailments and treatments to drug responses, to forecast potential health complications. Such predictive capabilities enable timely interventions, which may minimize or even avert the onset of certain health adversities [180].

Furthermore, integrating AI-driven insights derived from EHRs can elevate precision medicine to unparalleled heights, ensuring that medical care is not just reactive, but preemptive. The ensuing chapters will elucidate the methodologies, intricacies, and promise that AI brings to the realm of precision medicine, fostering a healthcare ecosystem that is both individualized and ahead of the curve [64].

2.3 Knowledge graphs: an enabling technology

Knowledge Graphs (KGs) are a transformative tool in the realm of data organization and representation. At their core, Knowledge Graphs are structured, graphical representations that connect entities (nodes) through relationships (edges), offering a semantic understanding of information. In essence, a KG captures knowledge in a format that is both machine-readable and inherently relational, providing a holistic view of intricate datasets.

Introduced by Google to enhance its search capabilities, Knowledge Graphs have found applications across various domains, with healthcare emerging as a particularly promising field [224]. In healthcare, KGs can be invaluable in providing a comprehensive, interconnected view of patient information, ensuring that data from disparate sources converges to offer a more complete understanding of patient health.

The task of leveraging Artificial Intelligence for predicting adverse health events can be augmented through the strategic use of Knowledge Graphs. By extracting vital concepts such as disorders, treatments, and tests from clinical notes, we can construct temporal knowledge graphs that systematically organize patients' medical histories. These temporal KGs serve as dynamic repositories, mapping the evolution of a patient's health status and treatments over time. Such an organization not only offers an unparalleled view of a patient's health journey but also serves as a structured input for advanced AI models.

2.4 Few-shot conditions: the challenge of Italian clinical notes

In the endeavor to construct robust knowledge graphs in the healthcare domain, processing clinical notes is imperative. These notes, replete with critical patient information, hold the key to extracting essential mentions and relationships that constitute the nodes and edges of a knowledge graph. Three fundamental tasks underpin this extraction process:

- *Biomedical Named Entity Recognition (BioNER).* This task involves pinpointing mentions of interest, such as disorders, treatments, and tests [75].
- *Relation Extraction*: Post BioNER, the relationships between the recognized entities are discerned, connecting disorders to treatments or diagnoses to specific tests, for instance [284].
- *Entity Linking*: This involves mapping the identified entities to unique identifiers in established biomedical ontologies, such as UMLS or SNOMED, ensuring that the extracted information is standardized and interoperable with other biomedical systems [44].

The biomedical domain, laden with domain-specific terminology and intricate relationships, inherently makes these tasks challenging [99]. However, the advent of natural language understanding methodologies, particularly the attention mechanism and transformer architectures, has ushered in significant advancements [249].

Yet, this thesis confronts an additional layer of complexity. The clinical notes under study are compiled in Italian, a language that, while rich and expansive, lacks extensive resources in the context of biomedical processing [191]. The dearth of publicly available datasets and the absence of pre-trained transformer architectures tailored for Italian biomedical data compound the challenge.

In light of the identified low-resource limitation, the focus transitions to the domain of few-shot learning [227]. This paradigm, which entails the training of models on a small dataset, presents a viable resolution under constrained resource scenarios. Given the limited availability of annotated Italian clinical notes, few-shot learning appears conducive for executing the Information Extraction tasks essential for knowledge graph development [26]. Through the adoption of this approach, the objective is to achieve entity extraction, relation, and linking at a level of precision comparable to models nurtured on larger datasets, thereby introducing a novel methodology for healthcare knowledge graph construction in low-resource settings.

2.5 Ethics and Privacy

The task of extracting valuable information from EHRs through Named Entity Recognition, Relation Extraction and Entity Linking, and the prediction of future adverse events, bring forth a plethora of ethical and privacy considerations. These concerns are augmented when viewed through the lens of regulatory frameworks such as the General Data Protection Regulation (GDPR) and the impending Artificial Intelligence Act (AI Act) by the European Commission.

Data privacy and consent GDPR mandates that personal data processing be done lawfully, fairly, and transparently, ensuring individuals' rights to data privacy. It stipulates obtaining explicit consent for data processing, where Article 4(11) defines consent as "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her". In the realm of EHR, obtaining informed consent is crucial before any data processing task, ensuring individuals are aware of how their data will be used, especially in identifying named entities or linking entities which could potentially re-identify anonymized data.

Data minimization and purpose limitation The GDPR's principles of data minimization (Article 5(1)(c)) and purpose limitation (Article 5(1)(b)) emphasize collecting data only for explicit and legitimate purposes and restricting further processing outside the initially stated purposes. Specifically, Article 5(1)(c) of the GDPR mandates that personal data should be "adequate, relevant and limited to what is necessary" concerning the purposes for which they are processed. This implies that companies should only gather the minimal amount of personal data needed to provide their services and should not collect more data than required for the data processing objective. On the other hand, Article 5(1)(b) indicates that personal data should be collected for "specified, explicit and legitimate purposes" and should not be further processed in a manner that is incompatible with those initial purposes. The aspect of incompatibility is context-dependent and hinges on the expectations of the data subject. However, further processing of the data is permissible for purposes like public interest, scientific or historical research, or statistical analysis, especially in accordance with Article 89(1) of the GDPR which provides certain conditions under which such processing can occur. Moreover, the principle of purpose limitation is intricately tied to other GDPR principles such as lawfulness, fairness, and transparency in data processing. This principle underscores the importance of clarity and legitimacy in the purposes for which data is collected and processed, and it also dictates that any further processing should align with these originally specified purposes

In the context of EHR, it is pivotal to adhere to these principles to avoid over-collection and misuse of sensitive health data, particularly in prediction tasks which may tempt extending data usage beyond the original consent.

Anonymization and pseudonymization GDPR encourages the use of anonymization and pseudonymization as a means to safeguard individuals' data. Article 4(5) of the GDPR defines the term "pseudonymisation" as the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person. Employing these techniques is fundamental in EHR information extraction to mitigate risks associated with data privacy and consent.

Transparency and Accountability The upcoming AI Act emphasizes human-centric AI, which aligns with GDPR's principles of transparency and accountability. The Act aims to ensure that AI systems are safe, respect existing laws on fundamental rights, and that they are grounded in Union values and principles. Transparency in how AI models operate and make predictions, especially in adverse event predictions, and accountability in cases of erroneous predictions or misidentifications are essential to uphold ethical standards and public trust.

High-risk AI systems The AI Act proposes a risk-based approach to regulate AI systems, identifying "high-risk" AI systems that pose significant risks to individuals' health, safety, or fundamental rights. These systems would need to comply with mandatory requirements for trustworthy AI and follow conformity assessment procedures. Given the sensitive nature of health data, the AI applications in EHR could be categorized as high-risk, necessitating strict adherence to the regulatory requirements proposed in the AI Act to ensure the protection of individuals' privacy and rights.

Below are the key regulatory requirements as outlined in various provisions and discussions surrounding the AI Act:

- *Classification and Conformity Assessment.* AI systems intended for use as safety components or are themselves products, and fall within the scope of Union harmonisation legislation listed in Annex II of the AI Act, must undergo a third-party conformity assessment before being placed on the market or put into service.

- *Quality Data, Documentation, and Traceability.* High-risk AI systems must adhere to requirements regarding high-quality data, documentation, traceability to ensure accuracy and robustness, which are crucial for mitigating risks to fundamental rights and safety.

- *Transparency and Human Oversight.* The AI Act mandates transparency and human oversight for high-risk AI systems to ensure they can be easily understood and controlled by human operators.

- *Risk Management System.* Providers of high-risk AI systems are required to maintain an appropriate risk management system throughout the lifecycle of the AI system.

- *Data Governance, Robustness, Accuracy, and Security.* The AI Act specifies legal requirements in areas such as data governance, robustness, accuracy, and security for high-risk AI systems.

- *Additional Requirements.* Additional requirements are set out in Chapter 2, Title III of the AI Act and in points 4.3., 4.4., 4.5., 4.6§5 of Annex VII of the AI Act concerning the conformity assessment procedures that high-risk AI systems must comply with.

These regulatory requirements are aimed at ensuring that high-risk AI systems are developed, deployed, and maintained in a manner that safeguards individuals' privacy, health, safety, and fundamental rights. By adhering to these regulations, providers of high-risk AI systems aim to build trust and ensure compliance with the legal framework proposed in the AI Act.

Bias and discrimination The AI Act emphasis on addressing the opacity, complexity, and bias in AI systems is crucial in healthcare settings, where bias in named entity recognition or adverse event predictions could lead to disparate impacts or discriminatory practices. Strategies to identify, mitigate, and communicate biases in data and model predictions are fundamental to uphold ethical standards and ensure equitable healthcare service delivery. These considerations underscore the imperative of a meticulous approach to data privacy, ethical considerations, and regulatory compliance in conducting information extraction tasks on EHR, as non-compliance could not only result in legal repercussions but also erode public trust in healthcare AI applications.

2.6 Conclusion

This chapter laid a foundational understanding of the various aspects that converge to form the basis for the ensuing explorations in this thesis. Initially, an examination of the AI era was conducted, delving into the genesis and evolution of artificial intelligence. Emphasis was placed on its potential and the hurdles encountered on the path towards achieving human-like cognition. The discussion on precision medicine underscored the synergistic interaction between AI and healthcare, leading to a more personalized medical treatment paradigm, powered by the rich data contained in Electronic Health Records (EHRs).

The introduction to Knowledge Graphs (KGs) elucidates the transformative capability of these structures in organizing and representing data, establishing a foundation for their application in healthcare to provide a more structured and comprehensive view of patient information. The exploration of Italian clinical notes highlights the linguistic and resource challenges encountered in information extraction tasks, suggesting few-shot learning as a potential solution in low-resource scenarios.

Furthermore, the section on ethics and privacy brought to the fore the crucial considerations that underpin the deployment of AI in healthcare. It underscored the importance of adhering to existing and forthcoming regulatory frameworks to ensure the ethical handling of sensitive health data, especially in tasks that involve the extraction and analysis of patient information from EHRs.

In proceeding to the subsequent chapters, the focus transitions to the practical facets of this thesis, specifically on the construction and analysis of medical Knowledge Graphs. Building on the theoretical foundations outlined in this chapter, the discussion will extend to the methodologies, challenges, and insights derived from the practical application of AI and KGs in healthcare. The forthcoming discourse seeks to explore not only the technical advancements but also to intertwine ethical considerations throughout the process of constructing and analyzing medical Knowledge Graphs, advancing towards the overarching goal of nurturing a more informed, personalized, and ethically-grounded healthcare ecosystem.

Part II

Building and Analyzing Knowledge Graphs

Chapter 3

Background: definitions, tasks and metrics

The extraction and interpretation of pertinent information from vast and diverse healthcare texts have assumed paramount significance in enabling evidence-based medical research, informed clinical decision-making, and improved patient care. The field of healthcare has undergone transformative changes with the integration of advanced technologies, particularly in the realm of machine learning (ML) and natural language processing (NLP). A definition for ML is provided as follows:

> **Definition 3.1** — Machine Learning [168, 170]
> A computer program is said to learn from experience E with respect to some classes of task T and performance measure P if its performance can improve with E on T measured by P.

In this chapter, definitions for knowledge graphs are provided, and the foundational tasks of NLP and ML central to the thesis — Named Entity Recognition (NER), Relation Extraction (RE), Entity Linking (EL), and Adverse Events Prediction (AEP) — are explored.

3.1 Knowledge Graphs

A knowledge graph (KG) depicts a systematic arrangement of information, made up of entities, their interconnections, and detailed semantic explanations. These entities can be tangible items or theoretical notions, while their connections indicate how they relate to one another. The detailed explanations about the entities and their connections include specific types and attributes that have a clear significance. Commonly, these are represented using property or attributed graphs where both nodes and their connections possess certain characteristics or attributes. A widely accepted definition for Knowledge Graph is provided as follows:

> **Definition 3.2** — Knowledge Graph [253]
>
> A knowledge graph is a multi-relational graph composed of entities and relations which are regarded as nodes and different types of edges, respectively.

Knowledge can be thus expressed as a sequence of triples:

$$\langle head, relation, tail \rangle,$$

where *head* and *tail* are two nodes and *relation* represents their connection.

3.1.1 Representing time

In this thesis, knowledge graphs are employed to encapsulate the multifaceted information pertaining to patients and the dynamic trajectory of their medical histories. Thus, the temporal dimension is particularly important, offering a lens through which to elucidate the progression of diseases. For instance, by analyzing the time-stamped data nodes within the knowledge graph, one can discern patterns and trends in disease evolution, which could be instrumental in prognostic assessments and the tailoring of therapeutic strategies. The refined knowledge representation, named *Temporal Knowledge Graph (TKG)* encapsulates a chronologically-ordered series of quadruples, expressed as:

$$\langle head, relation, tail, timestamp \rangle$$

The incorporation of the *timestamp* dimension furnishes a temporal context to the interrelations among entities, thereby facilitating a nuanced understanding of their dynamic interplay over time.

A formal definition is provided as follows:

Definition 3.3 — Temporal Knowledge Graph

A Temporal Knowledge Graph (TKG) can be formalized as a sequence of KGs, i.e. $\mathbf{H}_T = \{\mathscr{G}_1, \mathscr{G}_2, \ldots, \mathscr{G}_T\}$, where T is the length of the KGs sequence and each KG $\mathscr{G}_t = \langle \mathbf{V}, \mathbf{R}, \mathbf{E}_t \rangle$ at timestamp t is a directed heterogeneous graph, $\mathbf{V}$, $\mathbf{R}$ and $\mathbf{E}_t$ being the sets of entities, relations and facts at timestamp t, respectively.

3.2 Named Entity Recognition

This task focuses on identifying and categorizing key entities within a text. In the healthcare sector, NER is invaluable for pinpointing entities like medical terms, anatomical structures, drug names, and various clinical metrics. Properly extracting these entities is crucial for building structured knowledge databases and enhancing efficient information access and knowledge application.

Based on Definition 3.1, NER can be identified as the task T and E is a corpus of sentences annotated with entity mentions, so that a performance measure P (usually *Precision*, *Recall* and/or $F1$) of the machine learning model can be improved. Formally, NER is defined as in Definition 3.4.

Definition 3.4 — Named Entity Recognition [131]

Given a sequence of tokens (i.e. a sentence) $s = [t_1, t_2, \ldots, t_N]$, NER outputs a list of tuples $[I_s, I_e, t]$ representing named entities mentioned in s. Here, $I_s \in [1, N]$ and $I_e \in [1, N]$ are the indexes of start and end characters of the named entity mention, while t is the entity type.

It is important to acknowledge that the preceding definition is constrained to continuous spans, which represent a frequent occurrence. Nevertheless, when applying NER to real-world tasks, some cases may arise where named entities exhibit overlapping characteristics or are positioned within discontinuous textual spans. These particular situations have been subject to comprehensive investigation across various studies [174, 251, 60].

Due to its unique attributes, NER is a *token-level* classification task, with models generating predictions for each token separately. This can be likened to the analogous task of image segmentation in computer vision, where classification occurs pixel by pixel rather than being based on the entire image. Consequently, numerous techniques and model structures developed to address few-shot tasks in text classification cannot be directly employed in the context of NER. This situation necessitates additional endeavors, extensions, and adaptations to make them suitable for NER challenges.

3.2.1 Annotation schemes

A plethora of annotation schemes has been employed across scholarly literature. Nevertheless, the endeavor of identifying the most suitable annotation scheme represents a multifaceted conundrum [111]. These schemes are detailed as follows:

- **IO**: Serving as the most rudimentary scheme for this task, it assigns each token within the dataset one of two discerning labels: an "inside" tag (I) denoting named entities, and an "outside" tag (O) associated with ordinary words. It is important to note that this scheme exhibits a limitation, as it inadequately captures consecutive entities of identical types.

- **IOB**: This scheme, also referred to as BIO, ascribes categorical tags to each word, indicating whether it denotes the "beginning" (B) of a recognized named entity, resides "inside" (I) such an entity, or resides "outside" (O) the scope of any known entities.

- **IOE**: Analogous to IOB, IOE functions in a near-identical manner, yet diverges by signaling the "end" of an entity (via the E tag) instead of its inception.

- **IOBES**: Serving as an alternative to the IOB schema, IOBES offers a heightened granularity of boundary-related information concerning named entities. In tandem with designating words as "beginning" (B), "inside" (I), "end" (E), and "outside" (O) relative to a named entity, this scheme employs an "S" label to typify single-token entities.

- **BI**: This scheme bears semblance to IOB in its approach to entity tagging. Notably, it augments its function by characterizing the initiation of non-entity words through the "B-O" tag, subsequently assigning "I-O" tags to the following words.

- **IE**: Functioning congruently with IOE, the IE scheme diverges by allocating the "E-O" tag to demarcate the culmination of non-entity words, while designating the remainder with "I-O" tags.
- **BIES**: In a manner reminiscent of IOBES, the BIES scheme encodes entities, while simultaneously applying analogous encoding principles to non-entity words. This involves employing "B-O" to demarcate the inception of non-entity words, utilizing "I-O" to signify the interior of non-entity words, and introducing "S-O" to label individual non-entity tokens positioned amidst two entities.

Findings show that the IO annotation scheme is usually able to achieve the highest performance [11]. However, its main limitation is the inability to recognize consecutive entities. The most commonly used scheme is IOB, as it guarantees a good trade-off between performance and annotation efforts.

3.2.2 Metrics

As a token-level classification task, the evaluation of NER systems employs metrics such as precision, recall, and F1-score at the token level, providing insights into their token-wise classification abilities. However, as the application of predicted named entities extends to downstream tasks, there arises a need to assess system performance at a more comprehensive named-entity level. According to the *Language-Independent Named Entity Recognition* task introduced at CoNLL-2003 [205],

> **Definition 3.5** — NER metrics
>
> Precision is the percentage of named entities found by the learning system that are correct. Recall is the percentage of named entities in the corpus found by the system. A named entity is correct only if it is an *exact match* of the corresponding entity in the data file

SemEval'13[1] introduced a comprehensive framework encompassing four

[1] https://www.aclweb.org/portal/content/semeval-2013-international-workshop-semantic-evaluation

distinct methodologies for quantifying precision, recall, and F1-score outcomes grounded in particular metrics. These methodologies are outlined as follows:

- **Strict**: This approach entails a stringent evaluation involving an exact alignment of boundary surfaces, ensuring both string correspondence and entity classification congruence.
- **Exact**: Within this context, a meticulous comparison is undertaken, specifically concerning the accurate delineation of boundary extents across surface strings, without consideration for entity categorization.
- **Partial**: This methodology involves a judicious assessment of partial boundary alignments along surface strings, irrespective of the entity classification, allowing for partial matches.
- **Type**: In this mode, an essential requirement pertains to a certain degree of overlap between the entity identified by the system and the corresponding gold-standard annotation.

Each of these evaluative methodologies operates distinctively, encompassing correct, incorrect, partial, missed, and extraneous aspects, thus offering a nuanced perspective on performance assessment.

3.3 Entity Linking

Within the domain of natural language processing, the task of *entity linking*, commonly denoted as named-entity linking (NEL), named-entity disambiguation (NED), named-entity recognition and disambiguation (NERD), or named-entity normalization (NEN) assumes significance as an endeavor to establish a distinct and unequivocal identity for entities (such as persons, locations, diseases, or treatments) referenced in text. For instance, consider the sentence *"The patient displayed clinical indicators consistent with diabetes"*. In this context, the primary aim is to accurately determine that "diabetes" pertains to the medical condition rather than being misconstrued as a broader concept or denoting an individual's name. It is crucial to acknowledge that entity linking deviates from named-entity recognition (NER), which merely identifies the presence of named entities within text, devoid of the capacity to ascertain the precise entity being alluded to.

3.3.1 Biomedical ontologies

Entity Linking in the biomedical domain can leverage a large set of ontologies. Biomedical ontologies are structured frameworks that provide a formalized representation of knowledge in the domain of healthcare and life sciences. These ontologies serve as structured vocabularies, capturing the relationships between different biomedical concepts, enabling standardized data sharing, interoperability, and facilitating advanced analysis. Several prominent biomedical ontologies have been developed to cater to different aspects of the healthcare domain. Some examples of biomedical ontologies are reported as follows:

- **UMLS (Unified Medical Language System)**: UMLS is a knowledge integration system developed by the National Library of Medicine (NLM). It integrates a variety of biomedical vocabularies, classifications, and standards, including SNOMED CT, MeSH (Medical Subject Headings), ICD-10 (International Classification of Diseases), and more. UMLS aims to create connections between disparate terminologies, allowing researchers, clinicians, and developers to access and work with a unified representation of biomedical knowledge. It includes tools for mapping between different terminologies and enhancing cross-terminology querying.

- **SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms)**: SNOMED CT is an extensive clinical terminology and ontology that encompasses a comprehensive set of medical concepts, covering a wide range of clinical domains. It provides a standardized way of representing clinical information, facilitating communication and data exchange between different healthcare systems. SNOMED CT goes beyond a simple list of terms by organizing concepts into a hierarchical structure and capturing various relationships between them. It plays a vital role in electronic health records (EHRs), clinical decision support systems, and health information exchange.

- **Gene Ontology (GO)**: Focuses on the molecular functions, biological processes, and cellular components of genes and gene products. It aids in the annotation and analysis of genomic and proteomic data.

- **Human Phenotype Ontology (HPO)**: Captures phenotypic abnormalities and provides a standardized vocabulary for describing and annotating human phenotypes. It is crucial for clinical genetics and disease diagnostics.

- **Chemical Entities of Biological Interest (ChEBI)**: Focuses on chemical compounds and their role in biological processes. It's essential for research in drug discovery and molecular biology.

- **Protein Ontology (PRO)**: Represents information about proteins, their functions, and roles in various biological processes. It aids in understanding the relationships between genes, proteins, and diseases.

- **Anatomical Ontologies (e.g., FMA, Uberon)**: Capture the hierarchical structure and relationships among anatomical entities, facilitating the annotation of anatomical information in research and clinical contexts.

3.4 Relation Extraction

Relation Extraction (RE) is a NLP task that involves identifying and classifying relationships between entities mentioned in text. Entities can be people, organizations, locations, or any other named entities. The relationships between these entities are represented as specific types, such as "is-a," "part-of," "located-in," "causes," "treats," etc. Relation Extraction aims to extract these relationships from unstructured text and represent them in a structured format, making it easier for machines to understand and analyze the information.

In the realm of biomedicine, the accumulation of scientific literature, clinical reports, and biological data has reached an unprecedented scale. Extracting meaningful relationships from this vast amount of information is a daunting task for human experts alone. This is where Relation Extraction plays a critical role. Here are several reasons why Relation Extraction is of paramount importance in the context of building and analyzing biomedical knowledge graphs:

- **Knowledge Integration**: Biomedical research generates a multitude of heterogeneous data sources, including scholarly articles, clinical records, and experimental results. By extracting relationships between entities mentioned in these sources, Relation Extraction contributes to integrating this diverse knowledge into a unified framework.

- **Hypothesis Generation**: Relation Extraction can uncover hidden connections between biological entities that might not be immediately obvious to researchers. These newly discovered relationships can lead to the formulation of novel research hypotheses and guide further experimental studies.

- **Drug Discovery and Development**: In the pharmaceutical domain, identifying interactions between drugs, genes, proteins, and diseases is critical for drug discovery and development. Relation Extraction assists in identifying potential drug-target interactions, side effects, and disease mechanisms.

- **Clinical Decision Support**: Extracting relationships from clinical texts enables the creation of a knowledge base that can aid healthcare professionals in making informed decisions. For instance, identifying associations between symptoms, diseases, and treatments can assist in diagnosing and treating patients effectively.

- **Biological Pathway Reconstruction**: Relation Extraction contributes to reconstructing complex biological pathways by identifying interactions between genes, proteins, and metabolites. This aids in understanding cellular processes and disease mechanisms at a molecular level.

- **Data Mining and Literature Analysis**: Relation Extraction supports efficient data mining and literature analysis by structuring unstructured text into a graph-like representation. This allows researchers to perform advanced queries and analyze the relationships between entities in a systematic manner.

- **Semantic Enrichment**: Incorporating extracted relationships into a biomedical knowledge graph enhances its semantic richness. This enables advanced semantic searches, automated reasoning, and the discovery of previously unreported connections.

Relation Extraction acts as a foundational step that transforms unstructured biomedical information into structured knowledge, facilitating advanced data analytics, hypothesis generation, and decision-making in various biomedical domains.

3.5 Adverse Events Prediction

Adverse Events Prediction (AEP) aims to anticipate and forecast potential adverse events associated with pharmaceuticals, medical treatments, or interventions. Adverse events encompass any harmful or unintended effects resulting from the use of medical products, and predicting these events involves leveraging available data to identify patterns and relationships that can aid in foreseeing such occurrences.

In the context of biomedical research and healthcare, adverse events prediction holds paramount significance due to its potential to revolutionize clinical decision-making. Biomedical knowledge graphs, which represent complex networks of biomedical entities (such as drugs, genes, diseases, and their relationships), provide an ideal foundation for integrating diverse data sources and enabling advanced predictive analytics.

In this thesis, we aim predict adverse events by leveraging temporal knowledge graph (TKG) data structures to model medical histories, as described Definition 3.3.

Hence, AEP can be defined as follows:

Definition 3.6 — Adverse events prediction

Given the patient $s \in \mathscr{V}$ and its medical history $\mathscr{M}_t$ (represented as a temporal knowledge graph), the input to our model is a query

> $\langle s, r, ?, t+1 \rangle$, i.e. we want to predict the next adverse event associated to s.

The integration of adverse events prediction into biomedical knowledge graphs contributes to the realization of personalized medicine. By considering an individual's unique characteristics, genetic features, and medical history, clinicians can make treatment decisions that maximize benefits while minimizing risks.

3.5.1 Metrics

While the literature in the TKGs field aims to test a model's ability to predict the occurrence of an event at a future timestamp, of which the actual occurrence is known, our ground truth is composed of a multiplicity of events, meaning that there is not a single disease that will be associated with the patient in the future, but a list of diseases that could potentially be associated with the patient. This required us to use a different evaluation protocol and set of metrics to test our models. To assess the effectiveness of our model, a set of metrics has been employed, including Mean Reciprocal Rank (MRR), True Positive rate at k (TP_rate@k), Hits at k (Hits@k), Mean Recall at k (MR@k), Mean Averaged Precision at k (MAP@k), where k denotes the top-k ranked predictions made by the model. Metrics are detailed in the following.

Mean Reciprocal Rank MRR measures the average of the reciprocal ranks of the correct results in a set of queries or predictions. The formula for MRR is:

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank}_i}, \tag{3.1}$$

where N is the number of correct concepts and rank_i denotes the ranked position of the i-th concept.

The utilization of MRR as an evaluation metric presents several advantages. Firstly, it is scale-independent, making it an appropriate metric for comparing the performance of different models or tasks, regardless of the number of items in the list, which can decrease as the medical history progresses. Additionally, MRR places a strong emphasis on the first correct result, which is particularly relevant in tasks such as information retrieval, where users are less likely to scroll through long lists of results. Furthermore, it takes into account the entire ranked list, as opposed to only the top k results, as seen in metrics such as precision and recall. Due to these advantages, we employ MRR as a method of selecting the optimal model among the results obtained at each epoch of training.

True Positive rate To evaluate the usability of our model in real-world scenarios, where it would serve as an assistant for risk prediction or diagnosis suggestion, we investigate how likely one of the top-k predictions is correct, i.e. it appears in future steps of the medical history. Thus, we consider a prediction as a *true positive* if at least one of the top-k scored disorders is correct. The TP rate is then defined as the ratio between the sum of true positives and the total number of test samples:

$$\text{TP\_rate@}k = \frac{1}{N}\sum_{i=1}^{N} 1(\mathbf{y}_i, \hat{\mathbf{y}}_i^k), \tag{3.2}$$

where N denotes the number of test samples, $\mathbf{y}_i$ is the list of correct future disorders, $\hat{\mathbf{y}}_i^k$ denotes the first k disorders scored by the model and $1(\mathbf{y}_i, \hat{\mathbf{y}}_i^k)$ equals to 1 if at least one element of $\hat{\mathbf{y}}_i^k$ is also in $\mathbf{y}_i$, 0 otherwise.

Note that the number of correct concepts decreases as we move forward with the medical history, thus implying a possible decline of performance computed with this metric.

Hits Despite physicians having the knowledge and expertise to distinguish useful predictions among all the others, we would like every disorder predicted by the model to be correct. Hence, we compute Hits@k as the percentage of correct top-k predictions, i.e. its averaged precision:

$$\text{Hits@}k = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{TP}_i^k}{k}, \tag{3.3}$$

where TP_i^k denotes the number of true positives for the i-th test sample obtained by considering the top-k predictions.

As with TP_rate@k, performance could decrease as we proceed forward with the patient's timeline.

Mean Recall While TP_rate@k and Hits@k give us an idea of the *precision* of the model, i.e. of its ability to correctly identify future disorders, we are interested also in its *recall*, i.e. its ability to find all the future disorders. Mean Recall is defined as the fraction of correct items found in the top-k predictions:

$$\mathrm{MR@}k = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathrm{TP}_i^k}{\mathrm{TP}_i^k + \mathrm{FN}_i^k}, \tag{3.4}$$

where FN_i^k denotes the number of false negatives for the i-th test sample obtained by considering the top-k predictions.

Note that while this metric theoretically lies in the $[0, 1]$ range, values are usually small in practice because the ground-truth list of correct concepts is usually longer than k, making it impossible to reach high scores.

Mean Averaged Precision Ideally, we would like our model to score relevant future disorders at the top positions. MAP takes into account both the precision and recall of the recommendations made by the model and rewards first-loaded relevant recommendations, making it more informative than all the other metrics which does not consider the order in the rankings of predictions. It is computed as follows:

$$\mathrm{MAP@}k = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP@}k, \tag{3.5}$$

where AP@k is the average of the precision at each recall level for a particular test sample:

$$\mathrm{AP@}k = \sum_{i=1}^{k} \mathrm{Hits@}k \cdot \mathrm{rel}(i), \tag{3.6}$$

where $\mathrm{rel}(i)$ is an indicator function which is 1 if the i-th item is relevant, 0 otherwise.

Chapter 4

Few-Shot Learning for Information Extraction

Despite the burgeoning growth in data volume, there are significant challenges in ensuring the confidentiality of sensitive information, especially in sectors such as healthcare. Additionally, the labor-intensive nature of curating high-quality datasets underscores the need for alternative machine learning training methodologies that do not rely on extensive datasets [177, 250, 147, 283].

In the evolving landscape of few-shot learning, there is a marked shift towards refining and tailoring existing few-shot techniques specifically for Named Entity Recognition (NER) tasks. This observation is substantiated by the surge in academic publications on the subject, as depicted in Figure 4.1. Li et al. [131] underscore the constraints of transfer learning, particularly stemming from linguistic variances and differences in annotated textual datasets. Such constraints often result in diminished efficacy when a model trained on one dataset is applied to diverse textual datasets. The burgeoning theoretical frameworks aimed at bolstering model generalization within the few-shot learning domain, in conjunction with the growing research initiatives, underscore the need for a rigorous evaluation of the prevailing state-of-the-art. Such an evaluation is imperative to glean insights into the inherent merits of each approach.

In this chapter, we delineate the contemporary advancements in Few-

[1] https://scholar.google.com/

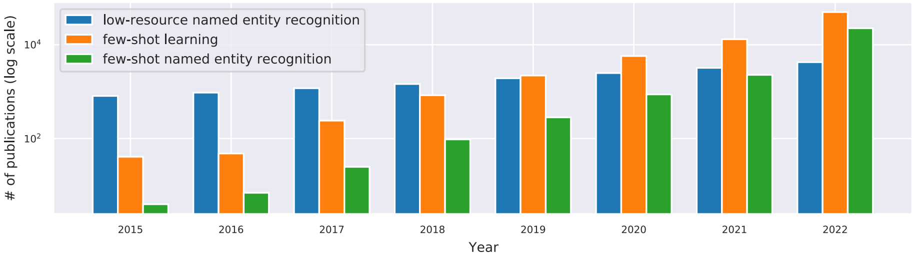


Figure 4.1. Evolution of the number of total publications whose title, abstract and/or keywords refer to few-shot learning during the last years. Data retrieved from Google Scholar[1](Feb 11th, 2023) by using the queries indicated in the legend.

Shot Named Entity Recognition (FS-NER). We commence by elucidating the concept of FS-NER and its relevance in the current research milieu. Subsequently, we introduce a classification schema, bifurcating the literature into two predominant categories: model-centric and data-centric. The model-centric category emphasizes the design and training of models optimized for few-shot scenarios, while the data-centric category delves into data manipulations aimed at enhancing or augmenting the available training datasets.

4.1 Few-shot NER: what, why, where and how

In this Section, we use some questions from the Kipling method[2] to provide an overview of FS-NER. Specifically, we provide answers for questions listed as follows:

- *What? (Problem definition)* — in Section 4.1.1 we formalize the few-shot NER problem in the context of machine learning and few-shot learning sub-field. Contextually, we will outline differences making few-shot NER a more challenging task worth of an in-depth analysis.
- *Why? (The need for FS-NER)* — we will provide a clear and concise example which shows the reasons behind the hype towards few-shot

[2] https://projectofhow.com/methods/the-kipling-method/

NER in Section 4.1.2.

- *Where? (Applications)* — in Section 4.1.3 we will describe real-world scenarios where FS-NER methodologies are needed.

- *How? (Taxonomy)* — we propose a taxonomy to categorize works on FS-NER, which will be described and analyzed in Section 4.2 and Section 4.3.

4.1.1 Problem definition

Few-Shot Named Entity Recognition (FS-NER) constitutes a distinctive subdomain within the broader landscape of Few-Shot Learning (FSL), which, in its entirety, resides within the overarching realm of machine learning (see Definition 3.1). The conventional trajectory of machine learning entails a reliance on substantial datasets infused with meticulously annotated information, commonly referred to as ground-truth, to achieve optimal training outcomes for models. In contrast, the specific pursuit of FSL is centered on the attainment of commendable performance even when confronted with a scarcity of annotated data instances. Formally, FSL can be defined as in Definition 4.1.

> **Definition 4.1** — Few-Shot Learning (FSL) [259]
>
> Few-Shot Learning (FSL) is a type of machine learning problems (specified by E, T and P), where E contains only a limited number of examples with supervised information for the target T.

In concrete terms, FSL aims at learning a classifier able to predict a label y for each input x (e.g. image classification [144], text classification [233]). In the light of Definitions 3.1 and 3.2, we can provide a clear and concise definition for FS-NER as described in Definition 4.2.

> **Definition 4.2** — Few-Shot Named Entity Recognition (FS-NER)
>
> Few-Shot Named Entity Recognition (FS-NER) is a sub-area of FSL where the machine learning problem specified by E, T and P is not only constrained by E containing a limited number of examples, but also by T being a NER task.

A pivotal distinction inherent in FS-NER lies in the needed association of multiple labels $\mathbf{y} = [y_1, y_2, \ldots, y_N]$ with each input sentence $\mathbf{s} = [t_1, t_2, \ldots, t_N]$, effectively corresponding to distinct tokens $t_i \in \mathbf{s}$. This intricate task configuration diverges markedly from conventional setups. Furthermore, the interdependence between two tokens, $t_i \in \mathbf{s}$ $(i \in [1, N])$ and $t_j \in \mathbf{s}$ $(j \in [1, N])$, extends beyond mere token-level adjacency to encompass semantic interconnectedness. This is instrumental in the construction of efficacious models, but this interdependence also implies challenges for the application of many FSL to FS-NER.

4.1.2 The need for FS-NER

Current strategies for Named Entity Recognition (NER) heavily rely on the underpinning frameworks of deep neural networks and Transformer architectures. These modern techniques have demonstrated remarkable prowess by obviating the need for extensive feature engineering, thereby attaining state-of-the-art benchmarks. Nonetheless, their practical applicability often encounters impediments due to their need for vast manually labeled training datasets. This reliance on abundant training data poses challenges when translating these achievements to real-world contexts. In FS-NER, the number of available training examples is small, thus impacting the reliability of the resulting NER model which, as a consequence, usually overfits data [259].

We present this phenomenon in Figure 4.2, wherein we illustrate the trends of F1 scores with respect to increasing NER training epochs. These trends are derived from validation sets associated with three extensively employed benchmark datasets: CoNLL-2003 [242], WNUT-17 [50], and WikiANN (en) [185].[3]

We can observe a discernible decrement in performance when the number of training instances, referred to as "shots," remains excessively small.

[3] To ensure reproducibility, we outline the specifics of this experimental setup. The models and datasets utilized are sourced from the HuggingFace repository [265]. For each dataset, a BERT base (cased) network [51] was fine-tuned over a span of ten epochs, with a *learning rate* set at 2e-5 and a *weight decay* of 0.01. All other hyperparameters retained their default configurations. The reported F1 scores are computed based on the validation sets. The training datasets for the few-shot experiments are composed by randomly sampling from the original training corpora.

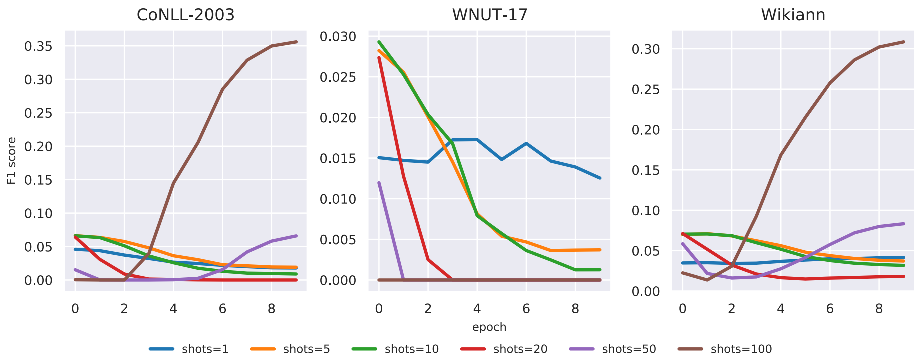


Figure 4.2. F1 score trends of few-shot NER models as the training epochs progress.

Specifically, for the CoNLL-2003 and WikiANN datasets, this decline becomes conspicuous when *shots* ≤ 20, whereas for the WNUT-17 dataset, it becomes evident when *shots* ≤ 100. This diminishing performance is indicative of the model's proclivity to overfit the limited training data it is provided. Furthermore, even as the F1 score exhibits an ascending trajectory with increasing training iterations, its eventual attainment remains lackluster due to the inherent limitation of the model in recalling all the instances of ground-truth entity mentions present within the test set.

These observations highlight the dual challenges faced: the mitigation of overfitting tendencies and the enhancement of NER model performance under the constraints of scarce labeled examples. It is from the convergence of these challenges that the exigency for Few-Shot Named Entity Recognition (FS-NER) methodologies becomes apparent.

4.1.3 Applications

In broad terms, the imperative for FS-NER is manifest across various application scenarios characterized by the scarcity of available training samples. This scarcity can be attributed to several factors:

- *Scarcity of resources* — a dataset to train our models NER models is not always available. For example, most of the current NLP research is focused on 20 out of the 7000 languages spread all over the world

[159], which are thus inevitably disadvantaged with the unavailability of data sources.

- *Difficult data sharing* — depending on the application domain, the possibility to share data and build training corpora may be a challenge. For instance, within the realm of healthcare, individuals, i.e., patients, often exhibit reluctance to disclose their sensitive information for research or commercial purposes. This hesitance persists despite the potentially transformative impact such data could exert.

- *Annotation costs* — the manual labelling process of NER data is expensive. This process entails the involvement of multiple annotators adhering to standardized protocols to mitigate annotation conflicts. Attaining high-quality training corpora necessitates a substantial temporal commitment to the annotation process. Additionally, specialized domain knowledge, such as healthcare or finance expertise, is often indispensable, thereby augmenting the overall costs incurred.

Hence, many real-world applications involve FS-NER. However, current state-of-the-art does not offer many successful use cases yet — methods are usually tested on benchmark datasets. Wang et al. [261] experiment their few-shot method on a private corpus of 1600 de-identified EHRs from cardiology, respiratory, neurology and gastroenterology deparments; Ni et al. [181] test their cross-lingual FS-NER approach on a custom multi-lingual dataset with over 50 entity types annotated to build cognitive question answering applications on top of the FS-NER system.

4.1.4 Taxonomy

In this section, we propose a taxonomy to categorize existing state-of-the-art FS-NER techniques. The first layer of the taxonomy is based on whether the methodology is focused on model architecture (*model-centric*) or data (*data-centric*), respectively. Input data sources and methodological flows change dependently on the technique, but we summarized high-level details in Figure 4.3.

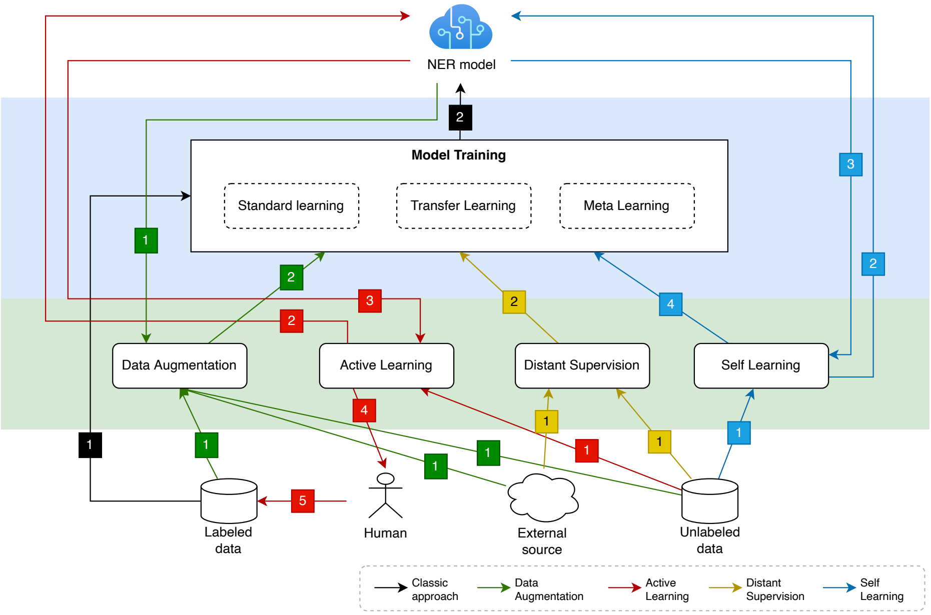


Figure 4.3. High-level methodological flows of FS-NER methods. The standard training approach (black arrows) receives labeled data as inputs; data augmentation techniques (green arrows) leverage labeled and unlabeled data, external sources and/or even the model itself to augment the size of the training corpus; active learning (red arrows) selects a subset of unlabeled data to be labeled by a human annotator by leveraging model predictions; distant supervision (yellow arrows) uses external sources and heuristics, while self-learning (blue arrows) uses model predictions to provide annotations to unlabeled data. Models are often trained relying on transfer learning or meta learning approaches. Numerical values are assigned to each data flow to indicate the sequence of operations.

Model-centric methods In this setting, model architectures are designed to make the most of the few available training samples. *Transfer learning* approaches leverage model weights learned in another domain or language. Differently from transfer learning, which leverage knowledge from the same task but a different domain, *Meta learning* aims to build models able to quickly adapt to new tasks without the need to be re-trained from scratch.

Data-centric methods The shift from model-centric to data-centric AI[4] is ongoing and increasingly widespread. This can be justified by the fact that the astonishing improvements brought by deep learning models on the state-of-the-art of several AI tasks has led the research community to find ever more better models, but now that a performance plateau has been reached, efforts are being made to deal with the other important aspect of AI systems: data. In the context of FS-NER, we identified four common methods to deal with the lack of data:

- *Data Augmentation* techniques leverage not only training samples but also external sources and unannotated data (when available) to increase the training corpus size
- *Active Learning* aims to select the most informative samples to be annotated from an unlabeled corpus in order to optimize the trade-off between performance and annotation costs.
- *Distant Supervision* consists in leveraging external data sources and heuristics to provide "weak" labels to data from an unlabeled corpus.
- *Self Learning* approaches use the model itself to provide a label to data from an unlabeled corpus.

In the following Sections we are describing in details the discussed methodologies.

4.2 Model-centric methods

In this section, we review model-centric FS-NER methods by separating them into two sections. Hence, we line up methods as a story and

[4] https://datacentricai.org

summarize their key characteristics and discuss similarities and differences under each category as well as limitations that have not been addressed yet.

4.2.1 Transfer Learning

In all the fields of Machine Learning, *Transfer Learning* is the standard approach to deal with the lack of data. The knowledge — i.e., their learned parameters — of models trained on huge datasets is "adapted" with new training iterations to make it possible for the model to perform well with a target domain where there is a lack of resources. Current Transfer Learning methods for NER are mainly based on deep neural networks and Transformer architectures and usually leverage *feature representation transfer* [188], which makes the model learn to map inputs from different domains in a close feature space, and *parameter transfer* [273], which makes the target model parameters close to those of the source model. We divide transfer learning approaches for FS-NER in three categories: *cross-domain*, *cross-lingual*, *fine-tuning*. In the following, we describe them in detail.

Cross-domain transfer In cross-domain Transfer Learning [89, 261], we aim to transfer the knowledge from a *source* specialty (e.g. Electronic Health Records, a.k.a. EHRs, from the department of cardiology) to a *target* specialty (e.g. EHRs from the department of orthopaedics). Figure 4.4 presents an example of inputs that can be used to train a cross-domain transfer learning system. The goal of this system is to enhance performance in a distinct target domain, which may feature a dissimilar set of entity types. The corresponding output is also depicted in the figure.

Cross-language transfer A high number of FS-NER works focus on leveraging cross-lingual information to improve the model performance [47, 181, 41, 267, 194]. Based on the availability of data, many transfer learning scenarios are possible, as shown in Figure 4.5. Most of the approaches rely on a projection-based transfer scheme [274, 104, 252, 181, 208]: one side of bitext is annotated with a tagger for a high-resource language and then the annotation is projected over the bilingual alignments obtained through unsupervised learning [183]. Projected annotations are then used

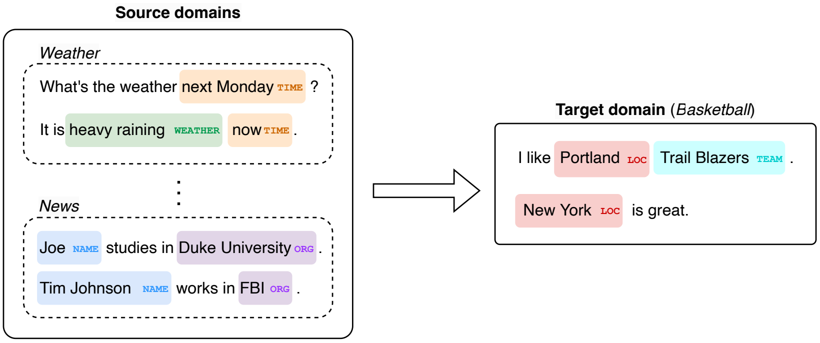

Source domains
Weather
What's the weather next Monday TIME ?
It is heavy raining WEATHER now TIME .
News
Joe NAME studies in Duke University ORG .
Tim Johnson NAME works in FBI ORG .
Target domain (Basketball)
I like Portland LOC Trail Blazers TEAM .
New York LOC is great.

Figure 4.4. Example of inputs for cross-domain transfer learning. Knowledge is being transferred to a new domain, which may even have a different set of entity types.

as weak supervision to train the tagger in the target language. However, paired sentences are not always readily available. In some cases, all that is available are individual sentences in a high-resource language, or a multilingual corpus that includes samples from multiple languages.

Prompt-based transfer Recent work in NLP has demonstrated the impressive gains obtainable with the pre-training and fine-tuning approach, especially when applied to transformer language models [249]. During the pre-training phase, a large unsupervised dataset is used to train the language model to find informative representations of inputs which can then be used to solve downstream tasks. Hofer et al. [82] show that pre-training on domain-specific corpora and reducing out-of-vocabulary words can significantly improve performance of NER models in few-shot settings. The number of publicly available pre-trained models in a variety of domains and languages is high and constantly increasing[5]. Just to mention one example, *BioBERT* [124] pretrains a BERT-based language model on PubMed abstracts and PMC full-text articles to apply the advancements of transformers for biomedical text understanding. However, how to replicate these contributions in low-resource languages, where also unsupervised text data is difficult to obtain, is an open challenge. Bondarenko et al. [22] fine-tune

[5] https://huggingface.co/models

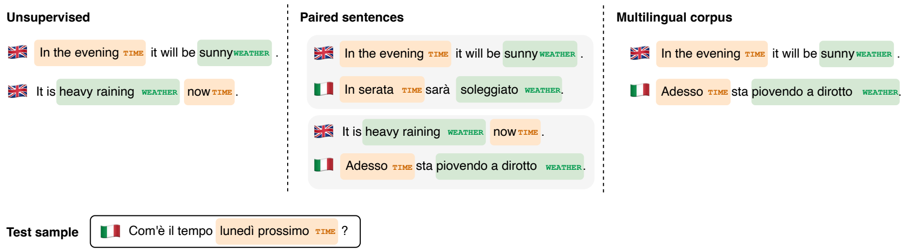

Unsupervised
In the evening TIME it will be sunny WEATHER .
It is heavy raining WEATHER now TIME .
Paired sentences
In the evening TIME it will be sunny WEATHER .
In serata TIME sarà soleggiato WEATHER.
It is heavy raining WEATHER now TIME .
Adesso TIME sta piovendo a dirotto WEATHER.
Multilingual corpus
In the evening TIME it will be sunny WEATHER .
Adesso TIME sta piovendo a dirotto WEATHER.
Test sample
Com'è il tempo lunedì prossimo TIME ?

Figure 4.5. Training corpora for cross-language transfer. In *unsupervised* scenarios, we only have access to high-resource language data, but need to extract entity mentions for a low-resource language. The ideal scenario is when we have both the high-resource language annotation and the corresponding low-resource language annotation available (*paired sentences*). In some cases, we can use a *multilingual* corpus that contains both high- and low-resource language samples.

a BERT language model pre-trained on russian data (RuBERT) to adapt it to NER and Relation Extraction. Schneider et al. [213] transfer learned the information encoded in a multilingual BERT model to a corpora of clinical narratives and biomedical scientific papers in Brazilian Portuguese. Reimers et al. [201] propose a knowledge distillation based approach to extend existing sentence embeddings models to new languages.

Recent work leverages *prompts* to exploit the knowledge acquired by such architectures during the pre-training phase by re-phrasing the task to a masked language modeling task which is closer to the target NER task [79, 67, 43, 88]. Figure 4.6 shows the workflow of PromptSlotTagging [88] as an example.

4.2.2 Meta-learning

Inspired by human intelligence, meta-learning (a.k.a. learning to learn) [212, 240] aims to quickly adapt models to new tasks without the need to re-train them from scratch and with only few data points. As humans, we are able to easily learn new skills after a few minutes or zero experience: for example, if we can ride a bike, we will easily learn to ride a motorcycle. This can be accomplished by ML models with a meta-learning phase during which the model learns to adapt to a large variety of

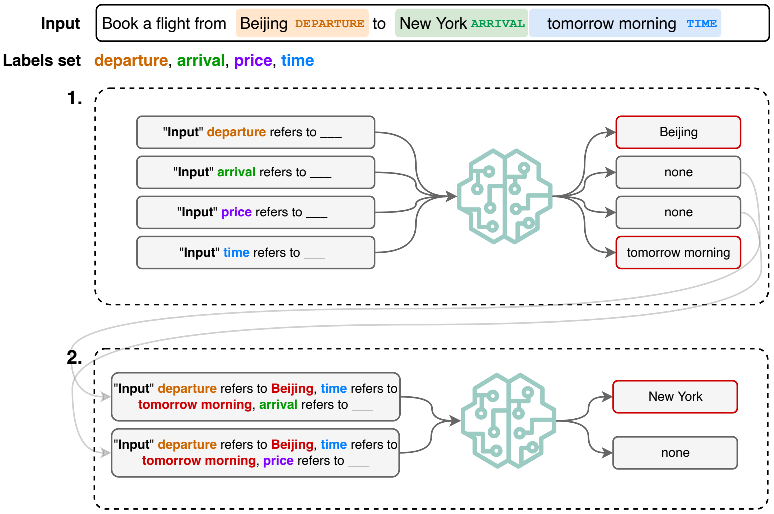

Input
Book a flight from Beijing DEPARTURE to New York ARRIVAL tomorrow morning TIME
Labels set
departure, arrival, price, time
1.
"Input" departure refers to ___
"Input" arrival refers to ___
"Input" price refers to ___
"Input" time refers to ___
Beijing
none
none
tomorrow morning
2.
"Input" departure refers to Beijing, time refers to tomorrow morning, arrival refers to ___
"Input" departure refers to Beijing, time refers to tomorrow morning, price refers to ___
New York
none

Figure 4.6. PromptSlotTagging model. In the first phase, the input sentence is embedded with inverse prompts and decoded by the language model. In the second phase, predictions are iteratively refined by reinforcing prompts with previously-predicted values.

tasks. While having being widely applied for few-shot image classification [59, 138, 195, 260], only recently meta-learning attempts have been made in NLP applications. Gu et al. [71] used meta-learning in neural machine translation, adapting the model to low-resource languages. Huang et al. [95] applied MAML to the query generation task. Qian and Zhou [193] proposed DAML, which learns general and transfereable information by combining multiple dialog tasks during training. Lin et al. [182] use meta-learning to generate personalized responses by leveraging just a few dialog samples. Recently, several works have applied meta-learning techniques to the NER task [63, 135, 117, 87, 184].

4.2.3 Summary and Discussion

Since when deep learning has started to proliferate, model architectures have been the focus of research. In the context of FS-NER, works are based on transfer or meta learning. The two approaches, while similar in their objective (i.e. improving the performance in the presence of scarse data resources), are slightly different in the mode in which they operate: while transfer learning uses a model trained with data from a similar domain or language to shift its knowledge to another model, meta learning focuses on training procedures allowing models to quickly adapt to new tasks.

When a model trained on a similar task is available, transfer learning can be easily applied without too many adaptation, and guarantees high-quality results, especially with transformer architectures. However, this ideal scenario is quite uncommon, since real-world task may be similar but deal with another language, or other entity types, thus requiring many adaptations which may also limit the resulting performance — e.g. multilingual transfer usually requires a machine-translation step whose error inevitably propagate across the transfer-learning framework. To overcome this, meta-learning models easily adapt when new tasks emerge (e.g. a new entity type is required to be recognized).

4.3 Data-centric methods

The ever-increasing attention towards Data-Centric AI [187] is reflected in a high number of FS-NER works focusing on data to improve performance. In this section, we review data-centric FS-NER approaches by separately focusing on *data augmentation*, *distant supervision*, *active learning* and *self learning*. Hence, we summarize their key characteristics and discuss similarities and differences under each category as well as limitations that have not been addressed yet.

4.3.1 Data Augmentation

One common way to deal with the lack of data is *data augmentation*, which consists in increasing the size of the available dataset with new samples generated by means of heuristics or external data sources. Augmentation methods explored in current literature for natural language processing (NLP) tasks usually manipulate words in the original sentence by word replacement [27], random deletion [262], word position swap [164] and generative models [276]. Applying these transformations to NER input samples is not possible due to the token-level classification implied by this task (each manipulation impacts labels). Thus, data augmentation

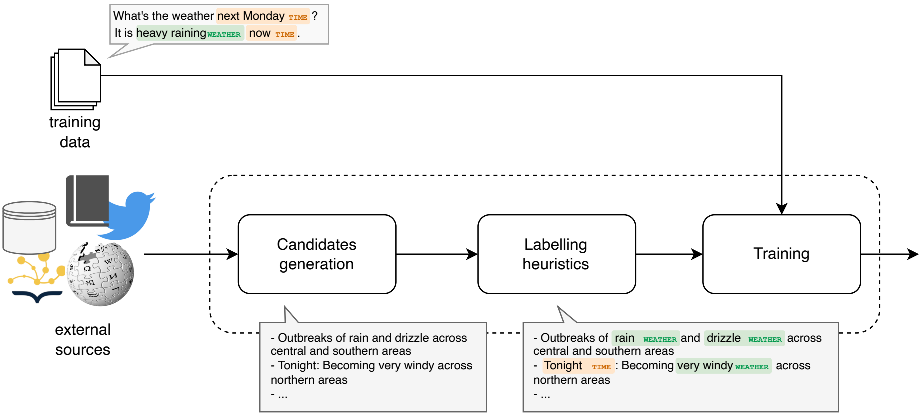

What's the weather next Monday TIME ?
It is heavy raining WEATHER now TIME .
training data
external sources
Candidates generation
Labelling heuristics
Training
- Outbreaks of rain and drizzle across central and southern areas
- Tonight: Becoming very windy across northern areas
- ...
- Outbreaks of rain WEATHER and drizzle WEATHER across central and southern areas
- Tonight TIME : Becoming very windy WEATHER across northern areas
- ...

Figure 4.7. General flowchart of a distant supervision approach

techniques for NER are comparatively less studied [45]. However, recent studies show that data augmentation is a promising direction to improve FS-NER performance [146, 279, 45, 16, 34].

4.3.2 Distant Supervision

In few-shot scenarios, labelled data could be retrieved from heuristics, different domains or laguages, external knowledge bases or ontologies. *Distant supervision* [166] aims to leverage such resources to heuristically annotate training data. For example, in biomedicine there are a lot of curated resources available: *NCBO Bioportal* [263] houses 541 biomedical ontologies, *Medical Subject Headings (MeSH)*[6] is a controlled vocabulary with 347, 692 classes of medical items, and so on. Combining ontologies is a difficult task due to their heterogeneous structures, concept granularities and overlaps or conflicts between definitions of entities. Generally, the main steps of distant supervision are (1) *candidate generation*, i.e. the identification of potential entities, and (2) *labeling heuristics* to generate noisy labels, as shown in the example of Figure 4.7. The use of distant supervision for FS-NER in low-resource languages has yet to be deeply explored. The amount of external information available in low-resource settings might be very limited: for example, the Wikipedia knowledge graph contains 4

[6] https://www.nlm.nih.gov/mesh/meshhome.html

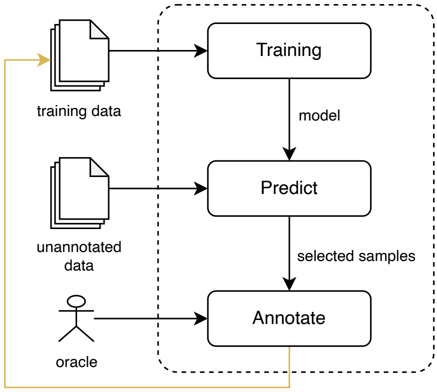


Figure 4.8. Active Learning workflow.

million person names in English while only 32 thousand in Yorùbá [1]. Furthermore, without further tuning under better supervision, distantly supervised models have low recall [28]. Several studies explore the benefits of distant supervision to the NER task [62, 220, 272, 143, 28, 155, 203, 121].

4.3.3 Active Learning

Active Learning aims to select *informative* sets of examples for training, actively querying the user for labels, as shown in the cycle depicted in Figure 4.8. The most common approach is *uncertainty sampling* [126], in which the model selects examples based on the uncertainty of its predictions (a general approach uses entropy as an uncertainty measure [223]). While its theoretical properties have been extensively studied in past works [48, 15, 14, 20], Active Learning approaches are recently spreading in natural language processing tasks. Zhang et al. [281] are the firsts to investigate active learning for sentence classification: they use the *Expected Gradient Length (EGL)* [217] as an active learning strategy aiming to select the instances which would result in the maximum change in the current model parameter estimates if their labels were provided. Shen et al. [222] are the firsts to explore AL methods on Deep Neural Networks (DNNs) for the NER task.

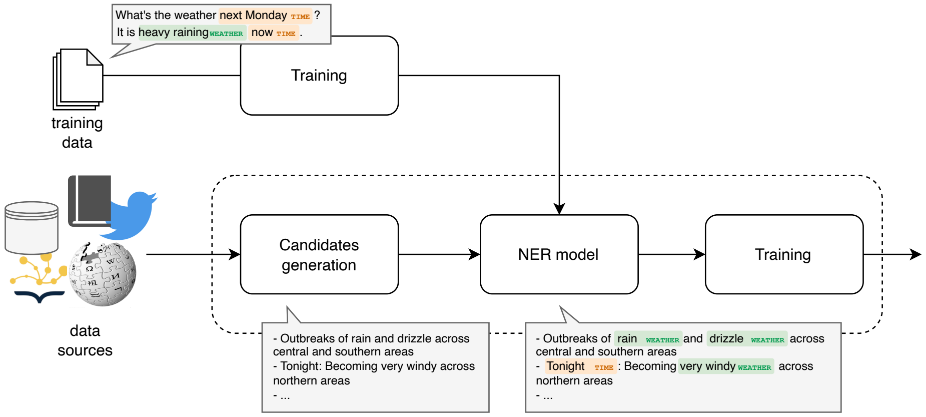

What's the weather next Monday TIME ?
It is heavy raining WEATHER now TIME .
training data
Training
data sources
Candidates generation
NER model
Training
- Outbreaks of rain and drizzle across central and southern areas
- Tonight: Becoming very windy across northern areas
- ...
- Outbreaks of rain WEATHER and drizzle WEATHER across central and southern areas
- Tonight TIME : Becoming very windy WEATHER across northern areas
- ...

Figure 4.9. General flowchart of a self-training approach.

4.3.4 Self-training

Self-training, also referred to as *self-learning*, is an approach similar to distant supervision, where the model is trained on examples labelled by the model itself. As shown in Figure 4.9, the difference is that the labelling heuristics is replaced by the model itself. Originally proposed by Scudder et al. [216], it is one of the earliest semi-supervised methods. In the NLP field, it has been successfully applied to neural machine translation [77] and sentence classification [175]. Zoph et al. [286] show that self-training guarantees improvements in performance in both high- and low-data scenarios, while data augmentation results sometimes in decreases of performance of pretraining. The potential of self-training to FS-NER has been explored in several works [145, 32, 278, 142, 91].

4.3.5 Summary and Discussion

Data-centric approaches for FS-NER try to make the most of the few available training samples or to leverage in the cleanest way possible an available but unannotated corpus, thanks to the intervention of humans, external resources and/or the model itself.

When the available few-shot training corpus is the one and only source of information available, data augmentation can be usually applied since it increases the size of the dataset by transforming the available samples.

However, the majority of research work assumes the presence of external resources such as a vocabulary of entities [279] or synonyms [45].

In general, an unannotated corpus is the key to achieve better results in few-shot contexts — methods basically differ on how they handle it to find greedy annotations. Active learning usually requires one (or many) human (s) to optimize the trade-off between the annotation efforts required and the resulting performance. On the other hand, distant supervision leverages external resources, such as heuristic rules and ontologies, to obtain weak labels. Similarly, self-learning gets weak labels by the model itself, which could be particularly useful when using language models to leverage the knowledge they acquired during the pre-training stage.

4.4 Conclusion

This chapter has provided a comprehensive review of state-of-the-art algorithms for few-shot Named Entity Recognition. The application field is analyzed and discussed in detail, and a taxonomy is proposed to summarize existing techniques into two macro-categories: model-centric and data-centric. Based on how models are defined and data are manipulated to address the few-shot learning task, methods in each macro-category are further categorized into subgroups.

Chapter 5

Information Extraction in Healthcare: the value of pre-trained language models

Knowledge graphs serve as a foundational framework for the representation and organization of heterogeneous information sources, such as clinical narratives and medical ontologies. These graphs can be harnessed for various downstream applications, including adverse event prediction and health risk assessment. The efficacy of knowledge graphs hinges on a meticulous information-extraction phase, which discerns and assimilates pertinent concepts from the data. Within this paradigm, the extraction of salient facts is paramount, as this phase profoundly influences the subsequent analytical processes. For instance, in the realm of electronic health record (EHR) analysis, the accurate extraction of patient medical events is crucial, as any oversight can culminate in flawed or incomplete insights.

However, the development of proficient information extraction systems in the healthcare domain is fraught with challenges. The intrinsic complexities of healthcare data, coupled with its unique challenges, pose significant impediments.

A primary challenge is the existence of synonyms, alternate orthographies, and polysemous terms in medical lexicon. Multiple synonyms or variant spellings for a single medical term can introduce ambiguity in data interpretation. For example, a medical condition might be denoted by diverse nomenclatures or acronyms across different medical documents, complicating the task for automated systems to consistently identify and interpret these nuances.

> **Example** — Synonyms and alternate spellings
> "Myocardial Infarction" is also known as "Heart Attack". Similarily, "Hemorrhage" can be referred to as "Bleeding" or "Haemorrhage."

Polysemous terms, which assume varied meanings based on their context, are prevalent in medical discourse. Such terms can lead to misconceptions if not contextualized appropriately by the extraction system.

> **Example** — Polysemous words
> Let us consider two clinical notes, where the word *cold* assumes different meanings:
>
> 1. The patient presented with a runny nose, sneezing, and a mild cough. The symptoms are consistent with a common *cold*. Advised rest, hydration, and over-the-counter cold medications.
>
> 2. The patient mentioned feeling a bit *cold* during the examination. Provided a warm blanket to ensure comfort.
>
> In the former, "cold" denotes a viral ailment, while in the latter, it pertains to the sensation of low temperature.

Furthermore, the paucity of annotated datasets poses a major challenge. Constructing machine learning models for information extraction necessitates voluminous labeled datasets. However, curating such datasets in healthcare is labor-intensive, necessitating domain experts to annotate data meticulously. This scarcity is even more pronounced for under-resourced languages. Additionally, the evolving nature of medical knowledge mandates perpetual updates to the extraction systems. With the incessant emergence of novel medical findings, it is imperative that the system remains updated, entailing a sustained and resource-intensive commitment.

In this chapter, the efficacy of transformer architectures is demonstrated through the proposed solution to the Disease Text Mining (DisTEMIST) challenge at BioASQ 2022 [208]. The presented solution secured the top-ranking position among nine participating teams in the Named Entity Recognition track. In Section 5.1, general information about the DisTEMIST challenge is provided, outlining its objectives and scope. Following that, in Section 5.2, a dive into the specifics of the winning approach is taken, with a detailed account of how transformer architectures played a pivotal role in achieving the outcome. Through this research, the significant advancements in NLP are highlighted, particularly the transformative impact of transformer-based models in tackling real-world challenges in biomedical text understanding, even in situations with limited available data.

5.1 The Disease Text Mining (DisTEMIST) challenge at BioASQ 2022

The Disease Text Mining (DisTEMIST) challenge [167] has been held at BioASQ[1] 2022, that invited researchers, biomedical professional and NLP experts to develop systems able to index the content about diseases in biomedical texts. The corpus contains a collection of 1000 clinical case reports written in Spanish, containing annotations of disease mentions on clinical notes, manually mapped to SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) codes.

The building process of the dataset is detailed in Figure 5.1. It was produced with meticulous attention to detail, complying to detailed annotation requirements and requiring a significant amount of manual text labelling done by clinical professionals. DisTEMIST documents from a variety of medical specialities, including cardiology, ophthalmology, infectious diseases, urology, oncology, paediatrics, tropical diseases, internal medicine, dentistry, and others, were purposefully chosen to represent a wide range of disorders. The goal of this diversity was to make it easier to transfer knowledge between other fields and textual sources.

The DisTEMIST Gold Standard corpus production process can be broken down into two separate phases. First, there was a manual text annotation phase in which the annotators located and labelled text references to diseases. Second, each of these observed disease mentions was given a

[1] http://bioasq.org

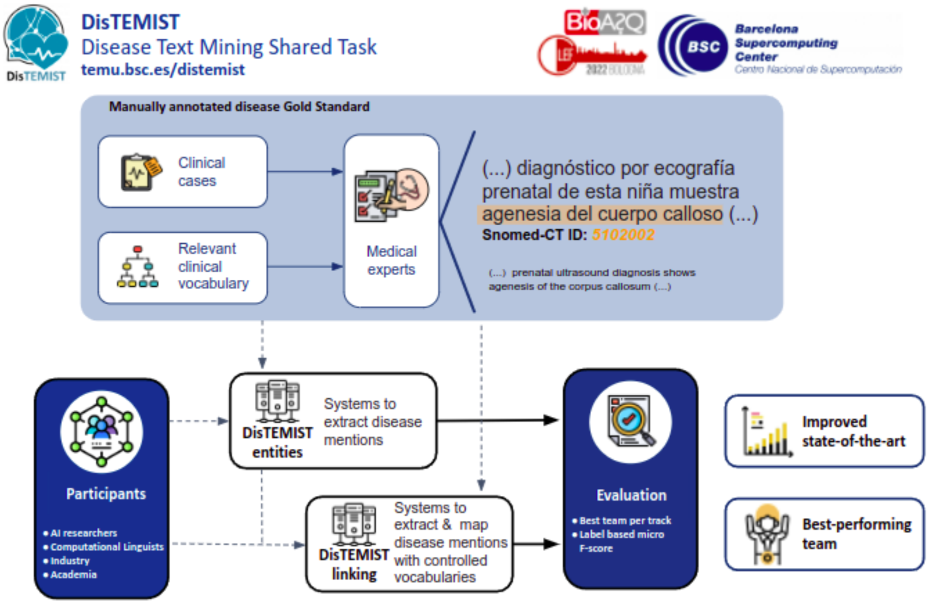

DisTEMIST
Disease Text Mining Shared Task
temu.bsc.es/distemist
BioASQ
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Manually annotated disease Gold Standard
Clinical cases
Relevant clinical vocabulary
Medical experts
(...) diagnóstico por ecografía prenatal de esta niña muestra agenesia del cuerpo calloso (...)
Snomed-CT ID: 5102002
(...) prenatal ultrasound diagnosis shows agenesis of the corpus callosum (...)
Participants
• AI researchers
• Computational Linguists
• Industry
• Academia
DisTEMIST entities
Systems to extract disease mentions
DisTEMIST linking
Systems to extract & map disease mentions with controlled vocabularies
Evaluation
• Best team per track
• Label based micro F-score
Improved state-of-the-art
Best-performing team

Figure 5.1. Overview of the DisTEMIST Shared Task. Illustration from [167].

unique SNOMED CT identity. There were, however, a number of difficulties in normalising the disease's references to SNOMED CT. These issues resulted from the wide range of terminology changes over time, variations in how clinicians expressed the same condition, and the complexity and diversity of clinical entities and expressions, some of which lacked an ideal SNOMED CT identifier.

5.1.1 Corpus statistics

Table 5.1 summarizes the characteristics of the DisTEMIST corpus. It consists of 1,000 documents with 16,678 sentences and 406,318 tokens in total. The training set and test set were split randomly into two sets, the latter of which had 250 records and was put aside for team evaluation. Each of the 10,665 instances of disease mentions in the corpus' documents was manually mapped to a corresponding SNOMED CT word, yielding 7,303 distinct codes. The annotated training set was made available during

Table 5.1. DisTEMIST corpus statistics

	Documents	**Annotations**	**Unique codes**	**Sentences**	**Codes**
Training	750	8,066	4,817	12,499	305,166
Test	250	2,599	2,484	4,179	101,152
Total	1,000	10,655	7,303	16,678	406,318

the challenge, however because to uncertainties regarding the quality of normalisation for the remaining 165 documents, only 585 documents had their disease mentions normalised to SNOMED CT. After the shared task, the annotation for these 165 documents was provided.

5.1.2 Gazetteer

To facilitate the normalisation process, challenge organizers provided a subset of SNOMED CT codes with concepts relevant to DisTEMIST diseases. A dictionary has been provided to participants, containing 147,280 entries, of which 111,177 are SNOMED CT main terms. A name and a description are associated to each code.

5.2 Methodology

Figure 5.2 shows an overview of the methodological flow of our solution for the DisTEMIST track. A Transformer backbone network pre-trained with Spanish biomedical corpora has been used in both NER and EL tasks. In the former case, it has been used to compute *token embeddings* for a classification head with a linear layer and a softmax activation function; in the latter, it computes *concept embeddings* for each concept within the *gazetteer*, which will be then used to link an entity mention to the nearest concept based on a measure of similarity. In this section, each module of our methodology will be extensively described.

5.2.1 Biomedical Transformer Backbone network

The Biomedical Transformer Backbone network used in this work has been pre-trained and made publicly available by Carrino et al. [29]. It uses

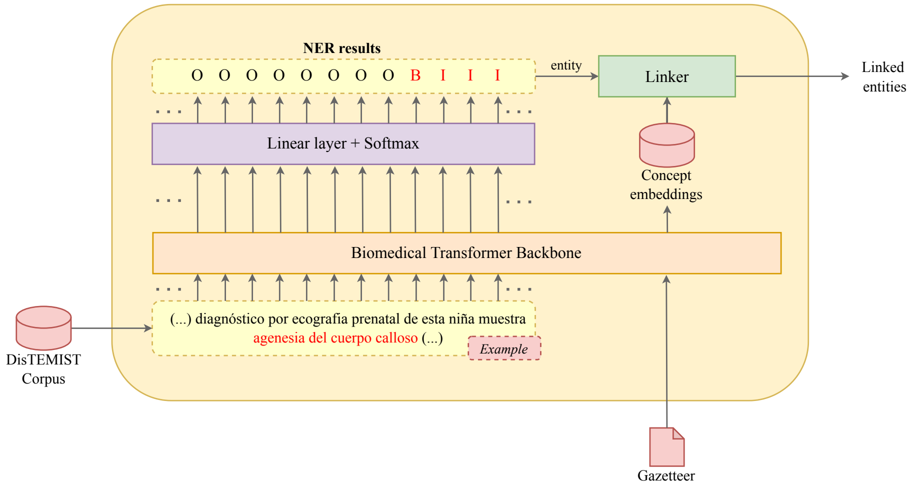


Figure 5.2. Overview of our NER + EL solution for the DisTEMIST track. A biomedical Spanish pre-trained Transformer backbone network is used to compute: (1) *token embeddings* to be classified by a classification head (linear layer + softmax); (2) *concept embeddings* for each concept within the *gazetteer* which will be used by the *Linker* to associate the nearest concept to an entity mention based on similarity measures.

a RoBERTa [151] base model with 12 self-attention layers with masked language modeling as the pre-training objective. The dataset used to pre-train the network consists in two corpora with different sizes and domains:

- *Clinical corpus*: it contains 91M tokens from more than 278K clinical documents (e.g. discharge reports, clinical course notes).

- *Biomedical corpus*: it contains data from a variety of sources, such as medical crawlers, PubMed[2] and Scielo[3] publications and patents. The entire corpus counts a total of 968M words.

[2] https://pubmed.ncbi.nlm.nih.gov

[3] https://scielo.org

5.2.2 Named Entity Recognizer

The Transformer-based backbone network is used to extract an embedded representation of each token x_j in an input sample $\mathbf{x}$, $\mathbf{z} = f_{\theta_{LM}}(x_j)$, θ_{LM} being the set of language model parameters. Thereafter, a linear layer (a.k.a. *classification head*) with parameters $\theta_L = \{\mathbf{W}, \mathbf{b}\}$ project the Transformer-based representation $\mathbf{z}$ into the label space, $f_{\theta_L}(\mathbf{z}) = Softmax(\mathbf{Wz} + \mathbf{b})$. The model parameters are then optimized by minimizing cross-entropy:

$$\mathscr{L}_{CE} = \sum_{(\mathbf{x},\mathbf{y})\in\mathscr{D}} \sum_{i=1}^{H} KL\Big(y_i \Big| q(y_i|x_i)\Big), \tag{5.1}$$

where $KL(p|q)$ is the Kullback-Leibler divergence between the two distributions p and q, and q is the prediction probability vector for each token:

$$q(y|x) = Softmax(\mathbf{W} \cdot f_{\theta_{PLM}}(x) + \mathbf{b}) \tag{5.2}$$

5.2.3 Entity Linker

Inspired by Kraljevic et al. [112], our EL approach relies on a *Concept Database (CDB)* component, i.e. a table representing a concept dictionary. To this end, e used the gazetteer provided by DisTEMIST track organizers. Even though not every concept within the gazetteer appears in our training set, we decided to keep all the concepts due to the unpredictability of concepts in the test set. Our linking approach is based on *context similarity*: we learn an embedded representation for each concept and for new documents, when an entity mention is detected by the NER model, its context is compared to the embedded representations of all the concepts in the CDB to choose the most appropriate one.

Concept Embeddings We learn concept embeddings in a supervised fashion. For each concept $c \in CDB$, we perform the steps described as follows to compute its concept embedding $V^c_{concept}$:

1. *Initialization*: given the concept name c_{name} and its description $c_{description}$ provided with the gazetteer, we initialize $V^c_{concept}$ with the

embedding of the concatenation of the two strings $[c_{name}, c_{description}]$ computed with the Biomedical Transformer backbone network.

2. *Context embeddings*: for each entity in the training set annotated with the concept c, we compute its context embedding $V_{context}$ with the Biomedical Transformer backbone network.

3. *Update*: for each entity in the training set annotated with the concept c, the concept embedding $V^c_{concept}$ is updated with the context embedding $V_{context}$. Specifically, the update criterion is described by the following equation:

$$V^c_{concept} = V^c_{concept} + lr \cdot (1 - sim) \cdot V_{context}, \tag{5.3}$$

where:

- lr is the *learning rate*, computed as $lr = \frac{1}{\mathcal{N}_c}$, $\mathcal{N}_c$ being the number of times the concept appears during training.
- sim is the cosine similarity between $V^c_{concept}$ and $V_{context}$,

$$sim = max\left(0, \frac{V^c_{concept}}{\|V^c_{concept}\|} \cdot \frac{V_{context}}{\|V_{context}\|}\right) \tag{5.4}$$

Linking Given the entity mention recognized by the NER model, we compute its context embedding $V_{context}$ by means of the Biomedical Transformer backbone network. Then, we compute its cosine similarity sim with all the concept embeddings $V_{concept}$. We eventually link the entity with the most similar concept.

5.2.4 Experiments

The performance of our proposed approaches for NER and EL has been evaluated by participating to the *DISease TExt Mining Shared Task (DisTEMIST)* track within the BioASQ 2022 challenge. In this section we show the performance results of our methodology on the final test set and some preliminary experiments on the training corpus provided by the challenge organizers.

Experimental setup

Evaluation Metrics Evaluation is done by comparing the automatically generated results to the results generated by manual annotation of experts. The primary evaluation metric for both the NER and EL subtracks will consist of micro-averaged precision (*MiP*), recall (*MiR*) and F1-scores (*MiF1*).

Configuration Both the NER and EL models were implemented using the HuggingFace Transformers library (v4.4.0) [264]. The biomedical Spanish Transformer backbone network[4] has been downloaded from the HuggingFace model repository. To deal with the limited length of input samples, we consider each sentence in a *clinical case* as a separate input samples for our models. In a preliminary phase to our submission, we studied the effects of various hyperparameters and the generalization error of our models by splitting the original corpus of clinical cases in three parts: (1) a *training set* (60% of the original corpus) used to train the model, (2) a *validation set* (20% of the original corpus) to evaluate the effects of hyperparameters and (3) a *test set* (20% of the original corpus) to evaluate the ability of our models to generalize to unseen data. We fine-tune our models with a Google Colab environment, which provided us a Tesla T4 GPU.

Results

NER hyperparameters and evaluation We studied the effects of different hyperparameters on our validation set:

- `learning rate`: the initial learning rate for AdamW optimizer. Initialized to `5e-5`.
- `weight decay`: the weight decay to apply to all layers except all bias and LayerNorm weights in AdamW optimizer. Initialized to `0` (no weight decay applied)
- `batch size`: the batch size per device (e.g. CPU, GPU) for training. Initialized to `16`.

[4] `PlanTL-GOB-ES/roberta-base-biomedical-clinical-es`

Table 5.2. NER hyperparameter selection (first stage)

`batch size`	`learning rate`	`weight decay`	**MiP**	**MiR**	**MiF1**
`16`	`5e-5`	`0.0`	0.7199	0.7759	0.7521
`16`	`4e-5`	`0.0`	0.7136	0.7677	0.7448
`16`	`3e-5`	`0.0`	0.7095	0.7749	0.7460
`16`	`2e-5`	`0.0`	0.6805	0.7672	0.7263
`16`	`1e-5`	`0.0`	0.6209	0.7175	0.6704
`16`	`6e-5`	`0.0`	0.7274	0.7836	0.7598
`16`	`7e-5`	`0.0`	0.7370	0.7822	0.7624
`16`	`8e-5`	`0.0`	**0.7400**	**0.7836**	**0.7642**
`16`	`9e-5`	`0.0`	0.7331	0.7827	0.7571
`16`	`1e-4`	`0.0`	0.7373	0.7754	0.7612
`16`	`8e-5`	`0.1`	0.7375	0.7885	0.7675
`16`	`8e-5`	`0.2`	**0.7428**	**0.7865**	**0.7694**
`16`	`8e-5`	`0.3`	0.7396	0.7846	0.7668
`8`	`8e-5`	`0.2`	**0.7479**	**0.7865**	**0.7722**

For each experiment, we train the NER model for two epochs — at the end of the selection process we will analyze the effects of an increased number of epochs. Our search for hyperparameters divides into two stages: in the first stage, we make hyperparameters vary in large ranges with the aim to detect a smaller range where we will perform a *grid search*. All the different configurations and associated performance results are listed in Table 5.2.

In the second stage, we perform a grid search based on a uniform distribution within the following hyperparameters ranges (which have been chosen based on the results of the first stage):

- `learning rate: [7e-5, 8e-5]`
- `weight decay: [0.1, 0.2]`
- `batch size: 8`

Given the best results from the grid search, we increased the number of training epochs with an *early stopping* criterion, by stopping training when the performance on the validation set does not increase for 5 con-

Table 5.3. Final preliminary NER results

epochs	batch size	learning rate	weight decay	**MiP**	**MiR**	**MiF1**	
18	8	8.516e-5	0.1844	0.7814	0.8031	0.7921	best hyper-parameters
18	8	8.516e-5	0.1844	0.7738	0.7931	0.7833	internal test set error

Table 5.4. Entity linking "internal" test set results with and without using the gazetteer.

System	**MiP**	**MiR**	**MiF1**
With gazetteer	0.7374	0.7374	0.7374
Without gazetteer	0.7374	0.7374	0.7374

secutive epochs. The final preliminary results and the generalization error are shown in Table 5.3.

EL evaluation We evaluated results of our linking module with and without the *gazetteer*: challenge organizers declared that it contains all the possibile links to all the entity mentions in the test set. However, its size (113609 concepts) is much higher w.r.t. the number of concepts in our training set (2430 concepts). When a concept does not appear in the training set, its embedding is determined by its name and description, which could result in many "noisy" concepts leading to wrong linking results. Table 5.4 reports results on our "internal" test set (a 20% subset of the training files provided for entity linking) obtained with and without the gazetteer, i.e. we considered only concepts appearing at least one time in the training set. Since results are equivalent, we decided to keep the gazetteer for our submission.

Leaderboard Official results of the DisTEMIST track are reported in Table 5.5 (NER) and Table 5.6 (EL). Specifically, we show (1) our results, (2) results from the best participant team (second-best in case of NER, since our team is the first ranked), and (3) median results (computed by considering the best submissions of each participant team). While the

Table 5.5. Official results of BioASQ DisTEMIST NER task. We show our result, the second-best result and median result (computed by considering just the best MiF1 score for each participant team).

System	**MiP**	**MiR**	**MiF1**
Ours	**0.7915**	**0.7629**	**0.7770**
Second-best participant	0.7434	0.7483	0.7458
Median	0.7146	0.6736	0.6935

Table 5.6. Official results of BioASQ DisTEMIST linking task. We show our result, the best result and median result (computed by considering just the best MiF1 score for each participant team).

System	**MiP**	**MiR**	**MiF1**
Ours	0.2814	0.2748	0.2780
Best participant	**0.6207**	**0.5196**	**0.5657**
Median	0.4795	0.2292	0.3102

domain-specific pre-training of the backbone network has been the key for a successful NER system, the EL solution seems to suffer from a design flaw. We can indeed observe a big discrepancy between results on our internal test set and the leaderboard, which may be caused by two main factors: (1) pipelined errors of NER and EL predictions and (2) the inappropriateness of the size of the training set: the gazetteer size (113609 concepts) suggests us that the leaderboard test set contains many concepts which are not present in our training set (2430 concepts). However, our context-based EL methodology computes embedded representations of concepts based on their occurrences in the training set, and all the other concepts are represented with their description provided with the gazetteer, which may be useless or even detrimental for similarity computation. Further investigations to handle the above-described problems are thus needed.

5.3 Conclusion

In this chapter, a solution for the Disease Text Mining (DisTEMIST) challenge at BioASQ 2022 has been presented. The described approach leverages the capabilities of transformer architectures, specifically a pre-trained RoBERTa model fine-tuned on Spanish biomedical corpora, to address the challenges of Named Entity Recognition and Entity Linking in the healthcare domain. This solution secured the top-ranking position in the NER track, exemplifying the efficacy of transformer-based models in extracting and linking pertinent medical information from clinical texts.

The challenges observed in the healthcare domain, like the presence of synonyms, alternate orthographies, and polysemous terms, underline the necessity of context-aware models. Transformer architectures, with their capacity to encapsulate contextual information, have demonstrated to be a substantial asset in overcoming these challenges. Moreover, the dearth of annotated datasets in the healthcare domain accentuates the importance of pre-trained models that can be fine-tuned for specific tasks with limited data.

The engagement in the DisTEMIST challenge yielded insightful experiences regarding the practical application of NLP techniques in the biomedical domain. The results garnered in the NER track authenticate the potential of transformer-based models in extracting relevant entities from clinical texts. Nonetheless, the discrepancies noted in the EL track between the internal test set and the leaderboard results signify a requisite for continued research and refinement in the entity linking process.

In conclusion, the advancements in NLP, especially the advent of transformer architectures, have inaugurated new pathways for information extraction in the healthcare domain. With the ongoing evolution of the medical field, the assimilation of NLP techniques is anticipated to be instrumental in harnessing the expansive amounts of unstructured data to distill meaningful insights and augment patient care.

Chapter 6

Data augmentation via context-based similarity

In Chapter 4, it was discussed that the training of NER models typically requires a substantial amount of annotated data to serve as a reliable reference. However, the process of generating accurate annotations is noted to be both time-consuming and costly, particularly in specialized fields such as law, history, or medicine, where a high degree of specialized knowledge is essential. These factors can pose challenges in securing funding for the acquisition of extensive annotated data. Moreover, experts in these specialized domains, being often engaged with their primary responsibilities, may have limited availability for annotation tasks.

Data augmentation, which entails expanding the size of the given dataset with additional samples produced by using heuristics or external data sources, is one popular method of addressing the scarcity of data. For Natural Language Processing (NLP) tasks, word replacement [27], random deletion [262], word position swap [164] and generative models [276] are the most common word manipulation techniques used in augmentation methods. However, the token-level classification implied by NER makes it impossible to apply these augmentation techniques to input data (each manipulation impacts labels). Consequently, NER data augmentation approaches are comparatively understudied [46], even though new studies have recently been carried out utilizing transfer learning to generate augmented datasets starting from domains with more resources[35] or using

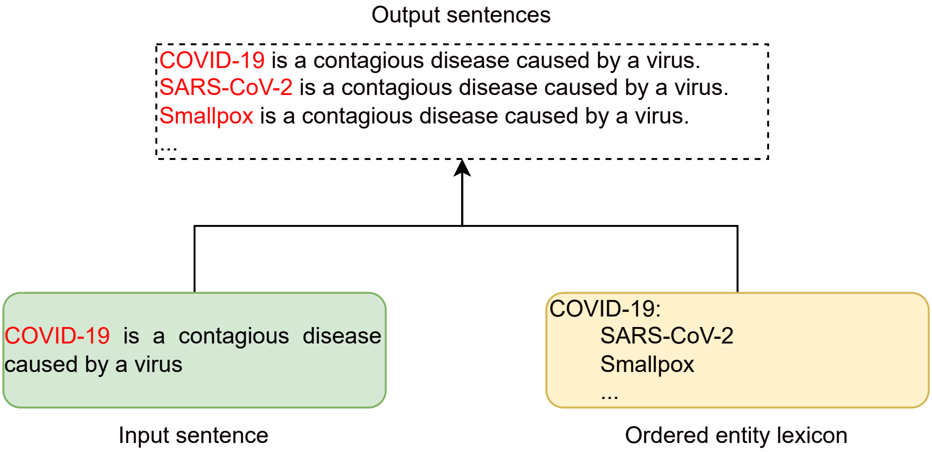

Output sentences
COVID-19 is a contagious disease caused by a virus.
SARS-CoV-2 is a contagious disease caused by a virus.
Smallpox is a contagious disease caused by a virus.
...
COVID-19 is a contagious disease caused by a virus
Input sentence
COVID-19:
SARS-CoV-2
Smallpox
...
Ordered entity lexicon

Figure 6.1. Example of NER data augmentation for an input sample. The mention *"COVID-19"* is replaced with mentions from the Entity Lexicon, which is ordered based on a context-based similarity metric

Masked Language Models (MLM)[285] to mitigate label misalignment.

Furthermore, although data augmentation has shown promising outcomes, the currently proposed data manipulation methods frequently produce an excessive amount of noisy and incorrectly classified samples since the added data may be syntactically and/or semantically erroneous.

To address the identified issue, the proposed solution is *COntext SImilarity-based data augmentation for NER (COSINER)* [16, 17], leveraging similarity metrics to generate augmented examples, thereby facilitating the creation of sentences that retain plausibility within a real-world context. This is achieved through a defined *context-based mention replacement* augmentation technique, where mentions within the input data are substituted with mentions from an *Entity Lexicon* likely to occur within the same contextual framework. Figure 6.1 illustrates an augmentation example applied to an input sentence.

Extensive experiments are conducted on three benchmark datasets in the biomedical domain to evaluate the performance of the proposed methods. Specifically, the performance of COSINER is compared against a set of baselines from existing literature, demonstrating that leveraging similarity for data augmentation yields improved performance. The choice of datasets from the biomedical domain is driven by the typical scarcity of data available for analysis in this field. However, the proposed methodology is versatile across various domains and can be deployed for any task involving named entity recognition (NER). It is further demonstrated that the enhanced performance achieved with COSINER is attributed to the first-ranked samples, obviating the need for large augmented datasets to improve results, thus offering a computational advantage.

6.1 Related Work

Data augmentation techniques based on manipulating the given corpus have been extensively investigated for problems requiring sentence-level classification [164, 262], but they have not yet received enough attention for tasks requiring token-level classification, where NER belongs. Data augmentation for NER is a promising research area, according to recent studies [46, 53].

State-of-the-art techniques are detailed below, with examples of augmented samples presented in Table 6.1.

- *Mention Replacement (MR)*: every entity mention from an input sample has a chance to be randomly replaced with a different mention from the original training set with probability p.

- *Label-wise token replacement (LwTR)*: each token from an input sample is randomly replaced (with a certain probability p) with any other word with the same label within the training dataset.

- *Synonym replacement (SR)*: each word from the input sample can be randomly replaced with probability p with a synonym, which could be retrieved by using WordNet [163].

- *Masked Entity Language Modeling (MELM)* [285]: pre-trained Masked Language Model (MLM) that has been fine-tuned to allow for mention replacement. After a linearization step used to minimize the number of tokens incompatible with the original labels, it performs masked entity prediction over the training set to generate new sentences.

Table 6.1. NER data augmentation examples. Disease entity mentions are in *italic*, while manipulated tokens are **bolded** (they may overlap).

Method	**Example**
Input example	*Breast cancer* can occur in both women and men.
MR	***Diabetes*** can occur in both women and men.
LwTR	*Breast cancer* can **complain** in both women and men.
SR	*Breast* ***tumor*** can occur in both women and men.
MELM	***Mammal cancer*** can occur in both women and men.
style_NER	**In patients with *cancer* the mortality rate is 10%**

- *Cross-Domain Named Entity Recognition (style_NER)* [35]: neural architecture that transforms sentences from a high-resource domain to a lower-resource one to generate augmented data. At each iteration the model pairs a random sentence from the source domain to the target domain, then it performs a denoising reconstruction to learn a compressed representation of the input. Finally, a step of de-transforming reconstruction is performed to convert sentences from one domain to another.

In few-shot conditions, the benefits of data augmentation strategies dramatically decline [35]. This is because text manipulation techniques may result in samples that are grammatically and semantically inaccurate, and the issue is made worse when the size of the augmented data outweighs the size of the original corpus. The key idea of the proposed system is to generate realistic augmented samples by using a similarity-based method in order to address this problem.

To the best of existing knowledge, this work represents the inaugural effort in examining *context-based similarity* for data augmentation purposes. Specifically, the approach entails the manipulation of input samples through the substitution of entity mentions with analogous mentions on both syntactic and semantic grounds, indicating a likelihood of these mentions appearing within comparable contexts.

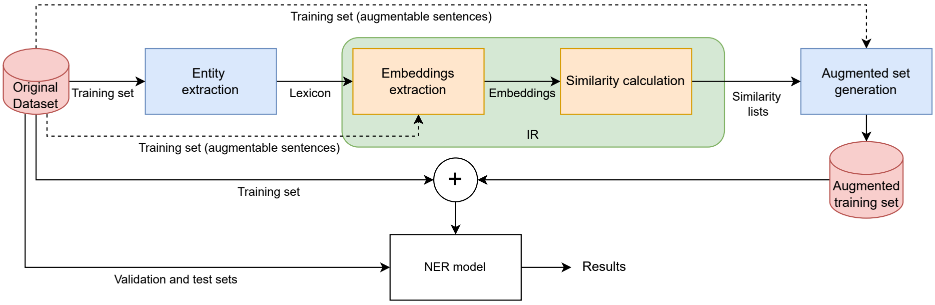


Figure 6.2. COSINER methodological flow (boxes represent steps). (1) Given the original training set, a Lexicon with all the entities is generated, (2) then all entities are mapped to an embedded space extracted from sentences with at least one mention. (3) Similarity values among embedding pairs are computed so as to link each entity to a ranked list of its most (least) similar entities. (4) The augmented training set is thus generated. (5) Finally, a model is trained by exploiting both the original and augmented training sets.

6.2 Methodology

In essence, COSINER uses mention replacement as a means of augmenting the initial training set. This technique was proposed and investigated by Dai et al. [46]. It involves choosing the entities of interest within the sentences of the dataset using a binomial distribution and replacing them with another item chosen at random from the same dataset. However, since this method is random, a lot of noisy or even incorrectly labeled samples are produced, which could have a bad impact on model performance. In order to replace the entity mention with the entities that are most similar to it in terms of syntax, semantics, and context, we use a similarity-based methodical flow. Figure 6.2 depicts the proposed methodical flow in broad strokes, and the following outlines each phase.

6.2.1 Lexicon generation

Each *concept* (i.e. entity mention) contained in the training set must be gathered so as to replace mentions. A *concept* may consist of one or more words, and we additionally store the frequency with which each word appears in the training set in the Lexicon $C_{concept}$. Depending on the number of mentions in the dataset, the number of entities in the Lexicon will vary. Note that the size of the Lexicon has a significant impact on how quickly similarity values between entity pairs are computed, nevertheless this impact is not a limitation as we conduct experiments under few-shot circumstances.

6.2.2 Embeddings extraction

"A numerical representation, denoted as $V_{concept}$, is needed for every concept in the Lexicon to determine the similarities between entities within the dataset. This is achieved by employing a pre-trained language model, as referenced in [52, 24], as a feature extractor. For each phrase in the Lexicon containing a specific mention, the extractor takes the phrase and maps each token to its corresponding word embedding, $V_{context}$. The $V_{context}$ represents the token in its specific context as an array of numerical features. When mentions consist of multiple tokens, the $V_{context}$ is computed by averaging the word embeddings of those tokens. Once a $V_{context}$ is determined, the overarching numerical representation, $V_{concept}$, is adjusted as delineated below:

$$V_{concept} = V_{concept} + lr \cdot (1 - sim) \cdot V_{context}, \tag{6.1}$$

where lr is a regularization term defined by the reciprocal of the number of times a mention appears in the whole dataset and sim is the cosine similarity between $V_{concept}$ and $V_{context}$:

$$lr = \frac{1}{C_{concept}} \tag{6.2}$$

$$sim(V_{concept}, V_{context}) = \max(0, \frac{V_{concept}}{||V_{concept}||} \cdot \frac{V_{context}}{||V_{context}||}) \tag{6.3}$$

$V_{concept}$ is initialized to the $V_{context}$ value of the first sentence where the mention appears.

6.2.3 Similarity computation

The cosine similarity between the embeddings $V_{concept}$ of every pair of entities in the Lexicon is computed so as to obtain a ranked list of similarity scores $z_{ij} = sim(V^i_{concept}, V^j_{concept})$ linked to each Lexicon entry. Two ranking criteria have been defined:

- Maximum (descending order): the first positions represent the concepts which are the most related. This increases the amount of data while keeping it as near to the training distribution as possible, allowing to generate realistic augmented samples that maintain the consistency of the context provided by the sentence;

- Minimum (ascending order): by considering the least similar entities first, we can include samples that lie as far as possible from the knowledge boundary, allowing to recognize and correctly classify extreme cases.

6.2.4 Augmented set generation

All the sentences with at least one mention are taken into account to create the augmented set. A similarity value s_m to each sentence is assigned, computed as the average of the entity similarity scores z_{ij} of the new entities within the phrase.

Two strategies have been defined:

- Local Augmentation: for each sentence, we generate k new samples. The advantage of this approach is that it takes every training sentence into consideration to generate the augmented set.

- Global Augmentation: for each sentence, k new samples are generated just like the previous strategy. Then all the new generated sentences are ranked in a single list by their similarity value s_m and select the first h elements. By doing so, there is a higher focus on samples that are closer to the initial training distribution.

In Figure 6.3 the differences between the two strategies are highlighted.

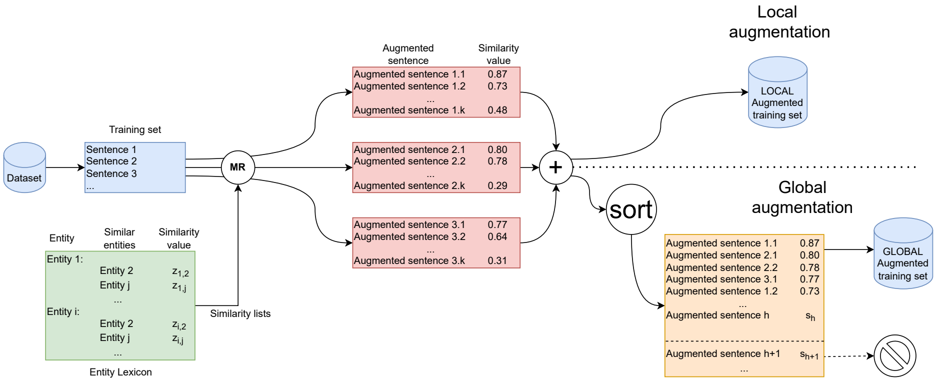

Local
augmentation
Augmented
sentence
Similarity
value
Augmented sentence 1.1 0.87
Augmented sentence 1.2 0.73
...
Augmented sentence 1.k 0.48
LOCAL
Augmented
training set
Training set
Dataset
Sentence 1
Sentence 2
Sentence 3
...
MR
Augmented sentence 2.1 0.80
Augmented sentence 2.2 0.78
...
Augmented sentence 2.k 0.29
+
sort
Global
augmentation
Augmented sentence 3.1 0.77
Augmented sentence 3.2 0.64
...
Augmented sentence 3.k 0.31
Entity
Similar
entities
Similarity
value
Entity 1:
Entity 2 $z_{1,2}$
Entity j $z_{1,j}$
...
Entity i:
Entity 2 $z_{i,2}$
Entity j $z_{i,j}$
...
Entity Lexicon
Similarity lists
Augmented sentence 1.1 0.87
Augmented sentence 2.1 0.80
Augmented sentence 2.2 0.78
Augmented sentence 3.1 0.77
Augmented sentence 1.2 0.73
...
Augmented sentence h s_h
Augmented sentence h+1 s_{h+1}
...
GLOBAL
Augmented
training set

Figure 6.3. COSINER Augmentation strategies. The initial steps of the local and global strategies are shared. First of all, k augmented sentences for each phrase with at least one mention are generated starting from similarity lists and the training set via Mention Replacement (MR), then the sentence similarity value s_m is calculated and assigned to each augmented example. For the local strategy all the augmented examples are used in the new training set. For the global strategy, a new list is created with all the examples ordered by their s_m and the first h sentences are selected for the augmented training set.

6.2.5 NER model training

The IOB2 scheme [197] is used for NER token-classification task, each token being thus associated to the B (beginning of an entity mention), I (inside) or O (outside) label. The original training dataset and the augmented samples are fed to a Transformer network backbone, [52, 24], an encoder-decoder structure that leverages stacked self-attention, point-wise and fully connected layers to extract the contextualized representation of each token x_j in an input sample $\mathbf{x}$, $\mathbf{z} = f_{\theta_{PLM}}(x_j)$, θ_{PLM} being the set of PLM parameters. Thanks to the auto-regression, the representation of the token x_j will be used as an additional input for the following token, in the BERT based architectures this relation is bidirectional, thus a token x_j will retrieve its context from both left and right. Thereafter, a linear layer (a.k.a. *classification head*) with parameters $\theta_L = \{\mathbf{W}, \mathbf{b}\}$ projects the Transformer-based representation $\mathbf{z}$ into the label space, $f_{\theta_L}(\mathbf{z}) = Softmax(\mathbf{Wz} + \mathbf{b})$. The model parameters are then optimized by minimizing cross-entropy.

6.2.6 Computational cost

Excluding lexicon generation and embeddings extraction from a formal analysis of computational times is justified as these tasks can be conducted off-line. Given:

- n represents the number of entities in the entity lexicon
- m denotes the fixed size of the embeddings extracted from these entities

It becomes necessary to compute the cosine similarity n^2 times for determining all the similarities among every entity pair. This results in a computational complexity of $O(mn^2)$ for calculating pair-wise cosine similarities across n embedding vectors of size m. The time required to generate new examples is minimal. Even with the quadratic complexity, the influence on processing duration remains slight due to the constraints of few-shot scenarios and a restricted quantity of examples, ensuring the procedure remains manageable and viable.

6.3 Experiments

In this section, the effectiveness and efficiency of the proposed method are assessed on three benchmark datasets from the biomedical field. Results indicate that the method outperforms selected benchmarks from existing literature in several datasets and few-shot scenarios. Additionally, the computing times are on par with simpler augmentation techniques and more favorable compared to complex approaches.

6.3.1 Experimental setup

Datasets and few-shot scenarios

"The method is trained and evaluated on three benchmark datasets sourced from biomedical articles. Details are as follows:

- NCBI-Disease[56]: consists of 793 PubMed abstracts, including 6,881 *disease* entity mentions;

- BC5CDR[132] consists of 1,500 PubMed articles, including 15,935 *chemical* and 12,852 *disease* mentions. We consider only *chemical* mentions in our experiments to add variety to entities since diseases have already been used in NCBI. Therefore our approach has been developed to be applied only to one entity type.

- BC2GM[225] consists of 20,000 sentences from PubMed abstracts, including 20,702 *gene* entities.

Three few-shot scenarios are defined based on the percentage of samples from the available corpora used for method application: 2%, 5%, and 10%. All experimental results are reported within these three few-shot contexts. Dataset statistics and details of these scenarios can be found in Table 6.2.

To assess the performance of the style_NER baseline, supplementary datasets with equivalent training set sizes to the benchmark datasets were employed:

- BC5CDR (Disease)[132]: we used the same BC5CDR corpus that is our target, but taking into account only disease mentions;

- CHEMDNER[115]: consists in a collection of 10,000 PubMed abstracts, including 84,355 *chemical* entity mentions;

- JNLPBA[39]: contains 2,000 abstracts from the GENIA corpus 3.02 with a taxonomy of 48 classes. We have considered only the genes mentions.

The experiments were carried out using a Kaggle notebook that provides a NVIDIA Tesla P100 GPU with 16 GB of memory and a 2-core Intel Xeon CPU with 13 GB of RAM in the configuration used.

Training details

Building on prior research in few-shot learning [209], it is assumed that the operating scenarios do not have data available for hyperparameter tuning. Consequently, hyperparameters are selected based on historical findings and pragmatic considerations. In particular, a pre-trained biomedical Transformer network [124] is utilized. All models and the elementary

Table 6.2. Statistics of the benchmark datasets used in our experiments.

		Dataset splits			Few-shot size		
Dataset	**Entity type**	*Train*	*Dev*	*Test*	*2%*	*5%*	*10%*
NCBI-disease	Disease	5425	924	941	108	271	542
BC5CDR	Chemical	4561	4582	4798	91	228	456
BC2GM	Gene	12575	2520	5039	251	628	1257

baselines are trained for 5 `epochs` with a `learning rate` of $5 \cdot 10^{-5}$, an AdamW `optimizer` [154], a `batch size` of 8, and a `maximum sequence length` of 512. For the MELM approach, a pre-trained RoBERTa model is employed as MLM to predict masked sentences. The style_NER approach utilizes a pre-trained BERT model as the foundational architecture, which is then trained for 5 epochs on source datasets. For both these strategies, a BERT model with the aforementioned parameters is trained on a new augmented dataset for 20 epochs. Each model undergoes training with five distinct seeds, and average outcomes along with 95% confidence intervals are documented. The efficacy of the methods is assessed via $F1$ scores, determined using the `seqeval`[1] Python framework.

6.3.2 Results

Comparison with baselines

Comparative results with notable baselines from the literature [46] are presented below. The benchmarks described are as follows:

For the 5 executions on NCBI-Disease in the 2% scenario using the optimal BERT:

- No Augmentation: This records outcomes achieved with the foundational training set, assessed using a pre-trained BERT model.

- No Augmentation (BioBERT): Outcomes from the foundational training set, evaluated using a specialized pre-trained BERT model for the biomedical domain, are reported here.

[1] https://github.com/chakki-works/seqeval

Table 6.3. Dataset used to transfer knowledge from a source domain to our target domain

Source Dataset	**Target Dataset**
BC5CDR (Disease)	NCBI-disease
CHEMDNER	BC5CDR
JNLPBA	BC2GM

- Mention replacement (MR): A method where a random mention from the foundational training set of the same entity type is selected for each mention in an instance.

- Label-wise token replacement (LwTR): In this method, for each word in a sentence, a decision is made randomly on its replacement with any other word from the dataset possessing the same label.

- Synonym replacement (SR): In this approach, each word in a sentence is subject to a binomial distribution decision regarding its replacement with a WordNet [163] synonym.

- Masked Entity Language Modeling (MELM): Here, a RoBERTa pretrained model is used as an MLM to forecast masks within the training dataset. Subsequently, a BERT model is trained on the augmented dataset derived.

- Cross-Domain Named Entity Recognition (style_NER): Knowledge is transferred from a source domain to a target domain for this benchmark. Supplementary data is necessary, and an appropriate dataset with congruent entity types for each target dataset is selected, as detailed in Table 6.3. The results are gauged using a BERT model.

Examples of generated samples per baseline are shown in Table 6.1. For each baseline, we generated one augmented sample per sentence (whenever possible), thus resulting in training datasets at most twice as large as the original.

Table 6.4 compares precision, recall and F1 scores of the different baselines with our method. Results indicate that COSINER, thanks to its effective use of context-based similarities, surpasses baselines in most of the scenarios and datasets. While it always guarantees the highest performance in terms of recall scores, meaning that the system is able to find more entity mentions that are present in the corpus, COSINER performs worse than SR in some scenarios in terms of precision, indicating that the augmentation process generates a higher number of false positives. Recall is crucial in biomedical named entity recognition NER because missing even a single entity can have significant consequences. For example, imagine a physician is using an electronic health record (EHR) system to review a patient's medical history to make a diagnosis. The EHR system relies on NER to identify relevant entities in the patient's medical records, such as their medical conditions, medications, and allergies. If the NER system misses a single entity, such as a patient's allergy to penicillin, this could lead to a prescription of a penicillin-based medication, which could result in a life-threatening allergic reaction. Therefore, it is critical for the NER system to have high recall to ensure that all relevant entities are identified.

The high scores of the SR baseline prove the importance of generating plausible augmented samples when transforming input sentences. Random replacements of MR and LwTR baselines result in too many noisy samples which, in some cases, may even decrease the performance obtained without applying any augmentation method, similarly MELM and style_NER surpass the BERT baseline without augmentation, but fail to overcome the use of a simple BioBERT model.

Effects of increasing the augmented set size

When generating an augmented dataset, the number of augmented samples is generally an important parameter to consider. Hence, the proposed method has been experimented with three different *budgets* for the augmented set: small (100 samples), medium (300 samples) and large (500 samples).

Figure 6.4 shows the results obtained on the three benchmark datasets. As expected, since — thanks to the similarity-based approach — the most informative examples are in the first ranked positions, there is no big difference in using higher budgets.

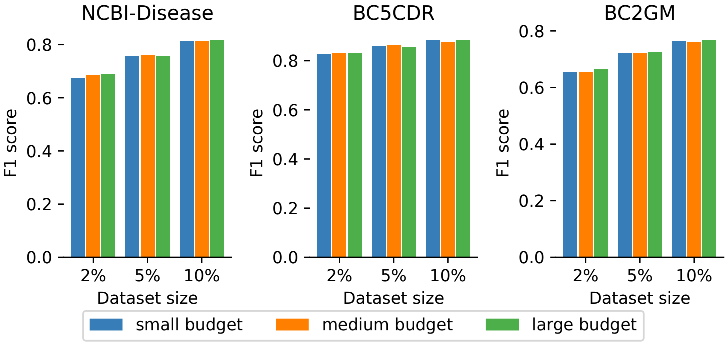


Figure 6.4. Comparative results between the small, medium and large budget of local augmentation strategy with maximum similarity technique

Table 6.4. Comparative results between the local augmentation strategy with maximum similarity technique and baselines.

Dataset size	**Method**	**NCBI-Disease**			**BC5CDR**			**BC2GM**		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
2%	No augmentation	0.430 ±0.193	0.403 ±0.169	0.461 ±0.225	0.628 ±0.179	0.625 ±0.185	0.634 ±0.215	0.510 ±0.036	0.448 ±0.015	0.592 ±0.082
	No augmentation (BioBERT)	0.651 ±0.122	0.619 ±0.100	0.688 ±0.162	0.792 ±0.067	0.799 ±0.058	0.786 ±0.110	0.644 ±0.031	0.600 ±0.057	0.695 ±0.022
	MR	0.666 ±0.084	0.626 ±0.1	0.710 ±0.067	0.813 ±0.032	0.806 ±0.06	0.822 ±0.071	0.640 ±0.02	0.593 ±0.062	0.696 ±0.049
	LwTR	0.677 ±0.101	0.637 ±0.125	0.723 ±0.08	0.828 ±0.019	0.808 ±0.052	0.850 ±0.075	0.642 ±0.037	0.591 ±0.059	0.704 ±0.019
	SR	0.692 ±0.103	**0.649** ±0.132	0.742 ±0.084	0.813 ±0.032	0.811 ±0.085	0.835 ±0.064	0.662 ±0.033	**0.619** ±0.058	0.710 ±0.029
	MELM	0.578 ±0.038	0.545 ±0.046	0.615 ±0.041	0.754 ±0.019	0.719 ±0.047	0.795 ±0.036	0.566 ±0.011	0.504 ±0.006	0.647 ±0.027
	style_NER	0.581 ±0.061	0.537 ±0.076	0.636 ±0.067	0.752 ±0.018	0.713 ±0.041	0.796 ±0.016	0.581 ±0.003	0.540 ±0.018	0.631 ±0.025
	COSINER (ours)	**0.692** ±0.081	0.640 ±0.076	**0.764** ±0.11	**0.832** ±0.022	**0.814** ±0.08	**0.853** ±0.066	**0.665** ±0.038	0.614 ±0.065	**0.724** ±0.025
5%	No augmentation	0.621 ±0.055	0.572 ±0.088	0.68 ±0.054	0.757 ±0.039	0.73 ±0.062	0.788 ±0.121	0.612 ±0.022	0.563 ±0.03	0.671 ±0.077
	No augmentation (BioBERT)	0.735 ±0.041	0.706 ±0.051	0.767 ±0.062	0.850 ±0.02	0.836 ±0.01	0.865 ±0.048	0.711 ±0.012	0.680 ±0.028	0.744 ±0.019
	MR	0.743 ±0.048	0.712 ±0.045	0.776 ±0.059	0.849 ±0.021	0.834 ±0.03	0.865 ±0.026	0.713 ±0.006	0.675 ±0.02	0.755 ±0.024
	LwTR	0.743 ±0.072	0.710 ±0.066	0.780 ±0.086	0.860 ±0.039	**0.846** ±0.017	0.876 ±0.067	0.699 ±0.012	0.660 ±0.024	0.742 ±0.029
	SR	0.758 ±0.044	0.719 ±0.049	0.800 ±0.049	0.858 ±0.03	0.841 ±0.033	0.875 ±0.067	0.719 ±0.011	0.684 ±0.023	0.758 ±0.019
	MELM	0.678 ±0.034	0.647 ±0.037	0.713 ±0.035	0.800 ±0.020	0.769 ±0.043	0.835 ±0.030	0.629 ±0.010	0.587 ±0.010	0.677 ±0.021
	style_NER	0.687 ±0.040	0.662 ±0.038	0.714 ±0.042	0.805 ±0.015	0.793 ±0.020	0.818 ±0.020	0.640 ±0.005	0.594 ±0.018	0.695 ±0.017
	COSINER (ours)	**0.765** ±0.035	**0.733** ±0.039	**0.810** ±0.057	**0.863** ±0.042	0.839 ±0.04	**0.892** ±0.058	**0.726** ±0.022	**0.692** ±0.013	**0.767** ±0.03
10%	No augmentation	0.712 ±0.056	0.670 ±0.065	0.76 ±0.046	0.804 ±0.032	0.781 ±0.046	0.829 ±0.054	0.669 ±0.019	0.626 ±0.026	0.720 ±0.045
	No augmentation (BioBERT)	0.791 ±0.028	0.760 ±0.024	0.825 ±0.036	0.875 ±0.013	0.858 ±0.02	0.892 ±0.028	0.759 ±0.017	0.734 ±0.019	0.786 ±0.016
	MR	0.794 ±0.018	0.761 ±0.025	0.831 ±0.019	0.874 ±0.034	0.859 ±0.038	0.889 ±0.04	0.754 ±0.01	0.724 ±0.013	0.787 ±0.032
	LwTR	0.789 ±0.023	0.756 ±0.034	0.825 ±0.036	0.882 ±0.017	**0.870** ±0.021	0.893 ±0.022	0.741 ±0.012	0.712 ±0.023	0.772 ±0.025
	SR	0.803 ±0.033	0.776 ±0.033	0.832 ±0.053	**0.883** ±0.018	0.862 ±0.016	0.904 ±0.021	0.763 ±0.012	0.738 ±0.019	0.788 ±0.02
	MELM	0.740 ±0.017	0.712 ±0.019	0.770 ±0.016	0.841 ±0.010	0.824 ±0.013	0.858 ±0.019	0.685 ±0.006	0.647 ±0.008	0.728 ±0.010
	style_NER	0.745 ±0.014	0.738 ±0.018	0.752 ±0.014	0.838 ±0.012	0.829 ±0.025	0.847 ±0.021	0.694 ±0.004	0.660 ±0.009	0.732 ±0.010
	COSINER (ours)	**0.816** ±0.066	**0.780** ±0.014	**0.856** ±0.068	0.882 ±0.007	0.861 ±0.022	**0.914** ±0.02	**0.767** ±0.023	**0.738** ±0.026	**0.798** ±0.015

Table 6.5. Comparative results between COSINER techniques with their best budget.

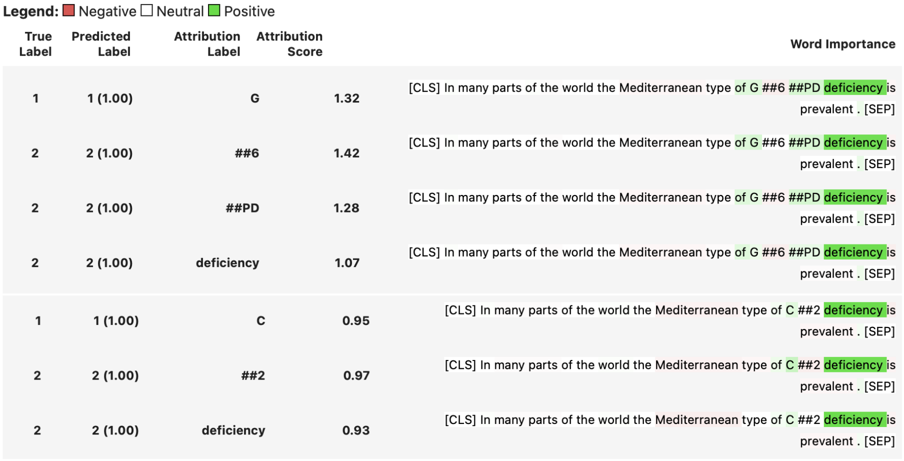
Dataset size	Similarity	Strategy	NCBI Disease	BC5CDR	BC2GM
2%	Maximum	Global	0.688 ± 0.077	0.83 ± 0.023	0.658 ± 0.036
	Minimum	Global	0.683 ± 0.086	0.823 ± 0.032	0.652 ± 0.027
	Maximum	Local	0.689 ± 0.088	**0.832 ±0.022**	**0.665 ±0.038**
	Minimum	Local	**0.692 ±0.081**	0.824 ± 0.015	0.659 ± 0.049
5%	Maximum	Global	**0.765 ±0.035**	0.858 ± 0.023	0.717 ± 0.007
	Minimum	Global	0.756 ± 0.028	0.853 ± 0.029	0.713 ± 0.009
	Maximum	Local	0.76 ± 0.031	**0.863 ±0.042**	**0.726 ±0.022**
	Minimum	Local	0.764 ± 0.041	0.86 ± 0.031	0.714 ± 0.007
10%	Maximum	Global	0.807 ± 0.038	0.88 ± 0.018	0.76 ± 0.02
	Minimum	Global	0.807 ± 0.029	0.873 ± 0.016	0.761 ± 0.012
	Maximum	Local	**0.816 ±0.066**	**0.882 ±0.007**	**0.767 ±0.023**
	Minimum	Local	0.807 ± 0.038	0.876 ± 0.016	0.76 ± 0.009

Effects of parameters for similarity computation and augmented set generation

Table 6.5 shows results obtained with different configurations of parameters for similarity computation (Maximum vs Minimum) and augmented set generation (Local vs Global) discussed in Section 6.2. As expected, the use of Maximum similarity computation leads to higher performance, since augmented samples are plausible and thus nearer to the test distribution. However, the high results obtained with the Minimum configuration show that sometimes it may be beneficial to consider "distant" entities to expand the scope of action of the NER model, especially in strongly limited few-shot settings. With regards to the augmentated set generation, Local criterion is generally better thanks to the augmentation of *all* the sentences in the original dataset.

Efficiency of data augmentation

The execution time required for data augmentation was compared across different baselines and budgets. Results in Table 6.6 indicate that COSINER not only surpasses baselines in terms of NER performance in limited scenarios (2%, 5%), but its computing times are also on par with simpler augmentation methods and better than more complex ones. It should be noted that the execution time is considerably influenced by the



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	1 (1.00)	G	1.32	[CLS] In many parts of the world the Mediterranean type of G ##6 ##PD deficiency is prevalent . [SEP]
2	2 (1.00)	##6	1.42	[CLS] In many parts of the world the Mediterranean type of G ##6 ##PD deficiency is prevalent . [SEP]
2	2 (1.00)	##PD	1.28	[CLS] In many parts of the world the Mediterranean type of G ##6 ##PD deficiency is prevalent . [SEP]
2	2 (1.00)	deficiency	1.07	[CLS] In many parts of the world the Mediterranean type of G ##6 ##PD deficiency is prevalent . [SEP]
1	1 (1.00)	C	0.95	[CLS] In many parts of the world the Mediterranean type of C ##2 deficiency is prevalent . [SEP]
2	2 (1.00)	##2	0.97	[CLS] In many parts of the world the Mediterranean type of C ##2 deficiency is prevalent . [SEP]
2	2 (1.00)	deficiency	0.93	[CLS] In many parts of the world the Mediterranean type of C ##2 deficiency is prevalent . [SEP]

Figure 6.5. Example of XAI tokens analyzed with IG method. In the first column is shown the ground truth, while in the second the predicted label for each token. The third indicates the token that has been examined following a tokenization phase carried out by the model, and the fourth displays the attribution score. The outcome of the IG algorithm applied to the sentence for that particular token is shown in the last column.

size of the training corpus, attributed to the larger entity Lexicon and increased number of pairwise similarities to compute. Additionally, the execution time for generating Lexicon and embeddings was not considered as these are one-time operations that can be executed off-line.

Interpretability of the model on augmented samples

In order to gain insights into the approach, explanations were derived for a sentence from the NCBI-Disease dataset using the Integrated Gradients (IG) method [235] on a model trained with 2% of available data. The sentence "In many parts of the world the Mediterranean type of G6PD deficiency is prevalent" was evaluated, highlighting the entity "G6PD deficiency" and its modified version where the term was replaced with "C2 deficiency." As depicted in Figure 6.5, tokens contributing to the outcome have comparable influence (tokens with positive impact are underlined in green, and those with negative impact in red). The similarity in attribution scores between the two sentences indicates the robustness of the

Table 6.6. Run times (s) for data augmentation with 95% confidence intervals. Comparison with baselines and budgets.

Dataset size	Method	NCBI Disease	BC5CDR	BC2GM
2%	MR	0.123 ±0.020	0.117 ±0.044	0.233 ±0.040
	LwTR	0.149 ±0.067	0.141 ±0.066	0.288 ±0.171
	SR	3.271 ±0.670	3.322 ±0.293	4.374 ±0.643
	MELM	231.4 ±21.355	608.8 ±18.151	468.2 ±29.597
	style_NER	420.2 ±5.916	2199.6 ±24.765	1096.6 ±14.760
	COSINER (small)	0.389 ±0.218	0.445 ±0.192	2.859 ±0.975
	COSINER (medium)	0.44 ±0.472	0.428 ±0.272	2.975 ±1.354
	COSINER (big)	0.529 ±0.491	0.586 ±0.202	3.415 ±1.765
5%	MR	0.212 ±0.065	0.198 ±0.091	0.436 ±0.204
	LwTR	0.287 ±0.171	0.264 ±0.111	0.656 ±0.251
	SR	3.703 ±1.493	4.016 ±0.893	4.494 ±1.137
	MELM	298.4 ±4.613	657.2 ±8.485	600.8 ±17.192
	style_NER	499.2 ±11.394	2327.4 ±32.540	1338 ±9.292
	COSINER (small)	1.541 ±1.002	1.578 ±0.936	15.555 ±5.233
	COSINER (medium)	1.678 ±0.811	1.601 ±1.134	17.257 ±7.973
	COSINER (big)	1.705 ±0.628	1.717 ±0.496	16.711 ±9.581
10%	MR	0.329 ±0.139	0.342 ±0.054	0.846 ±0.316
	LwTR	0.591 ±0.264	0.502 ±0.145	1.206 ±0.528
	SR	4.238 ±1.362	4.233 ±1.174	6.069 ±2.463
	MELM	407.2 ±17.458	759.2 ±37.747	874.6 ±22.291
	style_NER	639.4 ±15.172	4346.2 ±88.666	1854.4 ±30.320
	COSINER (small)	4.286 ±1.305	4.367 ±1.087	60.416 ±19.012
	COSINER (medium)	4.689 ±1.601	4.553 ±1.276	60.508 ±31.661
	COSINER (big)	4.864 ±0.916	4.961 ±1.841	62.386 ±22.487

example, suggesting potential model improvements from such inputs.

6.4 Conclusion & Future Work

In this chapter, a *context similarity*-based methodology was applied to generate plausible augmented data, aiming to enhance the performance of NER tasks. This method mitigates the challenges posed by noisy and mislabeled data that frequently arise with existing techniques.

Experiments conducted in the medical domain (a crucial context for data augmentation) demonstrated the efficacy of this methodology, outpacing several leading baselines while maintaining comparable execution times.

Future directions may explore the integration of this approach with techniques beyond Mention Replacement. Additional tests will be undertaken across varied application contexts and with diverse entity types.

Chapter 7

Learning how to augment data

Although the first attempts of NER data augmentation have shown promising results, the proposed methods of data manipulation may frequently generate a considerable amount of mislabeled and noisy samples, as the new data may not be syntactically and/or semantically accurate. For example, if we manipulate the sentence "*Hypotension* is a term that indicates low blood pressure" so as to replace the entity mention *hypotension* with another disorder (e.g. *dyspnea*, *hypertension*), the resulting augmented sample may be inaccurate and thus mislead the model in effectively identifying mentions.

In the presented chapter, the challenge of selecting highly informative samples from an augmented pool is tackled. Taking cues from policy-based active learning [58], a fixed heuristic is eschewed in favor of permitting the outlined framework to learn active data selection. This is achieved by framing the selection task as a reinforcement learning challenge. Specifically, for each instance within the augmented pool, a decision is to be made by an agent regarding its selection, grounded on the sample's attributes and model outcomes. The selection strategy is honed through the employment of a deep Q-network [169].

The method is experimented by simulating few-shot scenarios in BioNER applications, namely utilizing only k samples as the training data, with $k \in \{10, 50, 100\}$. Under such configurations, the framework demonstrates its capability to prioritize the selection of the most informative augmented samples, exhibiting encouraging results as evidenced by the comparison with the selected baselines. This approach introduces a novel avenue for investigating the potential of data augmentation in enhancing the performance of NER models amidst limited training data availability. The findings suggest a significant scope for improvement, given that the data augmentation technique deployed for generating the augmented pool could be substituted with more sophisticated and effective methods.

The remainder of this chapter is organized as follows: Section 7.1 summarizes the literature on NER data augmentation. Section 7.2 presents the augmentation framework, and Section 7.3 reports the experimental results. The chapter concludes in Section 7.4. The material in this chapter is based on the article "*Learning How To Augment Data: An Application To Biomedical NER*" [173], presented at the 6th International Workshop on Knowledge Discovery from Healthcare Data, co-located with the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023).

7.1 Related Work

Data augmentation aims to increase the amount of available training data by means of data manipulations, heuristics or external data sources. Dai and Adel [45] investigate the improvements in performance obtained by augmenting NER data with simple data manipulations, such as token replacements, mention replacements, and shuffling. However, these approaches may generate too many noisy samples which may in turn hinder the ability of the model to be effectively trained. Bartolini et al. [16] address this challenge by replacing entity mentions with the most similar entities retrieved by computing context-base similarity. Zeng et al. [280] address the poor generalization ability of few-shot systems to spurious correlations between an entity mention and its context by generating counterfactual examples. Chen et al. [34] leverage an external high-resource corpus to learn how to imitate language patterns (e.g. style, noise, abbreviations). All these works do not evaluate the impact of the noise produced by their proposed augmentation approach over models' performance.

The approach for selecting less noisy samples from an augmented pool is grounded in policy-based active learning [58]. Active learning (AL) is a recognized method for selecting highly informative unlabeled data for annotation, which in turn aids in training an optimized classifier, thereby enhancing the efficiency of human efforts. The strategies employed in AL are heuristic-based: for instance, uncertainty sampling [125, 243] operates by selecting data contingent on the uncertainty reflected in the model's outputs, while Seung et al. [218] opt for data based on the disagreement within a committee. In a different vein, Fang et al. [58] reconceptualize AL as a reinforcement learning challenge, wherein an intelligent agent employs a deep Q-network [169] to autonomously learn a selection strategy. In this endeavor, an intelligent agent autonomously acquires a policy for identifying the most beneficial samples from an augmented dataset, to enhance the overall model performance. Consequently, the agent is capable of selecting samples with a lower likelihood of misleading the model, eliminating the necessity for human input or guidance.

7.2 Methodology

The methodological workflow of the proposed framework is illustrated in Figure 7.1. In this section, we will first provide a formalization of the few-shot BioNER problem, and then describe each module in-depth, from the generation of a concepts vocabulary to the reinforcement learning cycles.

7.2.1 Problem formulation

The input of a NER system is a sentence $\mathbf{s}$, which can be represented as a sequence of tokens $\mathbf{s} = [t_1, t_2, \ldots, t_N]$. NER outputs a list of tuples $[I_s, I_e, t]$ representing named entities mentioned in $\mathbf{s}$. Here, $I_s \in [1, N]$ and $I_e \in [1, N]$ are the indexes of start and end characters of the named entity mention, while t is the entity type [136].

In practice, this task is usually accomplished by producing a paired sequence of categorical values $\mathbf{y} = [y_1, y_2, \ldots, y_N]$ as the output of the NER model, where $y_i \in \mathscr{Y}$ indicates the entity type of the i-th token. Hence, a NER dataset is defined as a collection of pairwise data $\mathscr{D} = \{(\mathbf{s}_i, \mathbf{y}_i)\}_{i=1}^{K}$, K being the number of examples.

For the purposes of this work, the IOB scheme will be used to identify entity mentions. Under this scheme, each input token is mapped to the beginning (B), inside (I) or outside (O) of an entity mention. Furthermore,

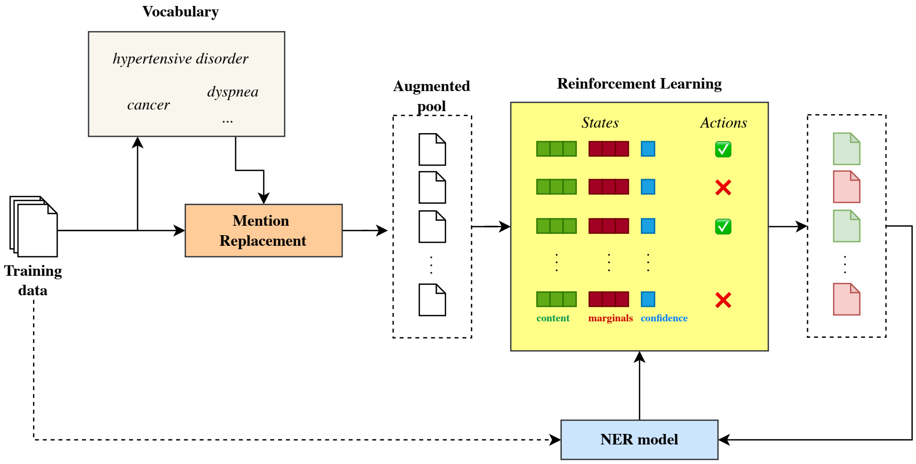


Figure 7.1. Methodological workflow for the augmentation of BioNER datasets. First, we collect the entity mentions occurring in training data, thus building a concepts *vocabulary*, which is then used to generate an *augmented pool* of data samples with a simple data augmentation technique named *mention replacement*. A deep Q-learning based approach iteratively assigns a state to each sample in the augmented pool and decides whether to select it or not to re-train the NER model according to a policy that is updated at each cycle based on a *reward* that measures the extent to which the addition of the new samples improves the quality of the model.

we will consider inputs from biomedical domains, where the NER task is known as *Biomedical NER (BioNER)*. Due to the data scarcity that usually affects such domains, the system will be tested in few-shot settings, i.e. the number of training instances K is small (e.g. $K \in \{10, 50, 100\}$).

7.2.2 Generation of a vocabulary of concepts

Based on the available training data, all entity mentions are extracted, thus building a vocabulary of concepts. In this work, the framework is tested by relying solely on the input training data, but this module can be easily extended to include concepts from biomedical ontologies or guided by domain experts. For example, physicians usually possess knowledge regarding the representation of medical concepts in clinical notes; hence, if

Input	Output
If untreated, **hemochromatosis** can cause serious illness and early death, but the disease is still substantially under-diagnosed.	If untreated, **mononucleosis** can cause serious illness and early death, but the disease is still substantially under-diagnosed.
When expressed in Escherichia coli, **SH-PTP2** displays tyrosine-specific phosphatase activity	When expressed in Escherichia coli, **PTPN6** displays tyrosine-specific phosphatase activity

Table 7.1. Examples of data augmentation via mention replacement. Here, entity mentions are reported in bold.

interested in recognizing mentions of a particular concept, a set of aliases can be provided to effectively augment the original training set.

7.2.3 Data augmentation via mention replacement

In each sentence of the training set, the determination of whether a mention should be replaced is made using a binomial distribution. If the outcome is affirmative, a replacement mention is selected from the concepts vocabulary. Subsequently, the corresponding IOB-label sequence is modified as needed. Examples of mention replacement are provided in Table 7.1.

The reason behind the choice of this augmentation technique lies in the high number of noisy samples it may generate, given the random nature of the mention replacement. This allows us effectively test the ability of our framework to discard samples that may mislead the model. However, the performance of the framework can be further improved with more sophisticated augmentation methods, e.g. based on context similarity [16] or learning patterns from cross-domain data [34].

7.2.4 Reinforcement learning cycles

We learn how to select data from the augmented pool with a module based on reinforcement learning. Our method is built upon the foundations of *Policy-based Active Learning* [58], which has been previously demonstrated to be capable of automatically learning an active learning strategy from data by formulating the active learning as a reinforcement learning problem where the *state* corresponds to the unlabeled data selected for labeling, and their label, and the *action* is the selection heuristic. Specifically, we adapt the method not to work with unlabeled data and human oracles, but with the augmented pool generated in the previous step, and to learn the best strategy to select the samples that may mostly benefit the performance of the NER model. Furthermore, while Fang et al. [58] make a streaming assumption, i.e. unlabelled data arrive one by one and the agent decides the action to take, we assume batch-based learning where the augmented pool is entirely available and the *reward* is computed on the set of actions that the agent has decided to take on the whole dataset in the i-th cycle.

In the remainder of this section, an in-depth details on the components of the reinforcement learning process is provided.

States

The representation of the state of each sentence $\mathbf{s}$ in the augmented pool at time i is determined by considering both an embedded representation of its content and the outputs of the NER model Θ_i trained over the selected data at time i. Specifically, the state s_k is defined as the concatenation of the three representations described as follows: content, marginals, and confidence. The set of states at time i is denoted by $\mathscr{S}_i$.

Content Adopting the approach delineated by Kim [105], each of the N tokens t_i in a given sentence is initially encoded, resulting in a matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. Subsequently, a convolutional neural network is applied, encompassing a series of filters executing linear transformations followed by ReLU activation functions. The terminal layer of the network undertakes a max-pooling operation, furnishing the representation of the sentence content $\mathbf{h_c}$.

Marginals Let $p_{\Theta_i}(\mathbf{y}|\mathbf{s})$ indicate the prediction outputs of the NER model given the input sentence $\mathbf{s}$. Another convolutional neural network is used to represent the predictive marginals, i.e. the probability distributions associated to all the tokens in $\mathbf{s}$. Following Fang et al. [58], the convolutional layer contains j filters activated with ReLU applied with a window width of 3 and height equal to the number of classes (3 in our case, i.e. I, O and B). Padding is used to endure a wide convolution, and mean pooling is used to allow the network to effectively capture the average uncertainty in each window. The final hidden layer outputs the representation of predictive marginals $\mathbf{h}_m$.

Confidence In accordance with Fang et al. [58], the confidence is represented by calculating the probability of the most probable sequence of labels as per the model, expressed as $C = \sqrt[n]{\max_{\mathbf{y}} p_{\Theta_i}(\mathbf{y}|\mathbf{s})}$, where n denotes the length of the sentence $\mathbf{s}$.

Actions

Given the state of each input sample, an agent has to decide whether to select it or not to re-train the NER model. Thus, for each sentence $\mathbf{s}_k$ in the augmented pool, the agent selects either to use it ($a_k = 1$) or not ($a_k = 0$). We denote the set of actions made at time i with $\mathscr{A}_i$.

Reward

The reward provides a feedback on the quality of the decisions made by the agent. At each step i, the reward is defined as the change in held-out performance:

$$\mathscr{R}_i(\mathscr{S}_{i-1}, \mathscr{A}_i) = \text{Performance}(\Theta_i) - \text{Performance}(\Theta_{i-1}), \tag{7.1}$$

where Performance$(\cdot)$ is a measure of the model's quality. In the conducted work, the F1 score is computed for the purpose of determining rewards. It is noteworthy that the value of $\mathscr{R}_i$ may be negative, indicating that the actions executed by the agent potentially exert a detrimental impact on the performance.

Deep Q-Network

A deep Q-learning [169] approach is adopted where the utility of choosing the action a_k from state s_k is evaluated by the Q function $\mathscr{Q}^{\pi}(s_k, a_k)$ according to the policy π. The Q-function is iteratively updated by the agent by considering the rewards obtained in each episode.

The deep Q-network (DQN) consists in a single hidden layer which takes the state vector of a single instance $s_k = [\mathbf{h}_c, \mathbf{h}_m, C]$ as input and uses a ReLU activation function to output two scalar values $\mathscr{Q}(s_k, a_k)$ associated to the two possible actions $a_k \in \{0, 1\}$.

The training objective is to minimize the difference between the estimated $\mathscr{Q}$-value and the true $\mathscr{Q}$-value for a given state-action pair. This is typically done by using a variant of the Q-learning algorithm known as the Bellman equation, which recursively defines the $\mathscr{Q}$-value for a state-action pair as the immediate reward plus the discounted future $\mathscr{Q}$-value for the next state-action pair. Mathematically, this can be expressed as:

$$\mathscr{Q}(s_i, a_i) = \mathbb{E}[r_i + \gamma \cdot \max_{a_{i+1}} Q(s_{i+1}, a_{i+1})], \tag{7.2}$$

where $\mathscr{Q}(s, a)$ is the $\mathscr{Q}$-value for state s and action a, $\gamma \in [0, 1]$ is the discount factor, and $\max_{a_{i+1}} Q(s_{i+1}, a_{i+1})$ is the maximum $\mathscr{Q}$-value over all possible actions in the next state.

The goal of the Q-learning algorithm is to update the Q-network weights θ to minimize the mean squared error between the estimated $\mathscr{Q}$-value $\mathscr{Q}(s, a; \theta)$ and the target $\mathscr{Q}$-value y:

$$\mathscr{L}(\theta) = \mathbb{E}\big[\big(y_i(r_i, s_{i+1}) - \mathscr{Q}(s_i, a_i; \theta)\big)^2\big], \tag{7.3}$$

where $y_i(r_i, s_{i+1}) = r_i + \gamma \cdot \max_{a_{i+1}} Q(s_{i+1}, a_{i+1}; \theta_{i-1})$ is the target $\mathscr{Q}$-value based on the current parameters θ_{i-1}, and results are averaged over a minibatch of samples. Learning updates are based on stochastic gradient descent.

7.3 Experiments

In this section, a comprehensive elucidation of the experiments conducted to evaluate the system's performance is presented. Initially, the experimental setup is delineated in Section 7.3.1, followed by a discourse on the experimental outcomes in Section 7.3.2.

7.3.1 Experimental setup

Datasets

The methodology is assessed on three widely recognized benchmark datasets from the biomedical domain, described as follows:

- *NCBI-Disease* [54]: 793 abstracts from PubMed, annotated with disorders entity mentions.
- *BC2GM* [226]: over 20,000 abstracts from PubMed annotated with gene mentions.
- *BC5CDR* [134]: over 1,500 abstracts from PubMed annotated with diseases and chemicals. For simplicity, we consider only chemical entity mentions in our experiments.

For each dataset, the original training, validation and test sets provided with their original release have been considered.

Few-shot simulations In order to emulate data scarcity conditions, a random sample of k sentences is taken from the training set, where $k \in \{10, 50, 100\}$. Given the potential variability in performance due to different training sample selections, each experiment is conducted 5 times and results are reported as an average.

Training details

Given the data-limited aspect of the work, it is assumed that data for tuning hyper-parameters is unavailable. Despite this, models are tested on the entire test set. Hyperparameters are selected based on prior work and practical considerations. Particularly, a pre-trained biomedical Transformer network [124] is utilized, and all models are trained for 3 `epochs` with a `learning rate` of $2 \cdot 10^{-4}$, utilizing an AdamW `optimizer`, a `batch size` of 5, and a `maximum sequence length` of 256. The reinforcement learning framework is executed for 5 episodes. Model quality is assessed based on the precision, recall, and F1 scores acquired with the `seqeval`[1] Python library.

Hardware configuration

All experiments were conducted on the platform Google Colab, using the Free tier plan, which provides a virtual machine with an NVIDIA T4 GPU with 16GB of RAM, an Intel® Xeon® processor with a frequency of 2.3GHz and 10 cores (but only one used by the VM instance), 12 GB of available memory and 78.19 GB of free disk. Due to the limits imposed by Colab's free plan, we were unable to pursue further improvements on the obtained results. Specifically, we could only run a maximum of 5 PAL episodes per experiment. Although this was sufficient to achieve the intended goals, conducting a greater number of episodes could have allowed for a more refined selection policy for the augmented instances, potentially leading to improved model performances.

Baselines

The proposed method has been compared with the two baselines for the selection of samples from the augmented pool described as follows:

- *Random*: a random set of instances is sampled from the augmented pool.

- *Uncertainty*: uncertainty-based active learning [125] is leveraged as an heuristic-based framework for the selection of samples, i.e. we rank augmented samples according to the uncertainty of the model in its predictions. Since model predictions are mapped to each token in the sentence, we aggregate them to obtain a single ranking value.

For each method, the initial model is always pre-trained with the available training data in the simulated few-shot scenario. Subsequently, the performance of that model is assessed when fine-tuned with the selected samples.

[1] https://github.com/chakki-works/seqeval

7.3.2 Results

The results in Table 7.2 offer a comprehensive comparison of various baselines performance across different k-shot scenarios and datasets for BioNER tasks. It is evident that the proposed method exhibits consistent and competitive results, notably securing top F1 scores in several instances. This signifies the method's capability in accurately pinpointing entity mentions. A notable enhancement over the random baseline is evident in the 10-shot scenario on the BC2GM dataset. Given the dataset's emphasis on the *gene* entity type, which can be expressed diversely (examples include mStaf gene, OBP, primase, V8 protease, MT), random mention replacements can introduce significant noise. As data becomes scarcer, this noise detrimentally impacts results. Yet, the inclusion of clean training data can mitigate this. Moreover, the presented method either surpasses or matches the top performers in precision, indicative of a dependable selection policy that curbs false positives.

Results indicate significant variations in standard deviations, linking model quality to the sampling of the few-shot training set.

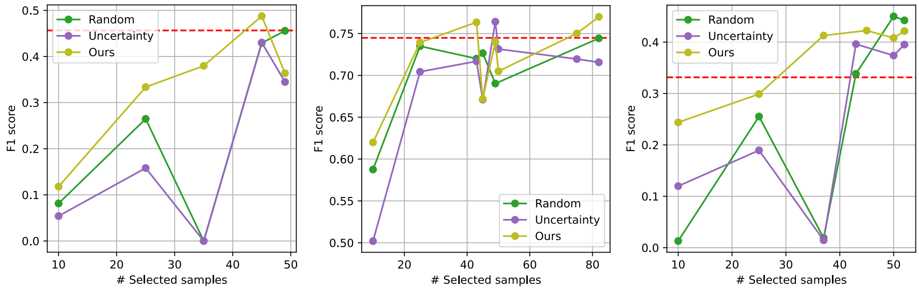
Figure 7.2 depicts F1 score performance trends when increasing the amount of data selected from the augmented pool in a 50-shots context. The method presented here consistently outperforms both random and uncertainty-based selection techniques. Notably, the performance advantage is more pronounced when fewer elements are selected. However, as the selection quantity grows, the differences between the curves become minimal. This observation aligns with the study by Fang et al. [58], suggesting that Policy-based Active Learning is more efficient with fewer selected elements. In each of the plots, the peak performance of the presented method surpasses that of a model without augmented samples, denoted by the dashed red line.

7.4 Conclusion & Future Work

The chapter introduces a new methodology for selecting pertinent samples from an enhanced pool, aiming to enhance Named Entity Recognition (NER) model outcomes in the biomedical sector when only limited training data is available. This method incorporates policy-based active learning

Shots	Dataset	Method	Precision	Recall	F1
10	NCBI-Disease	Random	11.94 ± 14.87	4.81 ± 9.66	6.30 ± 12.08
		Uncertainty	20.01 ± 9.16	**36.62 ± 20.58**	**25.05 ± 13.07**
		Ours	**21.33 ± 6.82**	26.31 ± 17.06	21.58 ± 13.77
	BC2GM	Random	7.03 ± 9.34	0.33 ± 0.35	1.02 ± 0.45
		Uncertainty	8.47 ± 7.95	25.38 ± 23.28	12.68 ± 11.82
		Ours	**21.32 ± 5.18**	**30.85 ± 21.05**	**23.11 ± 12.11**
	BC5CDR	Random	**79.27 ± 13.40**	49.84 ± 25.78	55.80 ± 18.26
		Uncertainty	46.91 ± 28.45	50.22 ± 45.76	39.14 ± 35.96
		Ours	62.15 ± 11.94	**74.99 ± 9.64**	**66.71 ± 6.92**
50	NCBI-Disease	Random	**30.47 ± 17.32**	41.52 ± 24.53	**34.82 ± 19.75**
		Uncertainty	25.24 ± 15.23	**47.42 ± 27.48**	32.26 ± 18.37
		Ours	29.16 ± 18.41	45.06 ± 27.31	34.37 ± 20.28
	BC2GM	Random	26.79 ± 13.92	40.17 ± 23.47	31.91 ± 17.58
		Uncertainty	25.17 ± 7.63	49.26 ± 27.83	31.28 ± 16.75
		Ours	**32.23 ± 2.41**	**52.70 ± 13.35**	**39.68 ± 5.90**
	BC5CDR	Random	62.02 ± 4.62	**86.39 ± 4.35**	72.02 ± 2.44
		Uncertainty	61.89 ± 5.27	85.59 ± 3.64	71.69 ± 3.32
		Ours	**67.26 ± 7.16**	83.38 ± 3.60	**74.18 ± 4.12**
100	NCBI-Disease	Random	49.74 ± 4.01	68.44 ± 4.09	57.45 ± 2.26
		Uncertainty	**50.66 ± 2.84**	69.37 ± 5.36	58.39 ± 1.65
		Ours	50.25 ± 8.34	**72.90 ± 4.59**	**58.92 ± 4.34**
	BC2GM	Random	**37.88 ± 3.02**	62.90 ± 3.46	**47.27 ± 3.30**
		Uncertainty	35.82 ± 1.28	**62.91 ± 6.38**	45.60 ± 2.64
		Ours	37.02 ± 2.88	62.04 ± 7.52	46.33 ± 4.22
	BC5CDR	Random	**67.19 ± 4.35**	88.65 ± 1.47	**76.36 ± 2.41**
		Uncertainty	60.46 ± 3.20	**90.44 ± 1.42**	72.45 ± 2.72
		Ours	64.53 ± 3.48	87.82 ± 4.54	74.28 ± 1.88

Table 7.2. Average results on the benchmark BioNER datasets in different k-shot scenarios, $k \in \{10, 50, 100\}$. For each method and score, we report the mean μ and standard deviation σ obtained across 5 repetitions, in the format $\mu \pm \sigma$. Results with the highest mean are reported in bold.



(a) NCBI-Disease **(b)** BC5CDR **(c)** BC2GM

Figure 7.2. Performance trends as the number of selected samples increases. The horizontal dashed red line is the performance of the original model without augmented samples. These results have been obtained in the 50-shots scenario.

[58] to craft a policy for pinpointing the most significant enhanced samples, which bolsters the NER model's capacity to generalize.

Assessments of this technique in simulated few-shot settings within BioNER tasks indicate its prowess in prioritizing the most significant enhanced samples. The results achieved are notable when juxtaposed with established benchmarks. This strategy sheds light on the untapped potential of data augmentation in amplifying the performance of NER models, especially in niches like the biomedical realm where labeled datasets are limited and domain expertise is paramount.

It would be beneficial for upcoming studies to gauge the resilience of this structure when applied to tangible biomedical datasets and to probe into the impact of diverse data enhancement methods on the technique effectiveness. While the current design of this structure utilizes a straightforward *mention replacement* method, it could be supplanted by more intricate techniques.

Additionally, there is potential in adapting this technique for other linguistic processing tasks outside of NER, such as the extraction of relationships and entity linkage. It would be insightful to contrast its efficacy with that of present leading-edge techniques. An exploration into the transparency of the deduced selection strategy could also provide a deeper understanding of the vital features of enhanced data samples that bolster the NER model's outcomes.

Chapter 8

Multi-task Biomedical Named Entity Recognition with Knowledge Distillation

Numerous industrial sectors, including healthcare, are being revolutionised by the uncontrolled growth of data produced by humans and machines as well as the availability of computing resources and algorithms able to handle and analyse it. In 2022, PubMed Central[1] provides open online access to 7.8 million full-text articles. Concomitantly, efforts are being made to collect and make available the unstructured health information associated with hospital admissions (e.g. EHRs, laboratory tests, medications). As a result, the field of biomedical text understanding can profitably benefit from the current advancements in Deep Learning and Natural Language Processing techniques.

Biomedical Named Entity Recognition (BioNER) consists in identifying mentions of biomedical entities (e.g. disorders, chemical compounds, genetic information) from unstructured text data. It is the first and essential step of many text understanding applications, such as the construction of knowledge graphs for data representation and analysis or conversational agents including research assistants and medical chatbots.

It is extremely difficult to develop a BioNER system that can recognise a wide range of entity types with high precision and recall for a number of

[1] https://www.ncbi.nlm.nih.gov/pmc/

Table 8.1. Example of TaughtNet output for the identification of disease, chemical and gene mentions, compared to the ground truth.Sample taken from the test set of the dataset *NCBI-disease*

Ground truth	Cycloheximide facilitates the identification of aberrant transcripts resulting from a novel splice-site mutation in COL17A1 in a patient with generalized athrophic benign epidemolysis bullosa.
TaughtNet	Cycloheximide facilitates the identification of aberrant transcripts resulting from a novel splice-site mutation in COL17A1 in a patient with generalized athrophic benign epidemolysis bullosa.

reasons, including:

- *Presence of synonyms, alternate spellings, polysemous words.* Biomedical datasets are characterized by a large number of synonyms or alternate spellings of entities, which are often referred to with non-standard abbreviations; polysemy is very common, i.e. the same token could represent different entities based on its context (e.g. the token "VHL" may refer to the Von Hippel-Lindau disease or to the gene name which causes the disease).

- *Lack of annotated data.* To guarantee high quality, the labeling process of healthcare datasets requires time, effort and domain knowledge. As a result, there is a lack of publicly available training data. Furthermore, the majority of datasets covers only one or two entity types, making it necessary to integrate different data sources.

- *Inference time and memory constraints.* Being (usually) a component of a larger pipeline architecture, the BioNER system has to be able to promptly provide its results when required. Moreover, in conversational agents, it may be necessary to deploy the system on devices with a limited amount of memory.

NER systems for biological text mining were used to be primarily dictionary- and rule-based, but they had a number of issues, including the *out-of-vocabulary* problem, i.e. they struggled to deal with unseen and/or polysemous words and had a low recall.

As a result of the availability of an increasing number of human-labeled datasets, BioNER systems evolved over time by means of deep learning techniques able to infer features from sentence contexts. These methods were typically based on Bidirectional Long-Short Term Memory networks with Conditional Random Fields (BiLSTM-CRF) [120, 204] and/or trying to capture character-level features of words [158, 37, 74, 257]. Recently, large-scale language models pre-trained on biomedical corpora and fine-tuned over BioNER datasets [18, 124, 92, 10, 73] have shown their remarkable potential to enhance the state-of-the-art of biomedical entity recognition and their promising prospects for improvement as the availability of training data increases [127].

Nevertheless, the above-mentioned models usually have hundreds of millions of parameters, and recent research demonstrates that as the training parameters are increased, performance on downstream tasks improves [25]. The expansion of model parameters implies computational and memory limitations, which may make it more difficult to use these systems in real-world settings.

This study seeks to harness the technological advances of Transformer models to develop a multi-task BioNER system capable of recognizing multiple entity types from its inputs, addressing the data shortage in the biomedical domain. Notably, many publicly accessible datasets in this domain are limited to tags for a singular entity type. The notion of creating, training, and implementing a distinct Transformer-based BioNER model for each dataset available is deemed impractical, primarily due to their extensive memory needs and potential issues stemming from overlapping predictions, such as a single mention receiving different entity type assignments from multiple models. Thus, this research introduces *TaughtNet*, a multi-task framework rooted in knowledge distillation, designed to refine a singular transformer architecture for the identification of various entity types (a sample output is illustrated in Table 8.1).

Contemporary studies by Khan et al. [101] and Yoon et al. [277] advocate for modifications in model architecture and training processes to achieve similar outcomes: the former integrates multiple models with shared layers to establish a "shared knowledge" across datasets, whereas the latter employs an ensemble of single-task models. Contrarily, *TaughtNet* yields a standalone Student Transformer model proficient in discerning an array of entity types. Within this framework, the Teacher models are not amalgamated into an ensemble for concurrent prediction. Instead, they function solely to transfer their expertise to the Student during the training phase.

Experimental results indicate that *TaughtNet* not only provides an efficient means of distinguishing between multiple entity types, maintaining peak performance across three benchmark datasets, but also exhibits adaptability to more compact and nimble Student models. Such adaptability proves advantageous in real-world applications, particularly where deployment on hardware with memory constraints or swift inferences are of concern. Moreover, the study highlights *TaughtNet*'s inherent capability to elucidate its prediction rationale, a feat often unattainable with the use of multiple models or intricate architectural alterations.

The subsequent sections are organized as follows: Section 2 revisits foundational concepts such as Biomedical Named Entity Recognition, Pretrained Language Models, Multitask Learning, Knowledge Distillation, and outlines prominent Related Works. The *TaughtNet* training framework is elaborated upon in Section 3, followed by a detailed account of experiments in Section 4. Conclusions and prospective research directions are discussed in Section 5.

This study is grounded in the article titled "*TaughtNet: Learning Multi-Task Biomedical Named Entity Recognition From Single-Task Teachers*" [172], featured in the IEEE Journal on Biomedical and Health Informatics (J-BHI).

8.1 Background

The focus of this paper is not to pretrain a novel language model, but rather to design a fine-tuning framework which, based on knowledge distillation, allows us to accomplish the NER task for multiple entities by exploiting pretrained language models and heterogeneous publicly available healthcare datasets, each of them referring to a different entity type.

8.1.1 Multi-Task Learning

Multi-Task Learning (MTL) aims to leverage multiple datasets that are similar to one another yet address various tasks [30]. The key idea is that the knowledge acquired by the model for solving a task (e.g., disease extraction) can help it in solving similar tasks (e.g. drug extraction).

In biomedical text mining, the first approaches (e.g. [42]) ignored the information of subwords which can be crucial to obtain high performance. Wang et al. [257] propose the combination of a multi-task BiLSTM-CRF model and a BiLSTM layer for modeling character sequences, obtaining promising results. To the best of our knowledge, [101] is the first work adopting the multi-task learning framework with a pre-trained language model.

Yoon et al. [277] highlight that despite the high recall obtained by MTL models, their precision is relatively low, i.e. they have difficulties in differentiating between entity types, primarily due to the presence of polysemous words in text which confuse the model. To solve such false-positive problem, the authors propose *CollaboNet*, a network composed of multiple models, each one built on a different dataset for a different task, which collaborates during training and inferences to output the final prediction. Despite the promising results, this framework requires "collaborator" models to be stored in memory at inference time and to provide their outputs when a prediction is required, resulting in low efficiency in computational and memory consumption terms.

To address the challenges of low precision and high computational and memory consumption, a training framework called *TaughtNet* was developed, drawing inspiration from CollaboNet. This framework facilitates the fine-tuning of a single transformer language model for multi-task BioNER using Knowledge Distillation. Put simply, single-task models are trained on distinct datasets. Instead of collaborating directly on prediction outputs, these models "teach" a central multi-task "student" model to predict entity types in their respective areas of expertise.

8.1.2 Knowledge Distillation

Knowledge Distillation (KD) has been originally proposed in [81] as a *teacher-student* framework which allows the knowledge embedded in a large "teacher" model to be shared with its small "student". Modeling the behavior of teacher and student with functions $f_T(\cdot)$ and $f_S(\cdot)$, respectively, the objective of KD is to minimize the following objective function:

$$\mathscr{L} = \sum_{x \in \mathscr{X}} L\Big(f_S(x), f_T(x)\Big), \tag{8.1}$$

where $\mathscr{X}$ is the training dataset and $L(\cdot)$ denotes the loss function computing the difference between the two behavior function outputs for the input $x \in \mathscr{X}$.

With the primary aim to "compress" the knowledge embedded in a large model — which shows good performance but is too large to be used in real scenarios — into a smaller one, the application of KG in NLP and pre-trained models has been extensively studied [106, 90, 207, 232, 234, 255, 97].

Research on the application of the KD framework for purposes other than model compression is restricted to a few works. Reimers et al. [201] try to transfer the knowledge embedded in an English BERT model to the German language. In [36], a fine-tuned BERT teacher is used as extra supervision to improve the text generation performance of conventional Seq2Seq student models.

TaughtNet is the first approach exploiting KD in a NER scenario to transfer the knowledge encoded in a variety of teachers, specialized in single entity types, into a single student, which learns to recognize all the entity types.

The *multi-teacher* scenario in the application of the KD approach has been thoroughly investigated [31, 65, 12, 152]. Fukuda et al. [65] hypothesize that the different "views" provided by various teacher distributions may help the student generalizing better while also capturing the complementary information embedded in each teacher stream. In [31], the teacher is an ensemble of models whose outputs are determined by the combination of the individual model predictions and the student learns to imitate its behavior by minimizing the Kullback-Leibler (KL) divergence [118] between student and teacher distributions (which the authors prove to be equal to minimizing the cross-entropy error between the two distributions). The use of an ensemble knowledge distillation framework in [12] results in better student accuracy thanks to the encouragement of heterogeneity in feature learning. [152] highlights the importance of assigning the proper weights to teachers when distilling their knowledge.

In contrast to traditional KD approaches, where teachers and students share the same tasks, this work aims to design a student able to handle all the tasks learned from teachers in a single model. Tan et al. [238] propose a similar approach, designing a multilingual translation system based on knowledge distillation from multiple individual teachers handling separate language pairs. Their experimental results, showing that the multilingual model reaches comparable performance with teachers — even outperforming them in many cases — further encourage our work.

8.2 Methodology

In this study, a set of publicly available healthcare datasets is utilized to develop a multi-task BioNER model. An extensive overview of the framework can be found in Figure 8.1. For clarity, the notation used throughout is summarized in Table 8.2, and each methodological step is exemplified.

8.2.1 Problem Formulation

Let $\mathbb{E}$ the set of entity types to be individuated, $e_i \in \mathbb{E}$ representing the i-th entity type (with $i \in \{1, \ldots, |\mathbb{E}|\}$). A corpus of annotated sentences $\mathscr{D}_i$ is associated to each entity type, $\mathscr{D}_i = \{(\mathbf{x}, \mathbf{y}) \in \mathscr{X}_i \times \mathscr{Y}_i\}$, $\mathscr{X}_i$ being the set of sentences $\mathbf{x}$ (sequences of tokens $x_j \in \mathbf{x}$, $j \in \{1, \ldots, \mathrm{H}\}$, where H represents the maximum sequence length) and $\mathscr{Y}_i$ being the relative set of labels. In this work, we will refer to the IOB2 annotation schema [199], assigning the "B" label to the *beginning*, the "I" label to the *inside* and the "O" to the *outside* of an entity mention.

Based on such datasets, our aim is to learn a model $f(\cdot)$ able to map each token x_j in a sentence $\mathbf{x}$ to its label $y_j \in \mathscr{Y}^{multi}$, where:

$$\mathscr{Y}^{multi} = \{B\text{-}e_1, I\text{-}e_1, B\text{-}e_2, I\text{-}e_2, \ldots, B\text{-}e_{|\mathbb{E}|}, I\text{-}e_{|\mathbb{E}|}, O\} \tag{8.2}$$

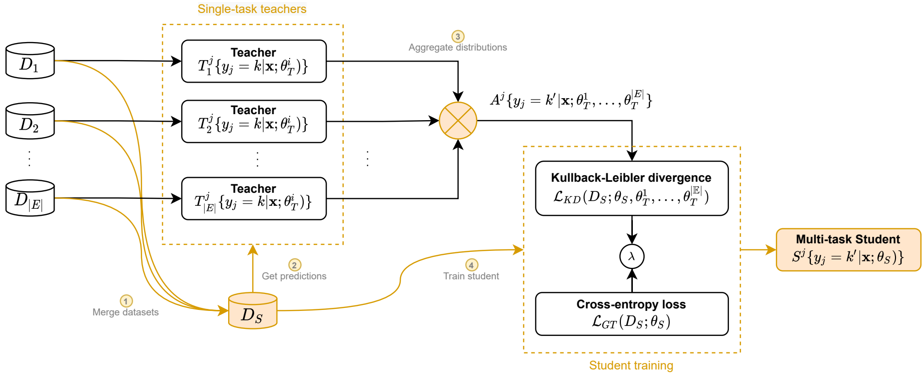

Single-task teachers
D_1
D_2
$D_{|E|}$
Teacher $T^j_1\{y_j = k|\mathbf{x};\theta^i_T)\}$
Teacher $T^j_2\{y_j = k|\mathbf{x};\theta^i_T)\}$
Teacher $T^j_{|E|}\{y_j = k|\mathbf{x};\theta^i_T)\}$
3 Aggregate distributions
$A^j\{y_j = k'|\mathbf{x};\theta^1_T,\ldots,\theta^{|E|}_T\}$
Kullback-Leibler divergence $\mathcal{L}_{KD}(D_S;\theta_S,\theta^1_T,\ldots,\theta^{|\mathbb{E}|}_T)$
λ
Cross-entropy loss $\mathcal{L}_{GT}(D_S;\theta_S)$
Multi-task Student $S^j\{y_j = k'|\mathbf{x};\theta_S)\}$
2 Get predictions
4 Train student
1 Merge datasets
D_S
Student training

Figure 8.1. Overview of *TaughtNet* training framework. (1) First, single-entity datasets are merged together to build a multi-entity corpus $\mathscr{D}_S$ used as a ground truth reference during the training of the Student; (2) then, each Teacher provides its predictions for each sample in $\mathscr{D}_S$ and (3) their output distributions are aggregated so as to build the corpus $\mathscr{A}$ used to distill the knowledge from Teachers. (4) Finally, the Student is trained to minimize two loss components referring to Teachers' knowledge and ground truth, respectively.

Running example

The set of entity types in our running example is $\mathbb{E} = \{e_1, e_2, e_3\} = \{disease, gene, drug\}$. Hence, the model trained with TaughtNET will learn how to predict one of the labels reported below to each token of input samples:

$$\mathscr{Y}^{multi} = \{B\text{-}disease, I\text{-}disease, B\text{-}gene, \\ I\text{-}gene, B\text{-}drug, I\text{-}drug, O\} \quad (8.3)$$

8.2.2 TaughtNet

The structure of this section reflects the procedural steps summarized in Figure 8.1 by comprehensively describing the phases involved in the training procedure: (1) datasets aggregation, (2) retrieval of teacher distributions, (3) aggregation of teacher distributions and (4) student training.

Table 8.2. Notations

Symbol	Description
$\mathbb{E}$	set of entities we aim to recognize
$e_i \in \mathbb{E}$	i-th entity
H	maximum sequence length
$\mathbf{x}$	sentence (sequence of tokens)
$x_j \in \mathbf{x}$	j-th token
$\mathbf{y}$	label list associated to a sentence $\mathbf{x}$
$y_j \in \mathbf{y}$	label associated to x_j.
$\mathscr{D}_i$	corpus of annotated sentences associated with the entity e_i or the student s, $\mathscr{D}_i = \{(\mathbf{x}, \mathbf{y}) \in \mathscr{X}_i \times \mathscr{Y}_i\}$
$\mathscr{X}_i$	set of sentences $\mathbf{x}$ (sequences of tokens $x_j \in \mathbf{x}$, $j \in \{1, \ldots, \mathrm{H}\}$) associated to $\mathscr{D}_i$
$\mathscr{Y}_i$	set of label lists $\mathbf{y}$ associated to each sentence $\mathbf{x} \in \mathscr{X}_i$
θ_T^i	model parameters of the teacher associated to the entity e_i
θ_S	model parameters of the student
$T_i^j\{\cdot\}$	output distribution of the teacher related to the entity e_i for the input token x_j
$\mathscr{A}^j\{\cdot\}$	output distribution resulting from the aggregation of teacher distributions for the input token x_j
$S\{\cdot\}$	student output distribution for the input token x_j
$\mathscr{L}(\cdot;\cdot)$	training loss function
$\mathscr{L}_{KD}(\cdot;\cdot)$	loss component based on teacher distributions (knowledge distillation)
$\mathscr{L}_{GT}(\cdot;\cdot)$	loss component based on ground truth
λ	hyper-parameter allowing to control the weight of the two loss components

Datasets aggregation Based on the available training datasets $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_{\mathbb{E}}$, we build an aggregated dataset:

$$\mathscr{D}_S = \{(\mathbf{x}, \mathbf{y}) \in \mathscr{X}'_S \times \mathscr{Y}'_S\}, \tag{8.4}$$

where $\mathscr{X}'_S$ results from the concatenation of the sentences contained in each single-task dataset $\mathscr{X}'_S = \mathscr{X}_1\mathscr{X}_2 \ldots \mathscr{X}_{|\mathbb{E}|}$, and the same goes for labels $\mathscr{Y}'_S$ with the only difference that B and I labels are diversified based on the corresponding entities, as described in Section 8.2.1.

The aggregated dataset $\mathscr{D}_S$ will serve as the data source to obtain the distribution representing the knowledge of teachers (used for knowledge distillation) and as a ground truth reference during student training.

Retrieval of Teacher predictions Let $\theta_T^1, \theta_T^2, \ldots, \theta_T^{|\mathbb{E}|}$ be the parameters learnt by *teacher* models on their corresponding single-task datasets. For each sentence token $x_j \in \mathbf{x}$, the i-th teacher will be able to provide the distribution T_i^j:

$$T_i^j\{y_j = k|\mathbf{x}; \theta_T^i)\}, k \in \{B, I, O\} \tag{8.5}$$

Distributions aggregation Thanks to knowledge distillation, a *student* model learns how to mimic the output distribution of a *teacher* model. Differently from the standard approach, our *student* has to learn from an heterogeneous set of teachers, each of them able to individuate a different entity type. Hence, we need an aggregation phase, where teacher distributions are merged in one single distribution to be used in the knowledge distillation framework.

Let $x_j \in \mathbf{x}$ be a token we have to aggregate distributions for. Let's denote with $p_k^i = T_i^j(y_j = k|\mathbf{x}; \theta_T^i)$ the probability which the i-th teacher assigns to the label k, where $k \in \{B, I, O\}$.

The probability of the token x_j being assigned to the label B-e_i, I-e_i and O can be respectively computed as the probability of the intersection of the events shown as follows:

$$P(B\text{-}e_i) \quad = \quad P\big((T_i \text{ assigns } B) \ \cap_{j \neq i} \ (T_j \text{ does not assign } B)\big) \tag{8.6}$$

$$P(I\text{-}e_i) \quad = \quad P\big((T_i \text{ assigns } I) \ \cap_{j \neq i} \ (T_j \text{ does not assign } I)\big) \tag{8.7}$$

$$P(O) = P\big(\cap_i (T_i \text{ assigns } O)\big) \tag{8.8}$$

Given the independence between teachers and the mutual exclusivity characterizing each teacher distribution, we can then compute the probabilities of the aggregated distribution $\mathscr{A}$ as follows:

$$\mathscr{A}^j(y_j = B\text{-}e_i|\mathbf{x}; \theta_T^1, \ldots, \theta_T^{|\mathbb{E}|}) = p_B^i \prod_{j \neq i} (p_I^j + p_O^j) \tag{8.9}$$

$$\mathscr{A}^j(y_j = I\text{-}e_i|\mathbf{x}; \theta_T^1, \dots, \theta_T^{|\mathbb{E}|}) = p_I^i \prod_{j \neq i} (p_B^j + p_O^j) \tag{8.10}$$

$$\mathscr{A}^j(y_j = O|\mathbf{x}; \theta_T^1, \dots, \theta_T^{|\mathbb{E}|}) = \prod_i p_O^i \tag{8.11}$$

Given a sentence token $x_j \in \mathbf{x}$, $j \in \{1, \dots, \mathrm{H}\}$, the output of this phase is the distribution of $\mathscr{Y}'$ labels:

$$\mathscr{A}^j\{y_j = k|\mathbf{x}; \theta_T^1, \dots, \theta_T^{|\mathbb{E}|}\}, k \in \mathscr{Y}' \tag{8.12}$$

Running example

Let $x_j \in \mathbf{x}$ be the input token and $T_{disease}^j = \{0.8, 0.1, 0.1\}$, $T_{gene}^j = \{0.05, 0.05, 0.9\}$, $T_{drug}^j = \{0.1, 0.1, 0.8\}$ its associated teacher predictions for labels B, I and O. Resulting from the aggregation of distributions we will have:

$$\begin{aligned}\mathscr{A}^j = \{&0.8 \cdot (0.05 + 0.9) \cdot (0.1 + 0.8),\\ &0.1 \cdot (0.05 + 0.9) \cdot (0.1 + 0.8), \dots, 0.1 \cdot 0.9 \cdot 0.8\} =\\ &= \{0.68, 0.09, 0.01, 0.04, 0.02, 0.09, 0.07\},\end{aligned}$$

where results for labels $\{B\text{-}disease, I\text{-}disease, B\text{-}gene,$ $I\text{-}gene, B\text{-}drug, I\text{-}drug, O\}$ are reported in order.

Student Training Let us represent the student model with its parameters θ_S and its output distribution $S\{y_t = k|\mathbf{x}; \theta_\mathbf{S}\}$, $k \in \mathscr{Y}'$. The fine-tuning procedure aims to minimize a loss function composed by two terms: the former measuring the distance of the student distribution from its teachers distribution, the latter representing its error on the ground-truth. Formally, we can define our loss as shown below:

$$\begin{aligned}\mathscr{L}(\mathscr{D}_S; \theta_S, \theta_T^1, \dots, \theta_T^{|\mathbb{E}|}) = \lambda \mathscr{L}_{KD}(\mathscr{D}_S; \theta_S, \theta_T^1, \dots, \theta_T^{|\mathbb{E}|}) +\\ + (1 - \lambda)\mathscr{L}_{GT}(\mathscr{D}_S; \theta_S),\end{aligned} \tag{8.13}$$

where $\mathscr{L}_{KD}$ and $\mathscr{L}_{GT}$ are the knowledge distillation and ground-truth loss, respectively, while λ is an hyperparameter controlling their weight on the overall loss $\mathscr{L}$.

Despite the Kullback-Leibler divergence being suitable for this knowledge-distillation task, similarly to [238] and in compliance with [31] which proves that minimizing the Kullback-Leibler divergence is equal to minimize the cross-entropy error between two distributions, it is sufficient to train the student model to minimize the following loss function:

$$\begin{aligned}&\mathscr{L}_{KD}(D_S;\theta_S,\theta_{\mathbf{T}}) = \\ &\qquad -\sum_{(x,y)\in D_S}\sum_{t=1}^{H}\sum_{k\in\mathscr{Y}'} T\{y_j = k|\mathbf{x};\theta_T^1,\ldots,\theta_T^{|\mathbb{E}|}\}\cdot \\ &\qquad\qquad \cdot\, logS\{y_t = k|\mathbf{x};\theta_{\mathbf{S}}\}, \qquad (8.14)\end{aligned}$$

where $\mathscr{H}$ is the sequence length and $S\{\cdot\}$ denotes the student distribution.

The ground-truth-based loss function is:

$$\begin{aligned}&\mathscr{L}_{GT}(D_S;\theta_S) = \\ &\qquad -\sum_{(x,y)\in D_S}\sum_{t=1}^{H}\nVdash\{y_t = k\}logS\{y_t = k|\mathbf{x};\theta_{\mathbf{S}}\}, \qquad (8.15)\end{aligned}$$

where the indicator $\nVdash\{\cdot\}$ represents the one-hot label annotated in the ground truth.

8.3 Experiments

In this section, an empirical evaluation of *TaughtNet* is presented. First, three student models are trained for three distinct biomedical entity types: diseases, chemical compounds, and genetic information. Subsequently, a variety of student architectures with differing sizes and parameters are trained, with results detailed in the *Results* subsection. The findings include: (1) a comparison of the top-performing student model against

Table 8.3. Datasets and performance of teachers used in the experiments

				Teacher		
Dataset	**Size**	**Entity type**	**# Mentions**	**Precision**	**Recall**	**F1**
NCBI	793 abstracts	Disease	6, 881	87.31	89.58	88.43
BC5CDR	1500 articles	Chemical	15, 935	94.38	94.19	94.28
BC2GM	20, 000 sentences	Gene	24, 583	85.24	86.17	85.70

several state-of-the-art benchmarks; (2) outcomes of students based on varying architectures and sizes; (3) a comparative analysis of all student models regarding prediction consensus; (4) an error assessment concerning various error categories; and (5) an exploration into the internal mechanism shifts from teachers to student models.

8.3.1 Experimental setup

Datasets and teachers

Performance evaluation of the proposed approach was conducted using three benchmark datasets. These datasets were constructed from PubMed abstracts: NCBI-Disease [55], BC5CDR [133], and BC2GM [226]. The datasets, inclusive of their training, development, and test splits, were sourced from: https://github.com/dmis-lab/biobert. Word labels were encoded using the IOB2 notation format [206].

For each dataset, teacher models were trained through fine-tuning over 30 epochs on a RoBERTa-large architecture. This architecture had been previously trained on PubMed, PMC, and MIMIC-III using a BPE Vocab derived from PubMed [127].

A summary of the datasets, in terms of size and entity-type, and of the teachers, in terms of their precision, recall and F1 scores, is provided in Table 8.3.

Evaluation details

For all the datasets, the same dataset splits as BioBERT [124] have been used, which are based on earlier publications for a fair evaluation. In particular, training/development/test splits of NCBI-disease and BC5CDR corpora are the same as their original version, while the training set of BC2GM has been modified because the original corpus does not provide a development set. Thus, 2,500 sentences are split off from the training data to generate the development set.

Metrics

Quality For evaluating the quality of named entity recognition approaches, the metrics *Precision*, *Recall*, and *F1* were utilized, calculated using the seqeval Python framework. In essence, *Precision* denotes the percentage of entities correctly identified by the system, while *Recall* represents the percentage of entities from the test set detected by the system. A system exhibiting low *Precision* struggles to differentiate between entity types. Conversely, low *Recall* signifies the system's inefficiency in recognizing entities.

The degree of agreement between different models was ascertained using the *Cohen's Kappa* metric, defined as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{8.16}$$

where p_o is the relative observed agreement among predictions, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.

Memory occupation and inference time The efficiency of models has been evaluated based on their *size* (in terms of MB of memory occupied) and the *samples-per-second* (SPS) required during the training and inference phases. A model with too many parameters is difficult to deploy on hardware systems with strict memory constraints, while a slow model is difficult to integrate in complex systems where the NER engine is just a step in a pipeline. Experiments have been performed on a Oracle Cloud Infrastructure (OCI) with an Intel(R) Xeon(R) Platinum 8167M CPU @ 2.00GHz (12 cores) and a NVIDIA Tesla V100 SXM2 GPU.

Settings and hyperparameters

The framework was developed utilizing the HuggingFace *transformers* library [264]. Various model architectures and weight sizes were ex-

Table 8.4. Comparative experiments. Best scores are reported in bold.

Dataset	Metric	Merged	MTM-CW	CollaboNet	MT-BioNER	TaughtNet
NCBI	Precision	83.88	85.86	85.48	86.73	**88.51**
	Recall	85.10	86.42	87.27	89.70	**89.90**
	F1	84.49	86.14	86.36	88.10	**89.20**
BC5CDR	Precision	94.08	89.10	94.26	88.46	**94.51**
	Recall	83.87	88.47	92.38	90.52	**93.40**
	F1	88.69	88.78	93.31	89.50	**93.95**
BC2GM	Precision	83.29	82.10	80.49	82.01	**84.90**
	Recall	78.94	79.42	78.99	84.04	**83.45**
	F1	81.06	80.74	79.73	83.01	**84.84**

plored. Specifically, models such as RoBERTa—large—PM—M3—Voc and RoBERTa—base—PM—M3—Voc—train—longer from Lewis et al. [127], along with huawei—noah/TinyBERT_General_4L_312D and distilroberta —base from the HuggingFace model hub were used.

For fine-tuning, the Adam optimizer was employed, having an initial learning rate of 5e-5, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e{-}8$. Batch sizes were set at 8, with a maximum sequence length of 128.

Regarding performance, while quality outcomes were satisfactory from initial epochs, optimal performance was typically observed after 20 epochs, aligning with findings from Lee et al. [124].

For the student model's training loss function, both KLDivLoss and NLLLoss PyTorch implementations were used for the knowledge distillation $\mathscr{L}KD$ and ground-truth $\mathscr{L}GT$ loss components.

8.3.2 Results

Comparison with baselines

The quality of the best student has been compared with several baselines, described as follows:

- *Merged*: the simplest way to train a multi-label NER model from single-entity datasets is to merge them in one aggregated dataset to be used for training and testing. We fine-tuned until convergence the same RoBERTa-large model architecture used for teachers on such dataset.

- *MTM-CW*: multi-task model built upon a single-task BiLSTM-CRF model with an additional context-dependent BiLSTM layer to model character sequences.
- *CollaboNET*: aggregates the results of *collaborator* single-task models, and uses them as an additional input to the target multi-task model.
- *MT-BioNER*: multi-task transformer-based neural architecture, where different models for different datasets share some layers to build a "shared" knowledge across tasks.

Table 8.4 reports results over the three benchmark datasets in terms of *Precision*, *Recall* and *F1* scores. Thanks to the utilization of high-performing teachers, the *student* model achieves the best results for each of the datasets. Interestingly, performance obtained for the *NCBI* dataset surpasses the related teacher thanks to the indirect positive effect of the (1) data augmentation obtained by merging all the dataset and the (2) joint training based on both the ground-truth and teacher predictions. A comparative discussion with baselines is provided in Section 8.3.3.

Smaller and smaller students

Thanks to its knowledge distillation based architecture, one of the advantages of using *TaughtNet* is its straightforward way to train multi-task small models by leveraging the knowledge of large and high-performing teachers. In experiments, results of different student architectures have been compared:

- *Large*: same as teachers', i.e. RoBERTa-large architecture pre-trained on PubMed and PMC and MIMIC-III with a BPE Vocab learnt from PubMed.
- *Base*: RoBERTa-base architecture pre-trained on PubMed and PMC and MIMIC-III with a BPE Vocab learnt from PubMed with an additional 50K steps.

Table 8.5. Performance of various student architectures with decreasing size. SPS stands for samples-per-second.

	SPS scores			*F1 scores*		
Model	**train**	**inference**	**Size (MB)**	**NCBI**	**BC5CDR**	**BC2GM**
Large	38	125	1,416.3	89.20	93.95	84.84
Base	111	324	495.5	87.62	93.63	84.25
Distil	202	566	265.5	80.71	85.24	77.45
Tiny	475	914	57	77.80	81.42	71.94

- *Distil*: distilled version of BERT base introduced by Sanh et al. [207]. It has 40% less parameters and runs 60% faster than BERT-base.

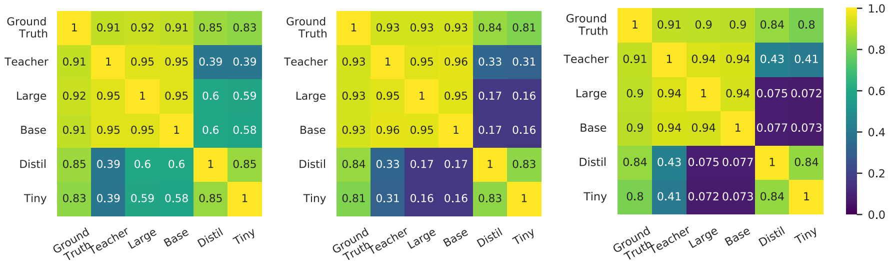
- *Tiny*: distilled version of BERT-base introduced by Jiao et al. [97], 7.5x smaller and 9.4x faster on inference than BERT-base.

Results are reported in Table 8.5 in terms of model size, samples-per-second (SPS) processed during the training and inference phase, and F1 scores over the three benchmark datasets. Interestingly, the *Base* architecture achieves F1 scores closely resembling its *Large* counterpart, probably resulting in the best choice in the trade-off between quality of predictions and model size / inference time. The distilled architectures (*Distil* and *Tiny*) result in lower F1 scores, but their considerable improvement in memory occupation and processing time could make them a suitable choice in limited-resource scenarios. In the experiments that follow, the differences between these students and their corresponding teachers will be presented.

Levels of agreement (Cohen's Kappa)

The Cohen's Kappa metric has been computed to measure the degree of agreement among models and the ground-truth[2]. Heatmaps in Figure 8.2 show agreements over the three benchmark datasets among the ground truth, the teacher, and the size-decreasing student architectures. Despite

[2] When computing the Cohen's Kappa between a teacher and a student, the *outside* (O) label has been assigned to predictions of entity types which are different from the type in which the teacher is specialized.



(a) NCBI (disease) **(b)** BC5CDR (chem) **(c)** BC2GM (gene)

Figure 8.2. Heatmaps of Cohen's Kappas among models and ground truth.

the disagreement between distilled models and their teacher — which highlights a limitation in distilling their knowledge, which will be explored in future work — results show an overall agreement between teachers and their students and among student architectures.

Error Analysis

We further explored the differences among models based on the number of correctly-retrieved entity mentions (CORRECT), new predictions deriving from the application of the framework (NEW) and their errors, which can be divided into five categories described as follows:

- *Complete False Positive (CPF)*: the model recognizes an entity which was not annotated as a named entity.
- *Complete False Negative (CFN*: the model does not recognize an entity which was annotated as a named entity.
- *Wrong Label Right Span (WLRS)*: the model correctly recognizes the presence of an annotated named entity, but assigns the wrong label.
- *Wrong Label Overlapping Span (WLOS)*: the model recognizes the presence of an annotated named entity, but assigns the wrong label and the span is wrong.

Table 8.6. Number of correct predictions, new predictions (for entity types not annotated in the ground truth) and different types of error.

Dataset	Model	CORRECT	NEW	CFP	CFN	WLRS	WLOS	RLOS
NCBI	Large	785	803	43	24	9	24	113
	Base	785	801	51	19	8	18	125
	Distil	742	172	61	126	1	7	79
	Tiny	720	185	83	158	3	5	69
BC5CDR	Large	4753	4612	190	189	58	93	279
	Base	4750	4654	194	168	50	97	307
	Distil	4187	307	143	995	48	42	100
	Tiny	4032	357	339	1120	48	41	131
BC2GM	Large	5330	3426	257	213	48	37	608
	Base	5306	3453	309	202	56	32	640
	Distil	4797	625	468	699	18	31	685
	Tiny	4393	764	617	987	35	48	767

- *Right Label Overlapping Span (RLOS)*: the model recognizes the presence of an annotated named entity, but the span is wrong.

It can be seen from the data in Table 8.6 that students trained with *TaughtNet* allow us to retrieve a considerable number of novel entity mentions which were not annotated in the ground-truth, thanks to the knowledge of the teachers employed. Concordant with the above-reported experiments, *Large* and *Base* students are able to detect a significantly higher number of new entity mentions w.r.t. distilled architectures. The highest limitation of distilled architectures w.r.t. to their "larger" counterparts is in the number of CFN errors, i.e. they are not able to identify mentions which are actually annotated.

The majority of errors fall in the RLOS category, meaning that models are able to identify an entity mention, but the range detected is not the same as the ground truth. However, previous works have shown that this type of errors are often a result of the subjectivity and inconsistency of span annotations [244, 179]. Some examples are shown in Table 8.7. It is important to note that many of the errors are due to the ability of our model to recognize multiple entity types: for example, the two words *gene* mention "estrogen receptor" (see WRLS, 2nd example) are assigned by our model to two different entity types ("estrogen" as a *chemical* compound, "receptor" as a *gene*).

Table 8.7. Samples annotated by the Large model which contain errors in disease, chemical and/or gene mention predictions. Errors are highlighted with a higher level of transparency (disease, chemical, gene). One input sample is randomly selected for each dataset and error.

Error	Student annotation	Ground Truth
CFP	Other complement components were normal during remission of lupus, but C1, C4, C2, and C3 levels fell during exacerbations.	Other complement components were normal during remission of lupus, but C1, C4, C2, and C3 levels fell during exacerbations.
	The encoded protein contains an amino terminal PDZ domain, followed by a predicted coiled-coil region, a PEST domain, and a carboxy-terminal SAM domain.	The encoded protein contains an amino terminal PDZ domain, followed by a predicted coiled-coil region, a PEST domain, and a carboxy-terminal SAM domain.
	We conclude that CNA and INA demonstrated similar profiles with regard to safety, morbidity, and mortality.	We conclude that CNA and INA demonstrated similar profiles with regard to safety, morbidity, and mortality.
CFN	If untreated, hemochromatosis can cause serious illness and early death, but the disease is still substantially underdiagnosed.	If untreated, hemochromatosis can cause serious illness and early death, but the disease is still substantially underdiagnosed.
	ORF3 encodes a putative periplasmic c-type cytochrome with a molecular mass of 94,000 Da and contains seven c-heme-binding motifs but shows no sequence homology to occ or ORF1.	ORF3 encodes a putative periplasmic c-type cytochrome with a molecular mass of 94,000 Da and contains seven c-heme-binding motifs but shows no sequence homology to occ or ORF1.
	BS pool size was decreased by 27% but total BS synthesis was not affected by EE in intact rats.	BS pool size was decreased by 27% but total BS synthesis was not affected by EE in intact rats.
WLRS	HFE is an MHC-related protein that is mutated in the iron-overload disease hereditary hemochromatosis.	HFE is an MHC-related protein that is mutated in the iron-overload disease hereditary hemochromatosis.
	Effects of long-term use of raloxifene, a selective estrogen receptor modulator, on thyroid function test profiles.	Effects of long-term use of raloxifene, a selective estrogen receptor modulator, on thyroid function test profiles.
	Nociceptin, also known as orphanin FQ, is an endogenous ligand for the orphan opioid receptor-like receptor 1 (ORL1) and involves in various functions in the central nervous system (CNS).	Nociceptin, also known as orphanin FQ, is an endogenous ligand for the orphan opioid receptor-like receptor 1 (ORL1) and involves in various functions in the central nervous system (CNS).
WLOS	Previous family studies suggested that these individuals may be compound heterozygotes for the common mutant TSD gene and a rare (allelic) mutant gene.	Previous family studies suggested that these individuals may be compound heterozygotes for the common mutant TSD gene and a rare (allelic) mutant gene.
	When expressed in Escherichia coli, SH-PTP2 displays tyrosine-specific phosphatase activity.	When expressed in Escherichia coli, SH-PTP2 displays tyrosine-specific phosphatase activity.
	Sub-chronic inhibition of nitric-oxide synthesis modifies haloperidol-induced catalepsy and the number of NADPH-diaphorase neurons in mice.	Sub-chronic inhibition of nitric-oxide synthesis modifies haloperidol-induced catalepsy and the number of NADPH-diaphorase neurons in mice.
RLOS	The evidence of a significant proportion of loss-of-function mutations and a complete absence of the normal copy of ATM in the majority of mutated tumours establishes somatic inactivation of this gene in the pathogenesis of sporadic T-PLL and suggests that ATM acts as a tumour suppressor.	The evidence of a significant proportion of loss-of-function mutations and a complete absence of the normal copy of ATM in the majority of mutated tumours establishes somatic inactivation of this gene in the pathogenesis of sporadic T-PLL and suggests that ATM acts as a tumour suppressor.
	A GT-rich sequence binding the transcription factor Sp1 is crucial for high expression of the human type VII collagen gene (COL7A1) in fibroblasts and keratinocytes.	A GT-rich sequence binding the transcription factor Sp1 is crucial for high expression of the human type VII collagen gene (COL7A1) in fibroblasts and keratinocytes.
	An increase in TDR by dl-sotalol facilitated transmural propagation of EADs that initiated multiple episodes of spontaneous TdP in 3 of 6 rabbit left ventricles.	An increase in TDR by dl-sotalol facilitated transmural propagation of EADs that initiated multiple episodes of spontaneous TdP in 3 of 6 rabbit left ventricles.

Explainability

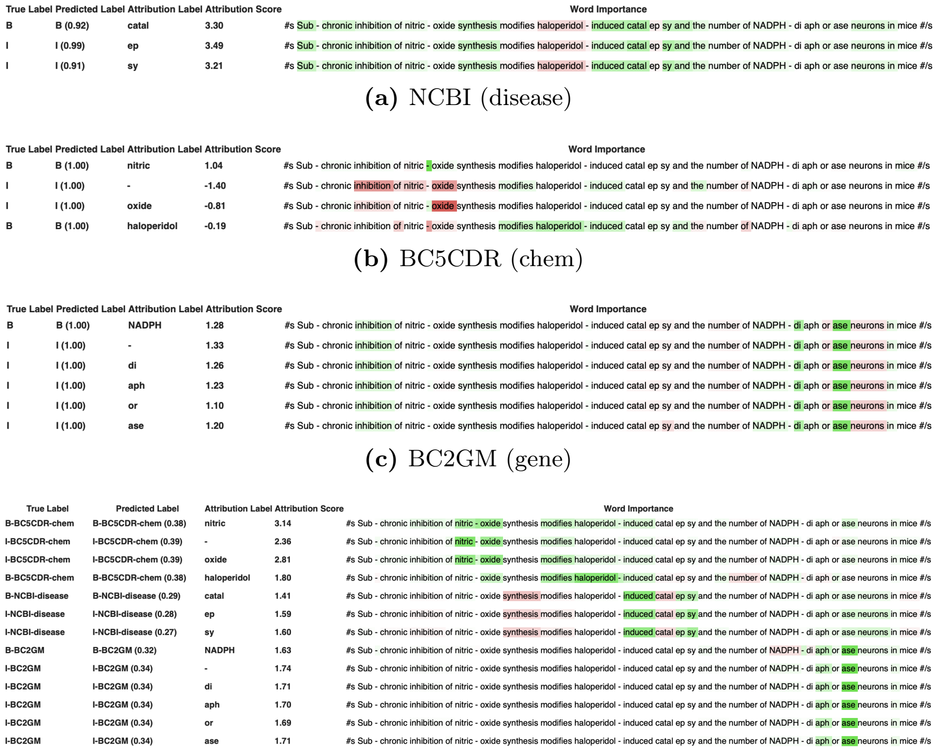
We apply *Integrated Gradients* [235] to assign an importance score to each input token by approximating the integral of gradients of the output w.r.t the inputs[3]. To investigate how the inner workings of the models change from Teachers to Student, we report in Figure 8.3 the explanations from the three *large* Teachers and the resulting Student to the sentence: *"Sub-chronic inhibition of nitric-oxide synthesis modifies haloperidol-induced catalepsy and the number of NADPH-diaphorase neurons in mice"*, which contains at least one mention per entity type. Interestingly, despite our experiment being carried out with just the aim to prove the effortlessly interpretability of our method — which does not modify the architecture of the Student model and thus can leverage off-the-shelf methods to explain its predictions —, we also observed that the explanations provided by the Student are better targeted (i.e. lower number of influential tokens) and understandable.

8.3.3 Discussion

In the conducted experiments, an in-depth analysis was performed on the effects of learning from various single-task transformer-based teachers, contrasting TaughtNet with notable baselines from existing literature.

As presented in Table 8.8, a methodological comparison of cutting-edge methods is accompanied by average precision, recall, and F1 scores for the benchmark datasets employed. The findings indicate that multi-task methods leveraging high-performing pre-trained transformer models consistently surpass CollaboNet in several scenarios. This is notable considering CollaboNet's capability to address the low-precision issue encountered in multi-task learning systems. This is achieved through its collaborative framework, comprised of single-task BiLSTM-CRF models, which also tackles the *type conflict* issue - the scenario where different models identify identical mentions. TaughtNet incorporates the benefits of both multi-task learning and transformers, addressing the same challenges as CollaboNet. The outcome is an efficient fine-tuned transformer model capable of recognizing mentions across various entity types. This offers two

[3] Integrated Gradients for Transformers interpretability, code: https://github.com/cdpierse/transformers-interpret



True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B	B (0.92)	catal	3.30	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (0.99)	ep	3.49	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (0.91)	sy	3.21	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s

(a) NCBI (disease)

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B	B (1.00)	nitric	1.04	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	-	-1.40	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	oxide	-0.81	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
B	B (1.00)	haloperidol	-0.19	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s

(b) BC5CDR (chem)

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B	B (1.00)	NADPH	1.28	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	-	1.33	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	di	1.26	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	aph	1.23	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	or	1.10	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	ase	1.20	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s

(c) BC2GM (gene)

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B-BC5CDR-chem	B-BC5CDR-chem (0.38)	nitric	3.14	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC5CDR-chem	I-BC5CDR-chem (0.39)	-	2.36	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC5CDR-chem	I-BC5CDR-chem (0.39)	oxide	2.81	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
B-BC5CDR-chem	B-BC5CDR-chem (0.38)	haloperidol	1.80	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
B-NCBI-disease	B-NCBI-disease (0.29)	catal	1.41	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-NCBI-disease	I-NCBI-disease (0.28)	ep	1.59	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-NCBI-disease	I-NCBI-disease (0.27)	sy	1.60	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
B-BC2GM	B-BC2GM (0.32)	NADPH	1.63	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	-	1.74	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	di	1.71	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	aph	1.70	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	or	1.69	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	ase	1.71	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s

(d) Student (disease, chem, gene)

Figure 8.3. Visualization of attribution scores computed by applying Integrated Gradients to our Teachers (a-c) and Student (d). For each token, we show its true and predicted label, its attribution score and the original sentence where each token is highlighted based on its contribution to the prediction (green if positive, red otherwise)

Table 8.8. Overview of state-of-the-art approaches for multi-task BioNER. Precision, Recall and F1 scores shown here have been averaged across the benchmarking datasets used in this work. Best results are reported in bold.

Method	Ref	Year	Model	Characteristics	Precision	Recall	F1
MTM-CW	[258]	2018	Multi-task BiLSTM-CRF model	(+) outperforms the previous state-of-the-art by leveraging multi-task learning; (-) type conflict problem	85.69	84.77	85.22
CollaboNet	[277]	2019	Collaborating BiLSTM-CRF models	(+) handles the low-precision problem of multi-task models; (+) no type conflict problem; (-) requires single-task models to be trained in advance and called for every inference	86.74	86.21	86.46
MT-BioNER	[101]	2020	Multi-task transformer model	(+) obtains state-of-the-art performance thanks to the use of a transformer-based architecture; (-) suffers from the low-precision problem of multi-task learning systems; (-) type conflict problem	85.73	88.08	86.87
TaughtNet	—	—	Transformer model	(+) combines the advantages of multi-task learning, transformer architectures and CollaboNet; (+) easy to lighten; (+) no low-precision problem; (+) no type conflict problem; (+) easy applicability of eXplainable AI techniques; (-) requires teachers to be trained in advance	**89.31**	**88.91**	**89.33**

key advantages: (1) the possibility to deploy lighter models such as DistilBERT and TinyBERT under constrained hardware and computational requirements, and (2) the straightforward application of existing explainability techniques, given that the architecture remains unaltered.

8.4 Conclusion & Future work

The difficulty in finding a single dataset with all the entities required for a Biomedical Named Entity Recognition System (e.g. diseases, genes, species, drugs) has laid the foundations of this chapter. TaughtNet has the objective to integrate various publicly available single-task healthcare datasets in a single BERT architecture which can be used as a fast and highly performing BioNER engine in real applications, such as conversational agents or knowledge graph development.

Experimental results demonstrate that not only does TaughtNet surpass strong state-of-the-art baselines, but it also is a valuable option when constrained by strict computational and memory requirements thanks to its ability to train lightweight models that distill the knowledge from high-performing single-task teachers. Furthermore, we have shown the potential of TaughtNet to provide explainability, which is a valuable advantage, especially when dealing with healthcare data.

There is abundant room for further progress in exploring the use and application of knowledge distillation to bring the student performance as close as possible to that of teachers. As a future work, we would like to integrate more datasets and to extend the framework not only to other downstream tasks, but also to other application domains, since the technique is not dependent on the biomedical domain.

Chapter 9

Multi-task learning for few-shot biomedical relation extraction

Relation Extraction (RE) is a subfield of text classification, a natural language processing (NLP) task that aims to automatically associate unstructured text to one [86, 85] or several [102] labels. Specifically, RE aims to identify and extract relationships between entities in unstructured text data. This task is crucial for various applications such as information retrieval [68], question answering [123], knowledge graph construction [253] and text summarization [160]. One of the major challenges in the field of relation extraction is the high variability and complexity of the language used to express relationships. To address this challenge, various methods have been proposed, including rule-based methods [178, 19], machine learning-based methods [8, 83], and hybrid approaches [94, 282] that combine both.

In recent years, there has been a surge in research in the fields of soft computing [2, 5, 4, 3] and relation extraction and their potential for various NLP applications, especially in biomedical text understanding. Applications include detecting protein-protein interactions (PPIs) [192] and extracting information on adverse drug events (ADEs) [72]. One major driving factor behind the advancements in relation extraction for biomedical text is the integration of attention mechanisms [128] into NLP models.

These mechanisms enable the models to concentrate on specific parts of the input, which is crucial when dealing with complex biomedical text that contains a high density of specialized terminology. Furthermore, the widespread availability of pre-trained biomedical language models has also been a key factor in enhancing the performance of relation extraction tasks. These models have been trained on vast amounts of biomedical data and can be fine-tuned for specific tasks, resulting in substantial improvements in performance [127]. Overall, the recent advancements in NLP and the availability of pre-trained biomedical language models have paved the way for a new generation of relation extraction models with improved performance. These models can extract valuable information from biomedical text with greater accuracy and efficiency, providing benefits for various biomedical applications.

While relation extraction for biomedical text has seen significant progress, the lack of large and high-quality annotated biomedical datasets remains a major challenge. The annotation process is time-consuming and requires extensive domain knowledge, making it expensive to obtain large amounts of annotated data. As a result, this has a significant impact on the performance of relation extraction models in real-world applications. To overcome these limitations, there is a growing need to shift focus from model-centric to data-centric AI, emphasizing the critical role of data in the learning process and the need to extract maximum value from it. Such a shift would enable the development of more effective and robust relation extraction models, addressing the limitations of limited annotated datasets.

Multi-task learning [30] is a technique that aims to address the issue of limited annotated training data by leveraging the similarities between different datasets. This approach involves training a single model on multiple related tasks, using the similarities between the tasks to improve the training process. This technique has been widely adopted in biomedical text understanding and has demonstrated its usefulness in several studies [189]. However, despite its advantages, multi-task learning can also result in a degradation of performance if the datasets used have different structures and objectives. The size and underlying properties of the datasets can also have an impact on the performance of the model [9]. Thus, careful consideration should be given to the choice of datasets used in multi-task learning to ensure optimal results.

In this chapter, a multi-task framework for biomedical relation extraction (RE) is introduced. This framework leverages a renowned multi-task learning approach [149] and utilizes three prominent multi-class datasets: DDI-2013, ChemProt, and I2B2-2010 RE. These datasets annotate relationships among various biomedical entities such as drugs, chemical compounds, proteins, medical issues, treatments, and tests. The design of the framework features a transformer-based model, incorporating shared layers for all three RE tasks while maintaining unique classification heads for each dataset. To augment performance, a training strategy rooted in knowledge distillation is implemented. Experimental outcomes explore the efficacy of this multi-task framework in situations with limited training data. The results indicate a significant enhancement, with F1 scores improving by as much as 65% over leading few-shot techniques when working with merely 10 training samples. This underscores the potential benefits of integrating multi-task learning in environments where procuring extensive annotated datasets is a challenge but acquiring smaller, analogous datasets from various clinical entities is more achievable. For those interested in further exploration, the implementation can be accessed at this GitHub repository: https://github.com/IoSylar/Multi-task-Learning-for-Biomedical-Relation-Extraction.

Key contributions from this work include:

- The introduction of a novel framework combining multi-task learning with knowledge distillation to address the challenges of few-shot learning in the domain of biomedical RE.

- Demonstrated superiority over contemporary benchmarks, consistently outperforming the current best methods in numerous few-shot learning contexts.

- A comprehensive, data-driven exploration into the principal elements influencing multi-task learning during the incorporation of varied biomedical datasets.

- A detailed performance analysis across scenarios with differing data availabilities, ranging from 1 to 1000 training samples.

The structure of this chapter is as follows: Section 2 delves into prior research focusing on relation extraction in the biomedical domain and its applicability in few-shot contexts. Section 3 provides a thorough description of the datasets employed and the methodological approach. Section 4 offers insights into the conducted experiments, their application in few-shot scenarios, and a discussion on the derived results. The chapter concludes in Section 5. The material presented in this chapter is based on the article "*Multi-task learning for few-shot biomedical relation extraction*" [171] published on the Artificial Intelligence Review journal.

9.1 Related Work

Relation Extraction (RE) has been thoroughly investigated in the realm of NLP and Information Extraction (IE). One of the most widely adopted rule-based methods is the use of regular expressions and lexical patterns to identify relationships, which rely on the manual creation of patterns that are specific to the target relationships and the domain of the text [178, 19]. While this approach has demonstrated good results, it is heavily dependent on the quality and coverage of the patterns. In contrast, machine learning-based approaches [8, 83] leverage supervised learning techniques to train models on annotated text data, enabling the models to learn to identify relationships based on context and features of the entities and their interactions. This approach is more robust and adaptable to new domains and relationships, but requires a substantial amount of annotated text data, which can be costly and time-consuming to obtain. Hybrid approaches [94, 282] combine the advantages of both rule-based and machine learning-based methods by using rule-based methods to pre-process the text and extract candidate relationships, which are then fed to a machine learning model for final classification. This approach can enhance performance and reduce the need for annotated data.

Relation extraction in biomedical applications presents a unique set of challenges compared to traditional NLP tasks. One of the key difficulties is the complexity of the domain-specific medical language, which often includes technical terms, acronyms, and abbreviations that are not found in general English text. Additionally, the relationships between entities in biomedical texts can be highly nuanced, with subtle differences in meaning that require a deep understanding of the biological and medical context. Despite these challenges, relation extraction has a wide range of potential applications in biomedical research, including the discovery of biological pathway [103] and associations between genes and diseases [161].

However, another important challenge is that annotated training data for relation extraction in the biomedical domain is limited, making it difficult to train machine learning models to accurately recognize relationships. While a vast amount of works on few-shot learning exist on image data [239, 236], these scenarios in RE are relatively under-studied. Hong et al. [84] propose a method based on distant supervision that automatically extract biomedical relations from large-scale literature repositories. Li et al [129] propose a joint model for named entity recognition and relation extraction based on a CNN for charactel-level representations and BiLSTMs. Chen et al. [33] introduce transformers as encoding layers of joint models to improve the performance in identifying patients suitable for clinical trials. Li et al. [137] explores the relatedness among multiple tasks by applying simple multi-task learning approaches.

Despite its advantages, when learning from multiple tasks it is possible that the performance of the resulting model may decrease compared to training a separate model for each task [9]. This can occur because the model may struggle to balance the optimization of multiple tasks, leading to sub-optimal performance on one or more tasks. Additionally, the tasks may have conflicting objectives or requirements, which can result in poor performance on some tasks. Furthermore, the model may over-generalize or over-fit to the training data, making it less effective at making predictions on unseen data. Therefore, it is important to carefully evaluate the trade-off between the potential benefits of multi-task learning and the potential risks to performance before choosing this approach for a given problem. In contrast to prior studies, this work goes beyond the evaluation of multi-task biomedical relation extraction models in few-shot scenarios and provides a comprehensive examination of the inter-task influences, both positive and negative, in our multi-task models.

9.2 Materials and Methods

In this section, we describe data, models and algorithms used to perform our experiments.

9.2.1 Datasets

The biomedical datasets used in this study are described in this section. We focus on three publicly available multi-class datasets for relation extraction: DDI-2013 [80], ChemProt [116], I2B2-2010 RE [246]. We use the same pre-processing procedure as in [127].

DDI-2013

This corpus consists in documents from the DrugBank database[1] and MedLine[2] abstracts annotated with pharmacological substances and their interactions. It is the first dataset highlighting (1) *pharmacodynamic (PD)*, i.e. the changes in pharmacological effects of a drug caused by the presence of another drug, and (2) *pharmacokinectic (PK)*, which occurs in presence of interference in the intake of one drug (i.e. the distribution or elimination of one drug from another).

The annotated relations are described as follows:

- *Mechanism*: describes the PK interference mechanism
- *Effect*: describes the effect of the intake of a drug or the PD mechanism
- *Advice*: highlights a recommendation or advice which regards interactions between drugs
- *Int*: indicates a drug-drug interaction without any additional information, explanations or advice

Size of training, development and test sets is: $\mid \mathscr{D}_{train} \mid= 29,334$, $\mid \mathscr{D}_{dev} \mid= 7,245$, $\mid \mathscr{D}_{test} \mid= 5,762$.

[1] https://go.drugbank.com

[2] https://www.nlm.nih.gov/medline/index.html

ChemProt

This corpus contains data from open source databases (e.g. CheMBL, BindingDB, PDSP Ki, DrugBank) annotated with chemical compounds, proteins and their interactions. We will consider the following groups of chemical-proteins relations (CPRs) in our study:

- *CPR 3*: indicates upregulation relations (activation, promotion, increased activity)
- *CPR 4*: indicates downregulation (inhibition, block, decreased activity)
- *CPR 5, CPR 6*: are related to interactions of type "agonist" and "antagonist", respectively.
- *CPR 9*: is related to substrate or part of relations. Therefore, this relation does not have particularly relevant features and is thus difficult to extract.

Size of training, development and test sets is: $| \mathscr{D}_{train} | = 19,461$, $| \mathscr{D}_{dev} | = 11,821$, $| \mathscr{D}_{test} | = 16,944$.

I2B2-2010 RE

This corpus focuses on relationships between medical concepts such as tests and treatments. The relation extraction task has 8 classes divided into 3 categories depending on the entities involves. We describes these categories as follows:

- *Medical problem-treatment relations*
  - *TrIP*: the treatment improves or cures the medical problem
  - *TrWP*: the treatment worsens the medical problem
  - *TrCP*: the treatment causes the medical problem
  - *TrAP*: the treatment is administered for the medical problem (the result is not mentioned in the sentence)
  - *TrNAP*: the tratment is not provided or is intermittently administered due to the medical problem

- *Medical problem-test relations*
    - *TeRP*: the test reveals the medical problem
    - *TeCP*: the test is conducted to investigate the medical problem (the sentence does not indicate the result but the reason for the test)
- *Medical problem-medical problem relations*
    - *PIP*: medical problem indicates medical problem

Size of training, development and test sets is: $\mid \mathscr{D}_{train} \mid = 21,385$, $\mid \mathscr{D}_{dev} \mid = 873$, $\mid \mathscr{D}_{test} \mid = 43,001$.

9.2.2 Method

In this section, we outline the methodology employed in our study. Specifically, we utilize a multi-task learning framework, MT-DNN [150], on three biomedical datasets for the purpose of Relation Extraction, as detailed in Section 9.2.1. As depicted in Figure 9.1, an Encoder based on a transformer architecture is shared among the tasks, and specialized classification heads are fine-tuned for each of the datasets. Subsequently, a knowledge distillation process is employed to enhance performance, as illustrated in Figure 9.2: multiple multi-task models are trained, with their predictions constituting the knowledge that is distilled by a single multi-task model.

Multi-task learning architecture: MT-DNN

We use a Multi-Task Deep Neural Network (MT-DNN) [150] as the multi-task framework for our experiments. The overall architecture is shown in Figure 9.1. The input $X = \{[CLS], x_2, \ldots, x_m\}$ is a word sequence of length m from one of the three analyzed datasets. The *Lexicon Encoder* maps each token x_i to its input embedding vector l_i obtained by summing the corresponding word, segment and positional embeddings. The pre-trained *Transformer Encoder* maps input embedding vectors into a sequence of contextual embedding vectors thus forming a shared representation across the different tasks. In this work, we use one of the

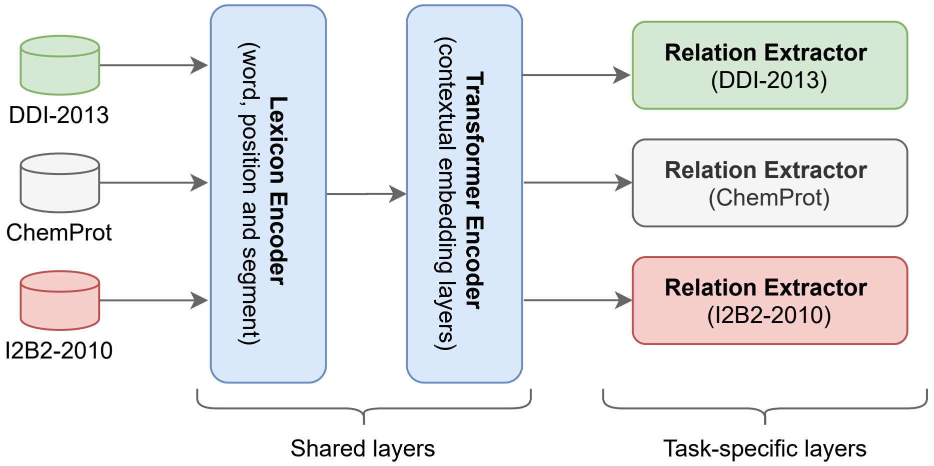


Figure 9.1. Overview of the multi-task architecture applied to our study. The Lexicon Encoder and Transformer Encoder are shared across the different tasks and maps the input first to a sequence of embedding vectors (one for each token) and then to shared contextual embedding vectors which take count of contextual information. A task-specific layer is then used for each dataset to generate dataset-specific representations.

pre-trained models made available by Lewis et al. [127] as the backbone of the multi-task framework. Task specific layers are defined as sentence classification models: the first token $[CLS]$ of each sentence X is a semantic representation of the sentence and the probability that X contains a relation between medical entities is predicted by a logistic regression with softmax:

$$P(isRelation \mid X) = softmax(\mathbf{W}_t^T \cdot \mathbf{x}), \tag{9.1}$$

where $\mathbf{W}_t^T$ is the parameter matrix for the task t.

Knowledge Distillation

The knowledge distillation (KD) method has been successfully used with multi-task learning to enjoy the advantages of ensemble learning while not needing to keep the entire ensemble of models but just one single model [148], our KD methodology is shown in Figure 9.2: we start by training three MT-DNN networks with three dropout values $p = \{0.1, 0.15, 0.2\}$

and each of them is then used as the backbone for a single-task network fine-tuned on each task dataset. Soft labels produced by teachers for each training example are then averaged to produce the *dark knowledge* to be distilled. We studied the effects of two types of KD loss: (1) Mean Squared Error (MSE) and (2) a hybrid loss based on Kullback Leibler divergence. MSE minimizes the mean squared discrepancy between the soft labels of the teacher and values estimated by the student network:

$$\mathscr{L}_{MSE} = \frac{1}{N} \sum (y - \hat{y}) \tag{9.2}$$

The hybrid loss is based on two contribution: the first is given by the Kullback Leibler loss which minimizes the divergence between two probability distributions, i.e. the soft labels of the teacher and the predictions of the student: the second contribution assumes that the teacher is not perfect and thus takes into account the ground truth by means of the cross-entropy loss:

$$\begin{aligned}\mathscr{L}_{hybrid} = \lambda \mathscr{L}_{CE}(y^i_\tau, f_\tau(x^i_\tau, \theta)) +& \\ (1-\lambda)\mathscr{L}_{KL}(f_\tau(x^i_\tau, \theta), f_\tau(x^i_\tau, \theta_T)),&\end{aligned} \tag{9.3}$$

where $\mathscr{L}_{CE}(y, \hat{y})$ denotes the cross-entropy loss, $y = y^i_\tau$ being the ground truth label for the i-th sample at time step τ and $\hat{y} = f_\tau(x^i_\tau, \theta)$ representing the predicted output for the i-th sample at time step τ, given the model parameters θ; $\mathscr{L}_{KL}$ denotes the Kullback-Leibler divergence between the output probability distribution from the model with parameters θ and the teacher with parameters θ_T; the parameter λ controls the weighting of the contribution of the knowledge distillation and ensures that the student also learns from the actual ground truth.

9.3 Experiments

Our analysis will be focused on answering the questions reported as follows.

- **Q1: Comparison with few-shot baselines.** ***How does few-shot MT-DNN perform as compared to few-shot learning***

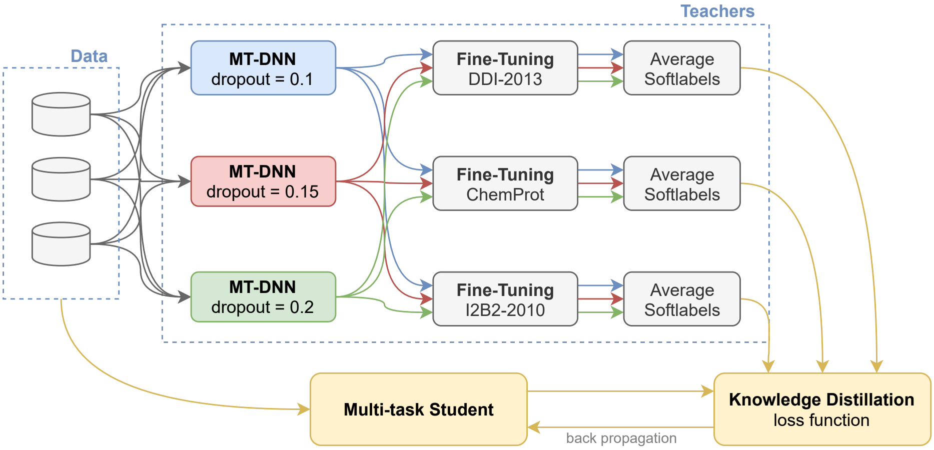


Figure 9.2. Overview of the knowledge distillation process applied in our study. First, MT-DNN networks are trained with different dropout values $p = \{0.1, 0.15, 0.2\}$. Each MT-DNN network is then fine-tuned on each dataset and all the soft-labels produced by teachers are averaged to produce the *dark knowledge* to be distilled. A single MT-DNN student is trained with a knowledge distillation loss which takes count of the knowledge acquired by teachers.

baselines? We use three few-shot learning baselines to perform a comparison with the multi-task architecture leveraged in this work: a Siamese network [109], ProtoNET [228], BioBERT [124], Clinical-BERT [10] and PET [210].

- **Q2: Effects of multi-task learning.** ***Can it improve the performance w.r.t. single-task models?*** We select one of the publicly available biomedical pre-trained transformer architectures as the base for our multi-task MT-DNN model, which is then enhanced with Knowledge Distillation and compared with single-task performance over the entire training-sets. Furthermore, we study how knowledge distillation impacts the overall performance by analyzing the effects of different values assigned to the loss weight λ

- **Q3: Tasks influence analysis.** ***What are the main influencing factors in multi-task learning?*** Different datasets can have a different impact over the multi-task performance. We will analyze similarities and differences between datasets to understand their effects on positive and negative transfer when training the multi-task model. On the basis of the above, we will analyze the mutual influence between different tasks by *pairwise training*, i.e. selectively excluding datasets from the training procedure to analyze their overall effects over the multi-task performance.

- **Q4: Few-shot scenarios.** ***How does the performance vary in few-shot scenarios?*** We are interested in the understanding of the value of multi-task learning when only a few set of data is available for each dataset, and how its effects vary when the training dataset increases. More in detail, we will train our multi-task models by simulating few-shot scenarios in where only k training examples are available for each dataset (with k varying from 1 to 1000) and we will test their performance over the entire test set.

9.3.1 Training parameters

In this section we report the training parameters used in our experiments. We fixed the `input sequence length` to 512 and the `batch size` to 8. We used the training parameters suggested in Liu et al. [150] for both the multi-task and single-task experiments. In particular, we conducted experiments by setting various hyperparameters such as learning rate, weight decay and optimizer using an initial random search and subsequently performing a greedy search focusing on the neighborhood of the default values on a subset of the training data, as commonly done in the literature. These preliminary experiments confirmed the suggested parameter values. Thus, we used an `Adamax` optimizer with `learning rate` set to 5e-5 and `weight decay` to 0.01 with adam `eps` to 1e-7. To avoid gradient explosion, the `grad clipping` parameter is set to 1.0. Additionally, we provide an empirical study on the value of the loss weighting parameter λ used in the knowledge distillation process.

When the training procedure involves the entire training dataset or at least 1000 examples, we set the number of `epochs` to 10 (in both the single-task and multi-task cases), while we set it to 20 in 1-, 10-, 50- and 100-shot scenarios.

The loss functions vary according to the type of approach: in single-task and simple multi-task learning, we use the cross-entropy loss; when using knowledge distillation, we experimented with MSE and a hybrid loss formed by cross-entropy and Kullback Leibler divergence.

The training parameters used for few-shot baselines are reported as follows:

- Siamese Network [109]: we use GloVe embeddings (`embedding size = 100`)

- ProtoNET [228]: `learning rate` is set to 1e-5, Euclidean loss is used and the support set varies depending on the number of shots. In 1-shot training, a support set equal to 1 is necessarily chosen; in 10-shot training we select a support set equal to 5 and this value remains the same in all the other scenarios due to RAM availability constraints.

- BioBERT [124] and ClinicalBERT [10]: same parameters used to train our multi-task networks.

- PET [210]: 5 `epochs` with 250 steps, `learning rate` set to 1e-4, `batch size` to 8, `weight decay` to 0.01. Furthermore, we initialize the weights of the transformer architecture with the biomedical checkpoint publicly made available in [127], which is the same we use for our MT-DNN models.

Note that the number of epochs and the learning rate were selected based on the complexity of the model and the amount of data available, and were determined through appropriate tuning to avoid overfitting,obtaining the best possible model on the validation set. It was observed that as the amount of data increased in few-shot tasks, fewer training epochs were required. To maintain fairness in comparing the results between the different tasks, common evaluation metrics such as F1, recall, and precision were used. The dependence on the number of shots and the initialization of the various networks was mitigated by sampling with 5 different seeds for each shot of training for each task, and initializing the network with these seeds during different trainings. This helps to increase the reliability and generalizability of the results and ensure a fair comparison between the different tasks.

9.3.2 Results

Q1: Comparison with few-shot baselines Tables 9.1, 9.2 and 9.3 report the comparison between our framework and state-of-the-art baselines in terms of precision, recall and F1 scores, respectively.

The results presented in Table 9.1 indicate that ProtoNET yields the highest precision in scenarios with extremely limited training data (1-shot and 10-shot). This method is based on a prototypical network that emphasizes on the representation of each relation type and the calculation of prototypes for each relation type, which enhances precision in relation identification when the training samples are relevant. However, when a slightly larger number of training samples are available, the multi-task learning approach demonstrates superior performance. This is due to the information shared among the three relation extraction tasks and the increased robustness and generalization capability of the model resulting from the larger number of training samples.

Despite its precision in identifying relations, ProtoNET exhibits a low recall as evidenced by the results presented in Table 9.2. The utilization of language models pre-trained with biomedical data as BioBERT and ClinicalBERT, the implementation of prompts in PET, which effectively leverages the knowledge gained by language models, and multi-task approaches that incorporate information from additional tasks may enhance recall and thus make these approaches more suitable for identifying a greater number of relevant relationships. Among these methods, our multi-task learning approach guarantees the highest results in terms of recall scores.

To sum up, our approach consistently produced the best results in 50-shot contexts with regard to precision, recall, and F1. In 10-shot contexts, our approach still achieved the best F1, as shown in Table 9.3, although precision was comparable or slightly lower compared to other baselines. However, our approach excelled in terms of recall, significantly outperforming other methods. This is attributed to the use of data from other tasks, which allowed us to identify a larger number of relevant relationships.

Q2: Effects of multi-task learning The results of utilizing MT-DNN and its extension through knowledge distillation are presented in Table 9.4. It is evident from the table that multi-task learning provides a signif-

Shots	Dataset	Siamese	ProtoNET	ClinicalBERT	BioBERT	PET	Ours
1	DDI-2013	4.96 ± 1.66	**23.42 ± 10.92**	5.80 ± 2.08	6.50 ± 2.11	8.19 ± 1.37	6.97 ± 2.01
	ChemProt	5.63 ± 1.87	**16.03 ± 5.39**	4.19 ± 0.77	4.64 ± 0.62	6.04 ± 1.32	7.57 ± 1.79
	I2B2-2010	3.21 ± 0.75	**14.19 ± 3.06**	1.98 ± 0.85	2.31 ± 0.80	2.55 ± 0.92	1.66 ± 0.79
10	DDI-2013	6.55 ± 0.76	**33.58 ± 4.44**	14.52 ± 1.42	15.04 ± 0.86	15.03 ± 1.45	16.00 ± 0.88
	ChemProt	5.26 ± 0.79	**20.71 ± 7.29**	10.43 ± 1.53	12.73 ± 2.15	11.30 ± 0.56	17.00 ± 0.90
	I2B2-2010	4.79 ± 1.42	**20.37 ± 3.58**	15.15 ± 2.97	14.47 ± 2.77	10.55 ± 5.04	18.39 ± 2.34
50	DDI-2013	11.10 ± 2.74	32.03 ± 4.34	24.12 ± 1.21	27.17 ± 1.83	22.26 ± 2.76	**35.58 ± 4.20**
	ChemProt	7.77 ± 2.20	18.62 ± 2.30	23.04 ± 1.83	27.51 ± 1.92	21.44 ± 1.67	**31.40 ± 1.19**
	I2B2-2010	14.02 ± 2.09	22.45 ± 2.93	25.89 ± 1.50	27.76 ± 3.23	22.55 ± 2.07	**29.82 ± 2.85**

Table 9.1. Comparison of precision scores (mean ± std values across five repetitions) with state-of-the-art baselines in k-shot learning scenarios, $k \in \{1, 10, 50\}$.

Shots	Dataset	Siamese	ProtoNET	ClinicalBERT	BioBERT	PET	Ours
1	DDI-2013	23.31 ± 14.04	5.04 ± 1.90	27.59 ± 8.65	35.92 ± 13.37	**44.17 ± 8.60**	34.58 ± 10.02
	ChemProt	18.06 ± 2.29	4.19 ± 0.85	17.72 ± 3.49	21.24 ± 2.90	23.15 ± 5.51	**32.05 ± 7.95**
	I2B2-2010	**18.42 ± 6.59**	2.73 ± 0.65	11.74 ± 6.99	14.47 ± 1.81	10.78 ± 6.57	6.64 ± 2.90
10	DDI-2013	26.92 ± 4.10	6.82 ± 0.64	60.30 ± 4.13	67.61 ± 6.14	62.96 ± 6.38	**74.22 ± 5.26**
	ChemProt	22.21 ± 3.24	4.99 ± 1.19	42.00 ± 6.04	49.21 ± 8.98	41.72 ± 4.98	**64.41 ± 6.24**
	I2B2-2010	27.20 ± 6.74	3.59 ± 0.50	58.88 ± 6.67	57.49 ± 7.53	61.67 ± 2.61	**68.23 ± 6.67**
50	DDI-2013	40.20 ± 9.77	7.18 ± 1.39	71.48 ± 2.49	78.44 ± 2.22	78.30 ± 2.34	**83.92 ± 2.04**
	ChemProt	29.75 ± 7.30	5.22 ± 0.92	68.35 ± 3.31	76.56 ± 3.86	67.32 ± 5.14	**82.31 ± 2.28**
	I2B2-2010	50.88 ± 3.24	3.92 ± 0.54	77.55 ± 2.12	78.13 ± 1.02	71.67 ± 2.61	**85.21 ± 1.38**

Table 9.2. Comparison of recall scores (mean ± std values across five repetitions) with state-of-the-art baselines in k-shot learning scenarios, $k \in \{1, 10, 50\}$.

icant improvement for the inference task on the ChemProt and I2B2-2010 datasets. However, it results in a decrease in performance when applied to the DDI-2013 dataset. The application of knowledge distillation is advantageous for all downstream tasks but fails to outperform the single-task performance on the DDI-2013 dataset. This phenomenon, referred to as *negative transfer*, will be thoroughly analyzed in research question Q3.

Furthermore, we analyzed the impact of knowledge distillation on the overall performance. In particular, we have performed hyper-parameter tuning on the weighting parameter λ which controls the contribution of ground truth to the knowlege distillation loss as in Eq. 9.3. Specifically, the tuning was conducted using shots 1, 10, and 50, while fixing the network initialization and shot extraction seeds to be the same across experiments with different λ values. The parameters used in these experiments are the

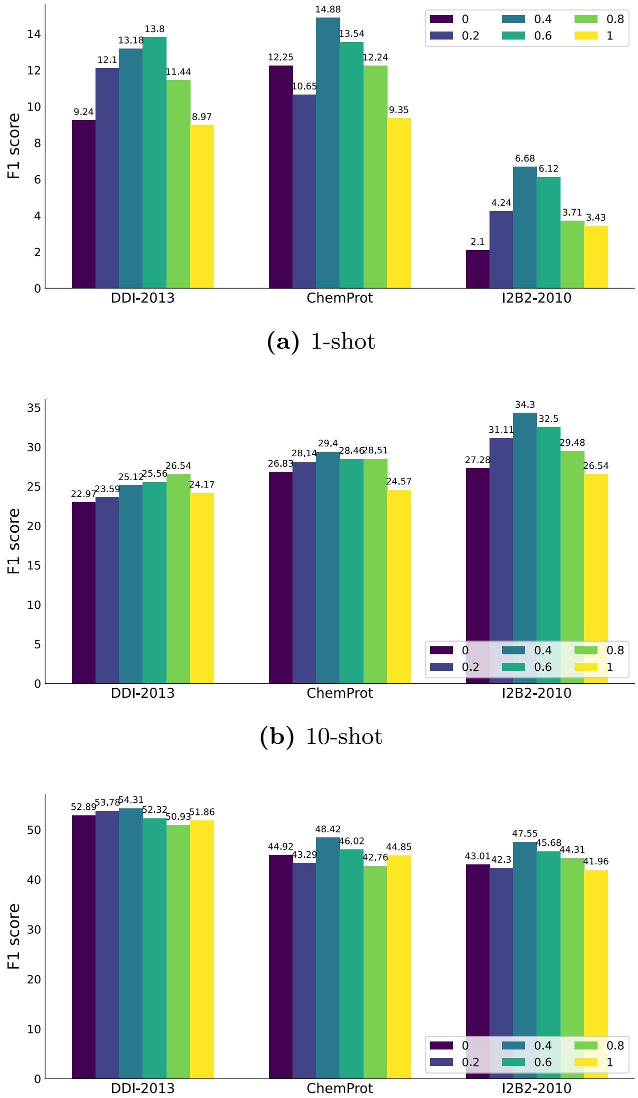
Shots	Dataset	Siamese	ProtoNET	ClinicalBERT	BioBERT	PET	Ours
1	DDI-2013	7.76 ± 3.06	8.20 ± 3.31	9.55 ± 3.22	10.62 ± 3.97	**13.82 ± 2.38**	11.55 ± 3.16
	ChemProt	8.48 ± 2.19	6.66 ± 1.50	6.76 ± 1.18	7.68 ± 0.80	9.51 ± 2.03	**12.06 ± 2.84**
	I2B2-2010	**5.40 ± 1.31**	4.53 ± 0.79	3.38 ± 1.55	3.30 ± 1.12	4.32 ± 1.61	3.17 ± 2.07
10	DDI-2013	10.49 ± 1.08	11.30 ± 0.84	23.34 ± 1.71	24.17 ± 1.57	24.22 ± 2.44	**26.32 ± 1.41**
	ChemProt	8.50 ± 1.25	8.00 ± 2.06	16.71 ± 2.41	20.21 ± 3.44	17.75 ± 0.94	**26.86 ± 1.27**
	I2B2-2010	8.14 ± 2.34	6.10 ± 0.87	24.05 ± 4.15	23.07 ± 3.93	17.52 ± 7.82	**28.92 ± 3.16**
50	DDI-2013	17.36 ± 4.19	12.07 ± 1.09	36.06 ± 1.55	40.34 ± 2.13	34.61 ± 3.50	**49.84 ± 3.90**
	ChemProt	12.38 ± 3.46	8.12 ± 1.19	34.42 ± 2.09	40.46 ± 2.53	32.51 ± 2.42	**45.60 ± 0.74**
	I2B2-2010	20.35 ± 1.62	6.66 ± 0.85	32.63 ± 15.10	40.41 ± 3.88	34.22 ± 2.18	**44.12 ± 3.22**

Table 9.3. Comparison of F1 scores (mean ± std values across five repetitions) with state-of-the-art baselines in k-shot learning scenarios, $k \in \{1, 10, 50\}$.

same as those used in our multi-task few-shot experiments. The λ values used for tuning are: 0, 0.2, 0.4, 0.6, 0.8, and 1. Results in Figure 9.3 show that the best F1 score is achieved with λ values that imply considering both the ground truth and teachers. In particular, the optimal value obtained in every few-shot scenario and with all the datasets — with the only exception of DDI-2013 (10-shot) — is $\lambda = 0.4$, slightly biased towards the teacher's additional knowledge. Hence, the student network can learn from the teacher how to capture more subtle and complex patterns in the data such as uncertainties and correlations between different classes and the nuances and complexities of the language. However, results degrade when the student network relies too heavily on the teachers' predictions.

Q3: Tasks influence analysis We first analyze the three tasks based on their similarities, and then study their mutual influence and effects in the multi-task learning framework used.

Differences in syntax. Initially, a vocabulary was derived from each dataset that encompasses the occurring words. The number of shared words between the tasks is depicted in the pie chart of Figure 9.4. It can be observed that the tasks of DDI-2013 and ChemProt exhibit the highest number of shared words, which is 42.9% of the total vocabulary. Conversely, the words in the I2B2-2010 dataset are distinct from those in the other two datasets, with a similarity of 30.8% and 26.3% compared to ChemProt and DDI-2013, respectively.



(a) 1-shot

(b) 10-shot

(c) 50-shot

Figure 9.3. Impact of the knowledge distillation on F1 scores in few-shot learning scenarios ($k \in \{1, 10, 50\}$. Results with varying loss weight λ. As λ increases, more weight is given to the ground truth instead of relying on teachers' knowledge.

Dataset	**Task**	**Precision**	**Recall**	**F1**
DDI-2013	Single Task	**83.37** ± **0.76**	**80.82** ± **0.66**	**82.07** ± **0.63**
	MT-DNN	83.05 ± 0.65	79.96 ± 0.79	81.47 ± 0.57
	MT-DNN+KD (Klb)	82.86 ± 0.49	79.67 ± 1.09	81.22 ± 0.56
	MT-DNN+KD (MSE)	83.32 ± 0.72	79.86 ± 1.06	81.55 ± 0.66
ChemProt	Single Task	74.41 ± 1.64	74.90 ± 1.81	74.62 ± 0.42
	MT-DNN	75.64 ± 0.74	75.38 ± 0.91	75.25 ± 0.26
	MT-DNN+KD (Klb)	**75.94** ± **0.44**	75.62 ± 0.52	75.75 ± 0.29
	MT-DNN+KD (MSE)	75.61 ± 0.85	**76.31** ± **0.82**	**75.95** ± **0.20**
I2B2-2010	Single Task	75.96 ± 1.78	75.64 ± 4.25	75.68 ± 1.35
	MT-DNN	76.88 ± 0.79	76.59 ± 0.84	76.73 ± 0.35
	MT-DNN+KD (Klb)	77.31 ± 0.70	76.74 ± 0.54	77.02 ± 0.10
	MT-DNN+KD (MSE)	**77.56** ± **0.82**	**76.78** ± **0.77**	**77.17** ± **0.12**

Table 9.4. Comparison of MT-DNN variants with single-task models over the entire training sets (results are reported in terms of mean ± stdDev). We experimented MT-DNN in its original version and with the knowledge distillation procedure described in Section 9.2.1 by using the MSE loss (MT-DNN+KD (MSE)) and the hybrid loss based on Kullback Leibler divergence (MT-DNN+KD (Klb)).

In Figure 9.5, the distributions of sentence lengths are presented, where the sentences are represented as a sequence of words. It is evident that, despite the similarities in median values across the various tasks, DDI-2013 exhibits a substantial quantity of lengthy sentences, with approximately 1000 instances surpassing 600 words. Conversely, sentences in I2B2-2010 tend to be comparatively shorter in comparison to those in other tasks.

Differences in semantics. The semantic similarity between various tasks was determined by computing the similarity between sentence embeddings generated with SentenceBERT [200]. This was achieved by utilizing BlueBERT [190] as the primary encoder. The method involved calculating the cosine similarity score between each sentence from each dataset and all the examples in each dataset, and then averaging the scores to obtain the similarity score between the target sentence and the three datasets. To obtain the similarity scores between datasets D_1 and D_2, the average similarity scores between sentences $s \in D_1$ and dataset D_2 were calculated.

The results presented in Figure 9.6 indicate that I2B2-2010 is the most

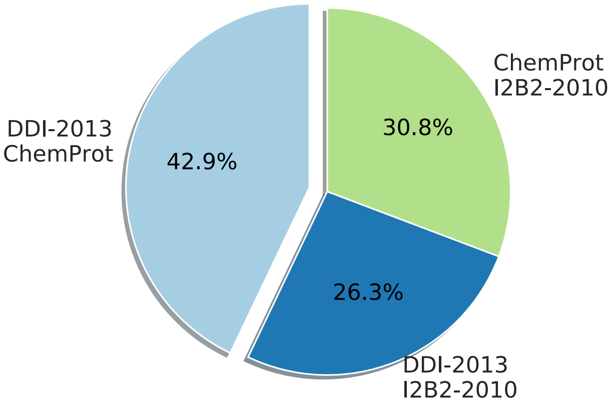


Figure 9.4. Percentage of words shared between pairs of datasets.

heterogeneous dataset, as evidenced by the low similarity score with itself. This is likely due to the fact that the data was collected from eight distinct hospitals. Conversely, ChemProt and DDI-2013 demonstrate a high degree of semantic similarity to each other.

We are interested in understanding the impact of semantic similarity and dissimilarity on performance when considering pairs of tasks. This investigation was conducted through the use of pairwise training [230]. The results presented in Table 9.5 show the scores obtained when multi-task training was performed with the task indexed in the row and the task indexed in the column (single-task performance is reported on the diagonal). The table reveals that while the performance of the other tasks is improved through multi-task training, DDI-2013 experiences a negative transfer, probably due to the absence of long sentences in other datasets, resulting in a decrease in performance compared to the single-task scenario. Additionally, the contributions made by DDI-2013 to the performance improvement of the other tasks are generally inferior compared to those made by the other tasks. On the other hand, the I2B2-2010 task, despite its inherent high variability, benefits the most from multi-task training.

Q4: Few-shot scenarios We examined the impact of multi-task learning on performance in scenarios with varying degrees of data scarcity. To

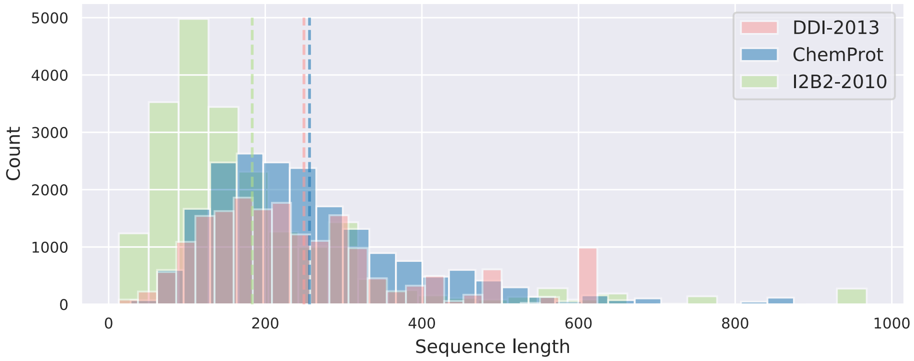


Figure 9.5. Sentence length distributions. Median values are marked with a dotted line.

Task	DDI-2013	ChemProt	I2B2-2010	All
DDI-2013	**82.07**	81.69	81.17	81.473
ChemProt	74.86	74.62	75.07	**75.25**
I2B2-2010	76.52	76.60	75.68	**76.73**

Table 9.5. Pairwise multi-task relationships between datasets. In the first three columns, single-task results are reported on the diagonal and pair-wise multi-task results obtained on the row-indexed dataset are reported when it is used in a multi-task setting with the column-indexed dataset. Multi-task results obtained by using all the datasets of this study are reported in the last column.

accomplish this, we measured the performance of multi-task models as the number of samples (k) increased ($k \in 1, 10, 50, 100, 1000$), and the results are presented in Figure 9.7 in terms of precision, recall, and F1 scores. In contrast to the results obtained in the pairwise experiments as described in Question Q3, we observed a generally positive transfer in performance. Specifically, while the DDI-2013 dataset experienced negative transfer when utilizing the complete training data, we noted a benefit from multi-task learning in low-resource scenarios for all datasets, with relative improvements ranging from 18.3% to 32.4% in F1 scores.

Furthermore, the improvement percentage typically increased as the

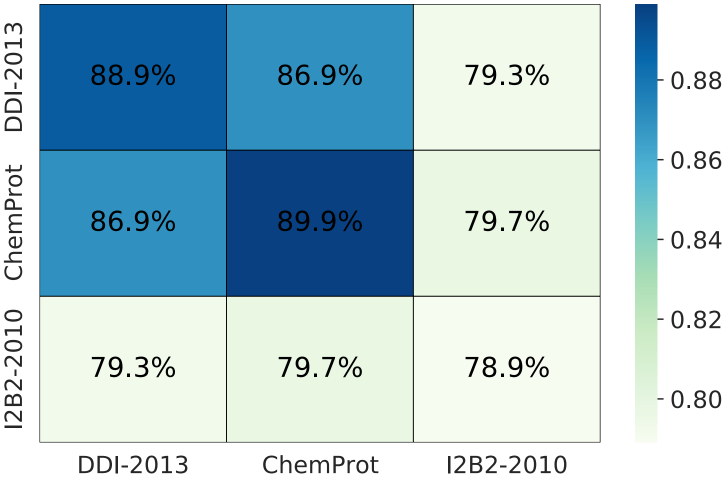


Figure 9.6. Heatmap showing the semantic similarities across tasks.

amount of training data decreased, reaching a maximum of 77.4% in F1 scores on the ChemProt data in the 1-shot scenario. This aligns with previous research [266, 230] that emphasizes the potential benefits of multi-task learning in few-shot learning contexts. Although the improvement in precision scores either remained constant or increased as the number of samples increased, there was a notable decrease in recall scores. This suggests that the advantage of multi-task learning in the few-shot scenarios investigated is mainly due to the improved ability of the trained model to differentiate between true positives and false negatives.

We conducted the pairwise experiment in few-shot learning scenarios to gain a deeper understanding of positive and negative transfer in few-shot scenarios. The results displayed in Figure 9.8 demonstrate that models trained in a pairwise manner have comparable scores to the multi-task models examined in Figure 9.7. The small differences across the pairwise results can be only observed in recall scores, where we can observe small decreases in performance when pairing ChemProt with other datasets. Additionally, the performance of the pairwise models is consistently higher than that of single-task models.

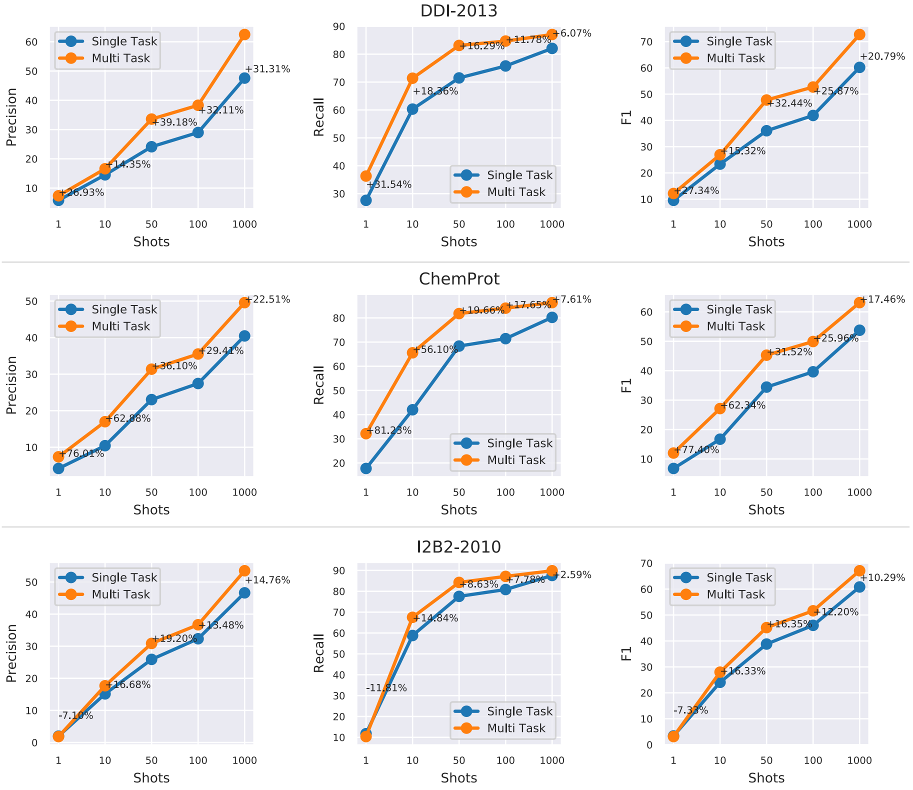


Figure 9.7. Few-shot comparison between single-task and multi-task networks. Performance on the three datasets under analysis (rows) is reported in terms of precision (first column), recall (second column) and F1 (third column). The improvement percentage of the multi-task network w.r.t. the single task network is reported for each k-shot setting.

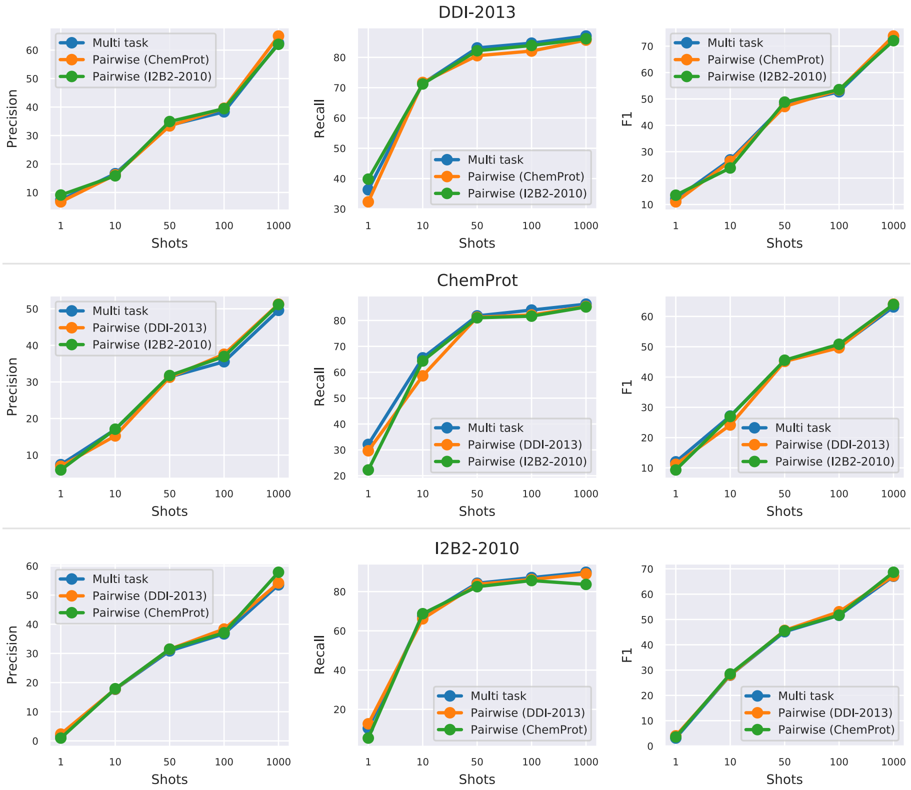


Figure 9.8. Pair-wise experiment in few-shot scenarios. For each dataset (rows), multi-task performance obtained with by using all the dataset is compared with multi-task performance obtained by using only one other dataset. Performance is reported in terms of precision (first column), recall (second column) and F1 (third column)

9.4 Conclusion & Future Work

In the presented chapter, a new framework for few-shot biomedical relation extraction is introduced, leveraging a transformer-based network combined with a multi-task learning method [150]. This method employs a shared layer across biomedical RE tasks and establishes a distinct classification head for each individual task. To optimize performance, a training structure rooted in knowledge distillation is implemented.

An analysis into the elements leading to positive and negative transfer in biomedical relation extraction shows that the introduced framework consistently attains positive transfer in scenarios with limited labeled data for the primary assignment. Furthermore, this strategy consistently outperforms contemporary few-shot learning benchmarks in a majority of tasks and settings, with a notable emphasis on recall metrics, reaching as high as 84% from just 50 training examples. This data indicates the model proficiency in accurately recognizing a significant majority of genuine positive relations.

Nonetheless, the precision metrics reveal potential areas for advancement, particularly in contexts demanding meticulous decision-making. To refine multi-task model precision, it's recommended to integrate supplementary attributes like dictionaries and medical ontologies. These resources offer structured vocabularies and semantic guidelines for discerning relations.

It is critical to underscore that the performance evaluation of the system utilized publicly accessible datasets. This may not provide a complete representation of its efficacy on real-world clinical datasets. As a result, more rigorous examinations using authentic clinical data are required to gauge the system's real-world effectiveness.

Chapter 10

Temporal Knowledge Graphs for Predictive Analysis of Patient Medical Histories

In the digital era, Electronic Health Records (EHRs) have become indispensable for healthcare providers, offering a comprehensive compilation of a patient's health history—encompassing demographics, medications, lab results, and treatment plans. This data repository not only augments the continuity of care and coordination among healthcare providers but also facilitates trend identification and data-driven decision-making to enhance patient care.

Numerous studies have ventured into analyzing the structured information within EHRs to predict potential medical issues [198, 139, 221]. However, a significant portion of EHR data is unstructured, thereby presenting a substantial challenge in extracting pertinent information for effective utilization. Natural Language Processing (NLP) techniques emerge as a solution, demonstrating capabilities in extracting relevant information from unstructured data and associating it with medical ontologies [271, 93, 61].

While transformer architectures excel in identifying intricate temporal patterns within data, their integration with static information—such as that derived from medical ontologies encapsulating medical scientific literature—remains challenging. This static information holds potential for forecasting future disorders. Conversely, Knowledge Graphs (KGs), which have recently shown promise in recommendation systems [256], information retrieval [153], and natural language processing, present an avenue for integrating both dynamic and static information. A traditional *static* KG is a structured knowledge representation utilizing a graph-based data topology to integrate factual information in the form of triples, $\langle s, r, o \rangle$, where s and o denote the subject and object entities, respectively, and r represents the relation between them. Medical ontologies, like SNOMED CT[1] and UMLS [21], are often structured as hierarchies of concepts, allowing for their representation in the form of KGs.

However, the static nature of traditional KGs contrasts with the continuously evolving nature of a patient's health status. This discrepancy underscores the necessity for alternative representations like Temporal Knowledge Graphs (TKGs), which adeptly capture the dynamic health status of patients by extending facts from a triple $\langle s, r, o \rangle$ to a quadruple $\langle s, r, o, t \rangle$, with a timestamp t appended. Consequently, a medical history can be modeled as a TKG, consisting of multiple snapshots that capture the patient's health status at different temporal junctures.

Recent work suggests that incorporating entities' static information, such as their types, enhances the quality of the model's entity representations [141]. In our endeavor, we amalgamate both the dynamic information of medical histories and the static information of medical ontologies in a learning framework, named MedTKG, with an aim to predict future disorders associated with a patient, i.e., the missing objects in the quadruple $\langle s, r, ?, t \rangle$, where s and r denote the patient and the disorder relation type, while t represents the timestamp of the query.

Figure 10.1 shows an example of a medical history and the related disorder diagnosis task. Each timestamp, denoted by t_i, represents a snapshot of the patient's health status at a particular point in time. These timestamps capture all the events that took place during a single day of the patient's hospital stay. The patient is connected to all the concepts extracted from clinical notes that were recorded at that timestamp. To ensure that the model does not predict repetitive or periodic events that have already occurred in the patient's timeline, we only store new facts that have not been previously recorded.

Contrary to existing literature on TKGs, the proposed methodology

[1] https://www.snomed.org

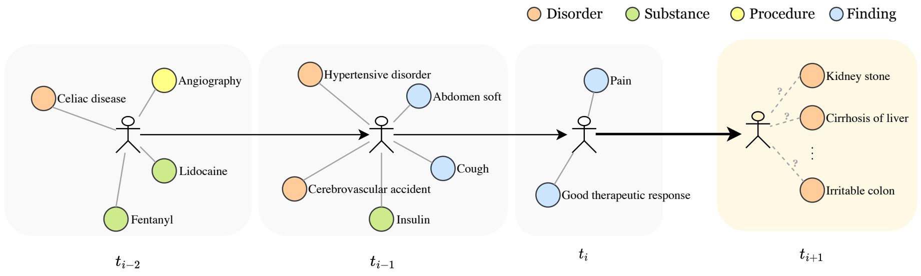


Figure 10.1. Modelling medical histories with Temporal Knowledge Graphs. To prognosticate potential disorders a patient may manifest at a subsequent time point t_{i+1}, we leverage historical data amassed from t_0 to t_i. Within each timestamp, this data is articulated in the form of a knowledge graph, where the patient is represented as a pivotal node interconnected to a myriad of medical concepts encompassing disorders, substances, procedures, and findings.

entails analyzing individual TKGs for each patient within the dataset, wherein a query $\langle s, r, ?, t\rangle$ does not have a unique object as the correct answer, but rather a list of all possible disorders that may transpire in the patient's future. This approach necessitates modifications to the training methodology and a revision of the evaluation metrics, leading us to leverage metrics commonly employed in the field of recommender systems.

Findings illustrate that integrating medical ontologies with the prediction model significantly bolsters its performance, manifesting a $+4.8\%$ relative improvement of mean recall (MR) and mean averaged precision (MAP) scores, among the others. These outcomes furnish valuable insights for ensuing research in the healthcare domain, and hold promise in augmenting the decision-making process for clinicians, ultimately aspiring to enhance patient outcomes.

10.1 Related Work

This section delineates a review of the extant literature pertinent to electronic healthcare record mining for subsequent disorder prediction employing deep learning techniques, and recent advancements in Temporal Knowledge Graphs (TKGs).

10.1.1 Deep learning for disorder forecasting

Predominantly, prior endeavors in prediction or forecasting have engaged structured datasets or structured data encapsulated in EHRs to prognosticate a restricted array of prospective events.

A substantive corpus of research has employed transformer-based models to scrutinize Electronic Health Records (EHRs). For instance, BEHRT [139] engages a subset of 301 disorders inherent in structured EHR data, albeit with a limitation to predicting disorders within a specified, predetermined temporal frame due to the requisite grouping of information by patient visits; its multi-label approach may encounter challenges with escalating numbers of concepts to be predicted. G-BERT [221] utilizes single-visit samples from EHRs, thus circumscribing its aptitude to capture extended contextual information and, akin to BEHRT, only leverages structured data. Med-BERT [198] is trained on structured diagnosis data encoded via the International Classification of Diseases but is evaluated on a narrow subset of disorders, thereby impeding a comprehensive assessment of performance. Conversely, MedGPT [114] exploits unstructured data within clinical narratives, initially executing a Named Entity Recognition and Linking (NER+L) task.

Despite the prowess of transformer-based models in discerning temporal patterns within data, they exhibit a deficiency in performance augmentation through integration with medical ontologies. These ontologies, entrenched in scientific literature, have been illustrated to foster more precise predictions via graph data structures. For instance, GRAM [38] leverages medical ontologies and the attention mechanism to learn robust medical code representations, KAME [157] prognosticates future visit information with medical ontologies, and CompNet [254] discerns correlative and adverse interactions between medicines, factoring in additional medical knowledge such as drug-drug interactions.

10.1.2 Temporal Knowledge Graphs

Temporal Knowledge Graphs (TKGs) augment the conventional Knowledge Graph (KG) representation by infusing time-awareness into event modeling. A variety of TKG methodologies have been proposed including TTransE [122] that enhances the translation-based score function employed in traditional KG embedding techniques [23] with an additional time embedding, HyTE [49] which projects entities and predicates onto a specific time hyperplane, DE-SimplE [69] employing diachronic entity embeddings to represent entities across different timestamps, ATiSE [270] learning time-aware embeddings of entities and predicates as a Gaussian distribution to signify time uncertainty, TeRo [269] extending HyTE by learning time-sensitive entity and predicate embeddings via rotation operations specific to various timestamps, and TComplEx [119] which refines ComplEx by scoring each event through a fourth-order tensor decomposition encapsulating time information. An array of models have been introduced that integrate Graph Neural Networks (GNNs) or Recurrent Neural Networks (RNNs) to discern spatial-temporal patterns, such as RE-NET [98], RE-GCN [141], HIP [78], and EvoKG [186].

10.2 Methodology

In this section, the proposed methodology to tackle the challenge of future disorder prediction is delineated. Our strategy takes inspiration from the work of Li et al. [141], where a Temporal Knowledge Graph (TKG)-based model is devised by contemplating the interrelationships among simultaneous facts, event patterns over time, and intrinsic attributes of entities. We adapt this strategy to our clinical context (i.e., employing clinical notes as input to model medical histories and medical ontologies as the repository of static information for medical concepts) and extend it to accommodate multiple independent TKGs, each representing a patient's medical history.

The architecture of the proposed *MedTKG* model is illustrated in Figure 10.2. Subsequent sections will furnish a detailed exposition of each constituent component.

10.2.1 Inputs

Medical History Commencing with the free text encapsulated in clinical notes, the first step in our architecture is the execution of Named Entity Recognition and Linking (NER+L), aiming to extract mentions of clinical concepts of interest and link them to a medical ontology. In our

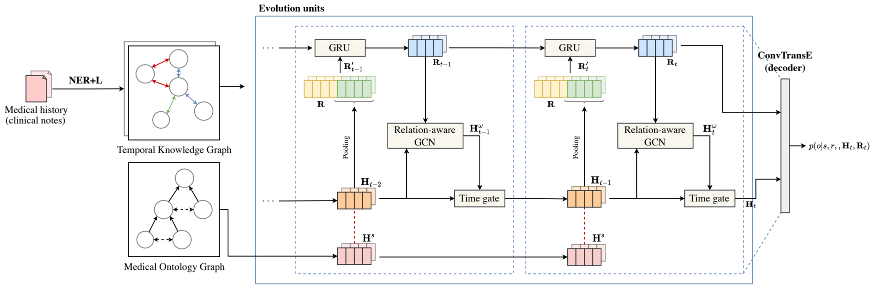


Figure 10.2. A methodological flowchart illustrating the steps involved in processing a single patient's medical history.

experiments, mentions of *disorders*, *procedures*, *substances*, and *findings* were extracted, and linked to the SNOMED-CT ontology employing the Medical Concept Annotation Toolkit (MedCAT) [113]. This module is designed for facile replacement to meet individual requirements.

Subsequent to the extraction of all pertinent medical concepts from a patient's medical history, this knowledge is represented in the guise of TKGs (refer to Section 3.1.1). While entities $\mathscr{V}$ and relations $\mathscr{R}$ are shared across timestamps and TKGs (i.e. the same concept can be present in different medical histories), facts $\mathscr{E}_t$ depend on the patient and the current timestamp. A fact $e \in \mathscr{E}_t$ can be formalized as a quadruple $e = \langle s, r, o, t \rangle$, where $s \in \mathscr{V}$, $o \in \mathscr{V}$, $r \in \mathscr{R}$ and t is the current timestamp. In this work, a fact $e = \langle s, r, o, t \rangle$ indicates that a patient s is related to the medical concept o of type r (e.g. disorder, medical procedure, medical substance) at timestamp t.

Medical Ontology Graph To assimilate the knowledge from the medical ontology into the learning framework, it is requisite to represent it as an Ontology graph, articulated as follows:

> **Definition 10.1 — Medical Ontology Graph**
>
> The ontology graph $\mathscr{G}^s$ is a static knowledge graph that models the knowledge encapsulated in a medical ontology. It is formalized as a graph $\mathscr{G}^s = \langle \mathscr{V}^s, \mathscr{R}^s, \mathscr{E}^s \rangle$, where $\mathscr{V}^s \subset \mathscr{V}$ is the set of concepts included in the ontology, $\mathscr{R}^s$ represents the set of possible relations

> between concepts, and $\mathscr{E}^s$ denotes the edges.

In this study, the SNOMED-CT medical ontology is employed, and the links between medical concepts are leveraged. Specifically, two types of relations are considered: (1) a *direct* relation, delineated as an "is a" relationship between two concepts as represented in the ontology, and (2) an *indirect* relation, delineated as a relationship between two concepts that are not directly linked in the ontology, yet share a common parent, i.e., they are both related to the same medical concept, also with an "is a" relationship.

10.2.2 Evolution unit

The Evolution Unit, grounded on the work of Li et al. [141], encompasses several elements employed to model the temporal dynamics of the patient's health status alongside the static information from a medical ontology. A Relation-aware Graph Convolutional Network (GCN) is deployed to capture the structural dependencies within the knowledge graph (KG) at each timestamp. The temporal evolution of the KG is orchestrated through the amalgamation of two gated recurrent components, specifically, a time-gated recurrent component and a Gated Recurrent Unit (GRU) component. These components facilitate the recurrent computation of the evolutionary representations of entities and relations at each timestamp. Moreover, a static graph constraint component introduces constraints between the static embeddings and the evolutionary embeddings of entities to integrate the static properties of the medical ontology. The objective of the evolution unit is to yield an entity-embedding matrix $\mathbf{H}_i$ for each graph $\mathscr{G}_i$ in the medical history.

In the following, we will describe each module of the evolution unit in detail.

Structural dependencies The structural dependencies among concurrent facts in a knowledge graph are captured to model the associations among the entities through the facts they participate in. Given their well-demonstrated ability to learn from multi-relational graph-structured data [140, 219, 275], a ω-layer relation-aware Graph Convolutional Network (GCN) is employed to model structural dependencies. This approach allows for a comprehensive understanding of the relationships and dependencies within the knowledge graph, which can be used to enhance performance on various knowledge-intensive tasks.

Specifically, given a KG $\mathscr{G}_t \in \mathscr{M}$ at timestamp t and the object entity $o \in \mathscr{V}$ at layer l, its embedding at the next layer $l+1$ is computed under a message-passing framework as shown as follows:

$$\mathbf{h}_{o,t}^{l+1} = \mathrm{RReLu}\Big(\frac{1}{c_0} \sum_{(s,r):\exists(s,r,o)\in\mathscr{E}_t} \mathbf{W}_1^l(\mathbf{h}_{s,t}^l + \mathbf{r}_t) + \mathbf{W}_2^l \mathbf{h}_{o,t}^l\Big), \tag{10.1}$$

where $\mathbf{h}_{o,t}^l$, $\mathbf{h}_{s,t}^l$ and $\mathbf{r}_t$ are the embeddings of the object o, subject s and relation r at layer l and timestamp t, respectively; $\mathbf{W}_1^l$ and $\mathbf{W}_2^l$ are the parameters for aggregating features and self-loop in the l-th layer; $\mathbf{h}_{s,t}^l + \mathbf{r}_t$ implies the translational property between s to o via the relation r; c_o denotes a normalization constant (in-degree of o); the RReLu activation function is from Xu et al. [268].

Historical dynamics Sequential patterns in the medical history are captured by stacking the ω-layer relation-aware GCN. To mitigate the over-smoothing problem [108] and the vanishing gradient problem caused by long medical histories, a time gated recurrent component is applied as in [130]:

$$\mathbf{H}_t = \mathbf{U}_t \otimes \mathbf{H}_t^{\omega} + (1 - \mathbf{U}_t) \otimes \mathbf{H}_{t-1}, \tag{10.2}$$

where $\otimes$ indicates the dot-product operation, $\mathbf{H}_t^{\omega}$ is the output of the final layer of the relation-aware GCN and the time gate $\mathbf{U}_t$ performs the non-linear transformation $\mathbf{U}_t = \sigma(\mathbf{W}_4\mathbf{H}_{t-1} + \mathbf{b})$, σ and $\mathbf{W}_4$denoting the sigmoid function and the weight matrix of the time gate, respectively. By using the time gate component, the entity embedding matrix $\mathbf{H}_t$ is obtained by taking into consideration both the output of the final layer of the relation-aware GCN $\mathbf{H}_t^{\omega}$ and the embedding $\mathbf{H}_{t-1}$ from the previous timestamp.

To capture the sequential patterns of relations, a GRU component is adopted. In particular, given a relation r at timestamp t and its related entities $\mathscr{N}_{r,t} = \{i | (i,r,o,t) \text{ or } (s,r,i,t) \in \mathscr{E}_t\}$, we compute the input for the GRU at timestamp t as the concatenation of (1) the result of a mean pooling operation over the embedding matrix of entities in $\mathscr{N}_{r,t}$, $\text{pooling}(\mathbf{H}_{t-1,\mathscr{N}_{r,t}})$ and (2) the embedding $\mathbf{r} \in \mathbf{R}$ of the relation r:

$$\mathbf{r}'_t = [\text{pooling}(\mathbf{H}_{t-1,\mathscr{N}_{r,t}}); \mathbf{r}] \tag{10.3}$$

Then, the relation embedding matrix is updated via the GRU:

$$\mathbf{R}_t = \text{GRU}(\mathbf{R}_{t-1}, \mathbf{R}'_t), \tag{10.4}$$

where $\mathbf{R}'_t$ contains the $\mathbf{r}'_t$ values for all the relations.

Medical ontology static dependencies

The static embeddings of entities in the medical ontology are obtained through a 1-layer R-GCN [211] without self loops. The update rule is defined as follows, where the ReLU activation function is employed to introduce non-linearity into the model:

$$\mathbf{h}_i^s = \text{ReLU}\Big(\frac{1}{c_i} \sum_{(r^s,j):\exists(i,r^s,j)\in\mathscr{E}^s} \mathbf{W}_{r^s}\mathbf{h}'^s_j\Big), \tag{10.5}$$

where $\mathbf{h}_i^s$ and $\mathbf{h}'^s_j$ denote the i-th and j-th lines of the output matrices $\mathbf{H}^s$ and $\mathbf{H}'^s$, respectively; $\mathbf{W}_{r^s}$ is the relation matrix of r^s and c_i is a normalization constant equal to the number of entities linked to i.

10.2.3 Scoring function

The scoring function is devised to compute the conditional probability delineated in Section 3.5. Particularly, given the medical history $\mathscr{M}_T$, the probability score of candidate triples (s, r, o) is desired, formulated as $p(o|s, r, \mathscr{M}_t) = p(o|s, r, \mathbf{H}_t, \mathbf{R}_t)$ since medical histories are represented with entity and relation embeddings. ConvTransE [219] is employed as a decoder due to its capability of handling multi-relational data in conjunction with GCNs [248]. The scoring function is thus computed as follows:

$$p(o|s, r, \mathbf{H}_t, \mathbf{R}_t) = \sigma\Big(\mathbf{H}_t\text{ConvTransE}(\mathbf{s}_t, \mathbf{r}_t)\Big), \tag{10.6}$$

where $\sigma(\cdot)$ is the sigmoid function while $\mathbf{s}_t \in \mathbf{H}_t$ and $\mathbf{r}_t \in \mathbf{R}_t$ are the embeddings of the subject s and relation r, respectively.

10.2.4 Formulation of the loss function

The loss optimized by our model amalgamates two terms emanating from the entity prediction task (i.e., $\mathscr{L}^e$) and the medical ontology constraint (i.e., $\mathscr{L}^s$). Medical histories corresponding to numerous patients are processed independently by the model. Thus, given M medical histories, the loss function is articulated as follows:

$$\mathscr{L} = \sum_{m=0}^{M-1} \lambda_1 \mathscr{L}^e_m + \lambda_2 \mathscr{L}^s_m, \tag{10.7}$$

where λ_1 and λ_2 are parameters controlling the loss terms. We will describe the two loss terms in the following and omit the medical history identifier m for the sake of simplicity.

Entity prediction loss The entity prediction task is treated as a multi-label learning problem. The loss function is computed as follows:

$$\mathscr{L}^e = \sum_{t=0}^{T-1} \sum_{(s,r,o,t+1)\in\mathscr{E}_{t+1}} \sum_{i=0}^{|\mathscr{V}|-1} y_{t+1,i} \log\, p_i(o|s,r,\mathbf{H}_t,\mathbf{R}_t), \tag{10.8}$$

where T is the length of the medical history, $y_{t+1,i}$ is the i-th element of $\mathbf{y}_{t+1}$ and p_i is the probability score of entity i.

Medical ontology constraint The medical ontology constraint confines the angle between the evolutionary embedding $\mathbf{h}_{i,t}$ and the static embedding $\mathbf{h}^s_i$ of the entity i at timestamp t not to exceed a threshold which increases over time. The loss of the medical ontology constraint component at timestamp t is defined as follows:

$$\mathscr{L}^s_t = \sum_{i=0}^{|\mathscr{V}^s|-1} \max\Big(\cos\theta_t - \cos(\mathbf{h}^s_i, \mathbf{h}_{t,i}), 0\Big) \tag{10.9}$$

Given a medical history of length T, the medical ontology constraint loss is $\mathscr{L}^s = \sum_{t=0}^{T} \mathscr{L}_t^s$.

10.3 Experiments

10.3.1 Experimental setup

EHRs Dataset

The dataset employed for the training and evaluation of our framework is the MIMIC-III dataset [100], curated by the MIT Lab for Computational Physiology. This dataset encapsulates patient data from individuals admitted to the critical care units of Beth Israel Deaconess Medical Center spanning the years 2001 to 2012 and is publicly available for research purposes. The corpus for our analysis was culled from the entirety of unstructured clinical notes, aggregating to 2,083,179 documents encompassing 46,520 patients.

Subsequent to the extraction of medical concepts via MedCAT, a preprocessing regimen was enacted. Initially, infrequent concepts, delineated as those manifesting fewer than 100 instances across the dataset, were expurgated to obviate the potential detriments of rare diseases and possible patient identification. The conserved concepts were then categorized per patient and chronologically structured. To augment the robustness and exhaustiveness of the medical histories, several iterative steps were employed: 1) A biomedical concept was preserved within a patient's medical history if it manifested at least biannually, thus enhancing the precision of our Named Entity Recognition (NER) and Linking (NER+L) instrument albeit potentially diminishing recall; 2) Parent concepts of extant concepts in the timeline, as delineated by the SNOMED ontology, were removed to attenuate noise and excise redundant information; 3) Medical histories were segmented into diurnal intervals, and repetitive concepts within a segment were expunged; 4) Medical histories exhibiting fewer than 10 concepts were precluded from further analysis.

Ultimately, the medical histories were mapped onto the Temporal Knowledge Graph (TKG) structure delineated in Section 3.1.1. While conventional research paradigms employ a singular TKG, typically bifurcated based on the chronological succession of events for training and testing, our

Table 10.1. Statistics of the dataset.

Nodes	train	dev	test
$\lvert\mathscr{V}_{\text{patient}}\rvert$	36,803	1,947	2,027
$\lvert\mathscr{V}_{\text{disorder}}\rvert$	1,376	1,330	1,322
$\lvert\mathscr{V}_{\text{procedure}}\rvert$	34	32	34
$\lvert\mathscr{V}_{\text{finding}}\rvert$	755	689	696
$\lvert\mathscr{V}_{\text{substance}}\rvert$	472	458	449

Facts	train	dev	test
$\lvert\mathscr{E}_{\text{disorder}}\rvert$	911,418	47,647	49,259
$\lvert\mathscr{E}_{\text{procedure}}\rvert$	72,511	3,747	3,896
$\lvert\mathscr{E}_{\text{finding}}\rvert$	596,900	31,844	32,470
$\lvert\mathscr{E}_{\text{substance}}\rvert$	421,551	22,151	22,738

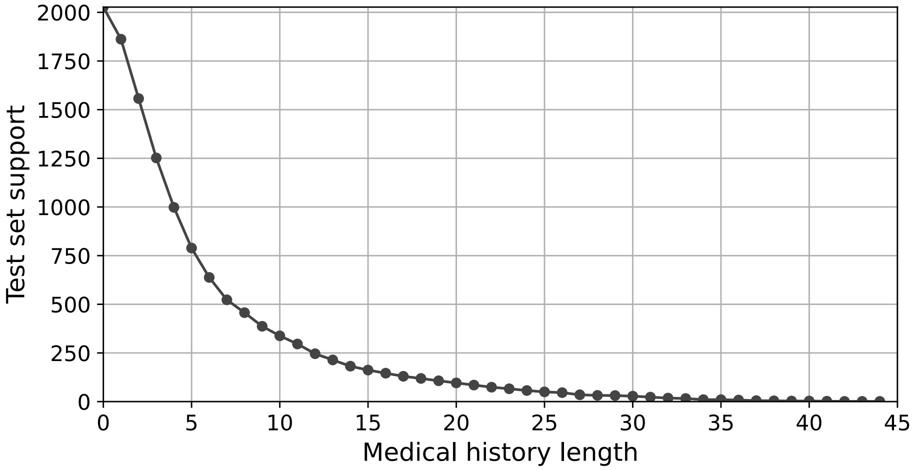
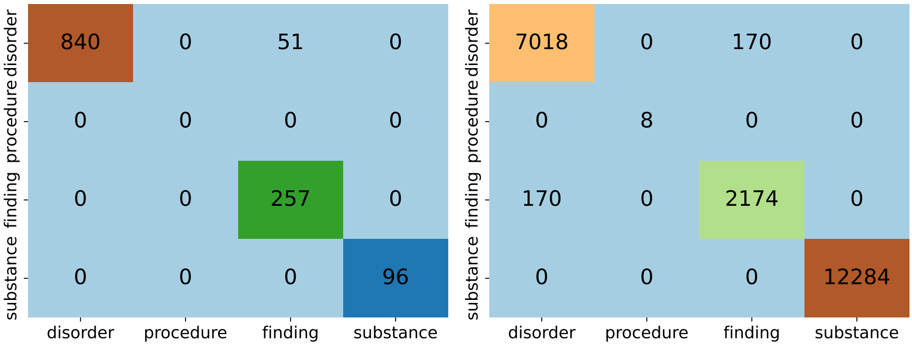


Figure 10.3. Trend of test set support as the length of medical histories increases.

methodology avails distinct medical histories, thereby engendering multiple TKGs. A segment of these TKGs was designated for model training, with the remainder allocated for testing. Specifically, the patient data was apportioned into training, validation, and testing sets, constituting 90%, 5%, and 5% of the dataset, respectively. It's noteworthy that each patient in the test set was further bifurcated into numerous test samples reflecting the entire spectrum of available medical history lengths. The intricacies of the graph data engendered through our study are elucidated in Table 10.1, offering a statistical overview of the dataset. Moreover, Figure 10.3 delineates the trend of diminishing support in the test set concomitant with the augmentation of medical history length, quantified in days.



(a) Direct (is a) **(b)** Indirect (common parent)

Figure 10.4. Relationships found in the medical ontology from source (rows) to destination (columns) concepts.

Medical Ontology

We employed the SNOMED CT ontology to facilitate a systematic mapping between all medical concepts and their corresponding codes. This methodology enabled the identification and analysis of various interrelationships among the encompassed concepts. We considered both direct and indirect relationships. Direct relationships are epitomized by the *is a* relationship between the source and destination concepts, whereas indirect relationships are characterized by a shared *is a* relationship between two concepts and a common ancestor. Figure 10.4 displays heatmaps depicting the quantity of direct and indirect relationships extracted from SNOMED CT.

Metrics

We used the metrics defined in Section 3.5.1 to evaluate the performance of our method: MRR, TP rate, Hits, MR, MAP.

Training parameters

Based on prior investigative endeavors [141] and empirical evidences, the training hyperparameters for the evolution unit were judiciously chosen to ensure optimal performance. The following specifications were adhered to:

- The embedding dimensionality, denoted as d, was designated a value of 200, aligning with conventional practice to ensure a rich representational space.
- The depth of the relation-aware GCN, symbolized as ω, was fixed at 2 layers, aiming to balance between model complexity and the capacity to capture relational dependencies.
- A dropout regularization with a rate of 0.2 was implemented across each layer of the relation-aware GCN to mitigate the potential overfitting phenomenon.
- The Adam optimizer [107], renowned for its efficacy in handling sparse gradients, was employed for parameter optimization with a specified learning rate of 0.0001.
- In the context of the *medical ontology constraint* component, the RGCN was configured with a block dimension of 2×2, and each layer was subjected to dropout regularization at a rate of 0.2.
- Pertaining to the ConvTransE model, the configuration was set to encompass 50 kernels with a kernel size of 2×3, alongside a dropout regularization rate of 0.2, to enhance the model's capability in capturing multi-relational nuances.
- The training regimen spanned across 10 epochs, executed on an NVIDIA A100 GPU, ensuring a comprehensive exploration of the model parameter space.
- Model selection was performed based on the highest Mean Reciprocal Rank (MRR) score attained on the validation dataset, thereby ensuring the generalizability and robustness of the model when transitioned to the testing phase.

This meticulous configuration of training hyperparameters was orchestrated to harmonize the trade-off between model expressiveness and computational efficiency, thereby facilitating precise and reliable prognostications.

10.3.2 Results

In this study, we explore the impact of incorporating static information, sourced from a medical ontology, into our machine learning paradigm. Specifically, we scrutinize the model's performance modulation in response to variations in two facets: (1) the influence of the medical ontology constraint on the learning loss, and (2) the temporal rate of alteration in the angle between evolutionary and static embeddings. Our evaluative procedure is executed employing the metrics delineated in Section **??**, supplemented by an assessment of the quantity of concepts leveraged by the model for prognostication, in tandem with the associated precision scores.

Performance comparison

Table 10.2 delineates the comparative performance outcomes corresponding to diverse medical ontology weights within the framework of the next-disorder prediction task. A discernible trend emanates from the data, indicating that the incorporation of medical ontology elicits consistent enhancements across the examined metric spectrum. Notably, an optimal balance appears to be achieved with a weight value of 0.6, which predominantly yields superior results relative to other weight denominations. This optimum is perceived to arise due to a harmonious interplay between the imposed medical ontology constraint and the network's aptitude for assimilating evolving information, where an excessive weight could potentially thwart the network's learning trajectory. Moreover, the elevated true positive rate and hits values are indicative of a precise model formulation, albeit the modest recall highlights a conceivable challenge stemming from the voluminous concept spectrum considered. A plausible remediation for this caveat might entail the prediction of concepts residing at higher echelons within the SNOMED-CT hierarchy, notwithstanding the consequential decrement in the model's response specificity.

Subsequent to the medical ontology weight impact on the training loss, the model's capacity for synthesizing static and dynamic information is modulated by the rate of augmentation in the threshold angle separating static and dynamic embeddings over progressive timesteps. The insights from Table 10.3 reveal that elevated pace threshold values (≥ 15) are synonymous with enhanced performance in the model's top predictions, owing

Table 10.2. Impact of the medical ontology weight. Best and second-best results are reported in bold and underlined, respectively. The last row (+%) reports the relative improvements obtained when using the medical ontology graph.

		TP rate				Hits				MR				MAP			
Weight	**MRR**	*@1*	*@3*	*@5*	*@10*	*@1*	*@3*	*@5*	*@10*	*@1*	*@3*	*@5*	*@10*	*@1*	*@3*	*@5*	*@10*
1.0	**7.25**	43.27	65.48	74.07	82.36	43.27	37.38	**34.27**	28.97	3.51	8.83	12.94	20.08	3.51	6.82	8.8	11.65
0.8	7.07	41.73	64.73	74.43	82.39	41.73	36.33	33.69	28.62	3.46	8.66	12.97	20.05	3.46	6.69	8.7	11.55
0.6	7.15	43.4	**66.61**	**75.05**	**82.44**	43.4	**37.85**	34.23	28.52	**3.7**	**9.09**	**13.3**	20.13	**3.7**	**7.09**	**9.08**	**11.85**
0.4	7.13	**43.54**	65.34	73.86	81.79	**43.54**	37.13	33.65	28.43	3.44	8.75	12.84	19.99	3.44	6.75	8.7	11.54
0.2	7.02	42.14	64.52	73.71	82.39	42.14	36.26	33.06	28.38	3.46	8.69	12.75	20.17	3.46	6.74	8.64	11.55
0.0	7.21	43.01	65.26	74.35	82.07	43.01	36.92	34.15	**28.98**	3.53	8.75	13.04	**20.23**	3.53	6.82	8.85	11.81
+%	↑ 0.5	↑ 1.2	↑ 2.1	↑ 0.9	↑ 0.4	↑ 1.2	↑ 2.5	↑ 0.3	↓ 0.03	↑ 4.8	↑ 3.9	↑ 2.0	↓ 0.3	↑ 4.8	↑ 4.0	↑ 2.6	↑ 0.3

Table 10.3. Impact of the pace of the angle threshold between dynamic and static embeddings. The weight of the medical ontology graph is set to 1.0. Best and second-best results are reported in bold and underlined, respectively.

		TP rate				Hits				MR				MAP			
Angle	**MRR**	*@1*	*@3*	*@5*	*@10*	*@1*	*@3*	*@5*	*@10*	*@1*	*@3*	*@5*	*@10*	*@1*	*@3*	*@5*	*@10*
1	7.09	42.88	63.74	73.18	81.57	42.88	36.13	32.9	28.03	3.31	8.33	12.3	19.43	3.31	6.44	8.27	11.05
5	7.2	43.15	65.57	**74.98**	**82.94**	43.15	37.36	34.26	**29.16**	3.62	8.8	**13.13**	**20.6**	3.62	6.88	**8.9**	**11.92**
10	**7.25**	43.27	65.48	74.07	82.36	43.27	37.38	**34.27**	28.97	3.51	8.83	12.94	20.08	3.51	6.82	8.8	11.65
15	7.18	43.06	**66.66**	74.86	82.52	43.06	37.1	33.77	28.72	**3.64**	**8.98**	13.01	20.25	**3.64**	**6.93**	8.85	11.72
20	7.19	**43.49**	65.8	73.56	81.71	**43.49**	**38.14**	33.86	28.12	3.5	8.75	12.53	19.51	3.5	6.85	8.68	11.44

Table 10.4. Impact of the medical ontology weight on the number of concepts ever predicted ($CEP@k$, k being the number of top-ranked predictions considered).

Weight	CEP@1	CEP@3	CEP@5	CEP@10
1.0	75	152	208	366
0.8	81	146	222	383
0.6	**103**	**202**	**280**	**460**
0.4	70	146	217	396
0.2	74	144	223	393
0.0	79	165	251	439

to its augmented learning potential from the training dataset, unencumbered by the medical ontology constraint. Inversely, diminished threshold values (≤ 10) foster superior outcomes across a broader spectrum of top predictions, thereby empowering the model to retrieve an expanded array of correct concepts. This dichotomy underscores the intricate balance between the model's learning freedom and the guiding framework provided by the medical ontology, thereby illuminating avenues for fine-tuning the model to achieve a desired performance contour.

Impact on concepts predicted

In Table 10.4, the extent of conceptual diversity captured by models with differing weights allocated to the medical ontology graph is elucidated. The data delineates that the prudent integration of the medical ontology, contingent on an apt weight assignment during the training phase, augments the model's competency in recognizing an expansive array of concepts. Additionally, the validation trajectory illustrated in Figure 10.5 elucidates a stark contrast between the models. The model devoid of the medical ontology's guidance initially discerns a voluminous number of concepts, albeit subsequently succumbs to overfitting, thereby constricting its conceptual scope. Conversely, the employment of the medical ontology enables a methodical assimilation of diverse concepts, thereby progressively enriching the model's conceptual comprehension.

In tandem with evaluating the quantity of concepts prognosticated by the model, a meticulous examination of the precision in predicting these

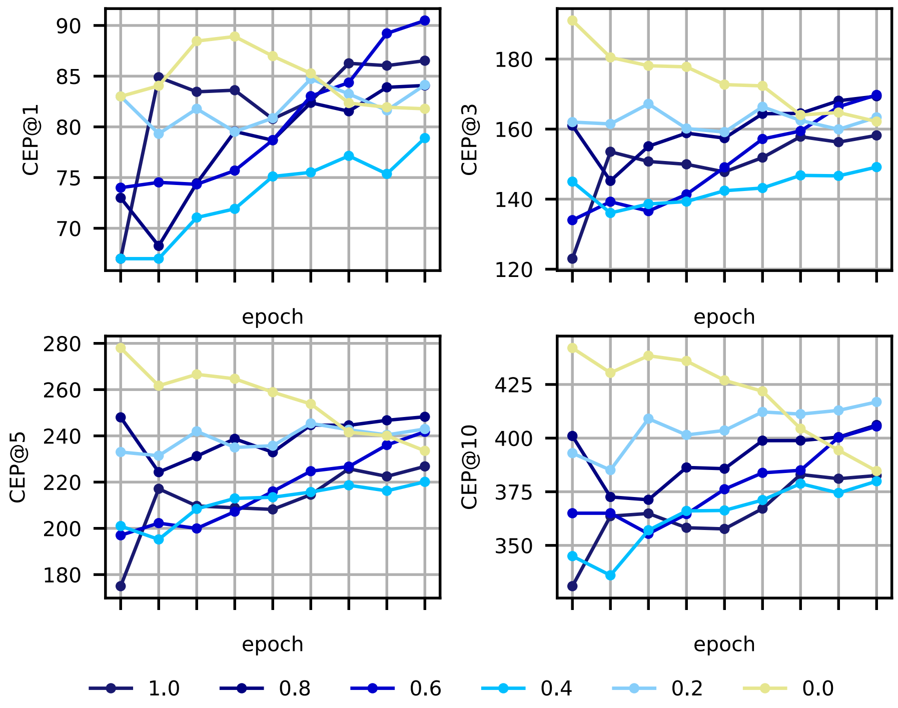


Figure 10.5. Trends of concepts ever predicted (CEP) with different Medical Ontology weights on validation data over training epochs.

concepts was executed, considering the incidence rate of their manifestation among the patient cohort. The elucidated outcomes, delineated in Figure 10.6, substantiate a robust correlation between the model's efficacy and the prevalence of the predicted concepts within the patient populace. As anticipated, the model exhibits augmented performance on concepts that are frequently encountered throughout its training regimen, attributable to their elevated support. Intriguingly, the integration of biomedical ontologies yields superior outcomes, notably for concepts with diminished support (as evidenced in the results ranging between 0.6 and 0.8), underscoring the advantageous impact of employing medical ontologies in such scenarios.

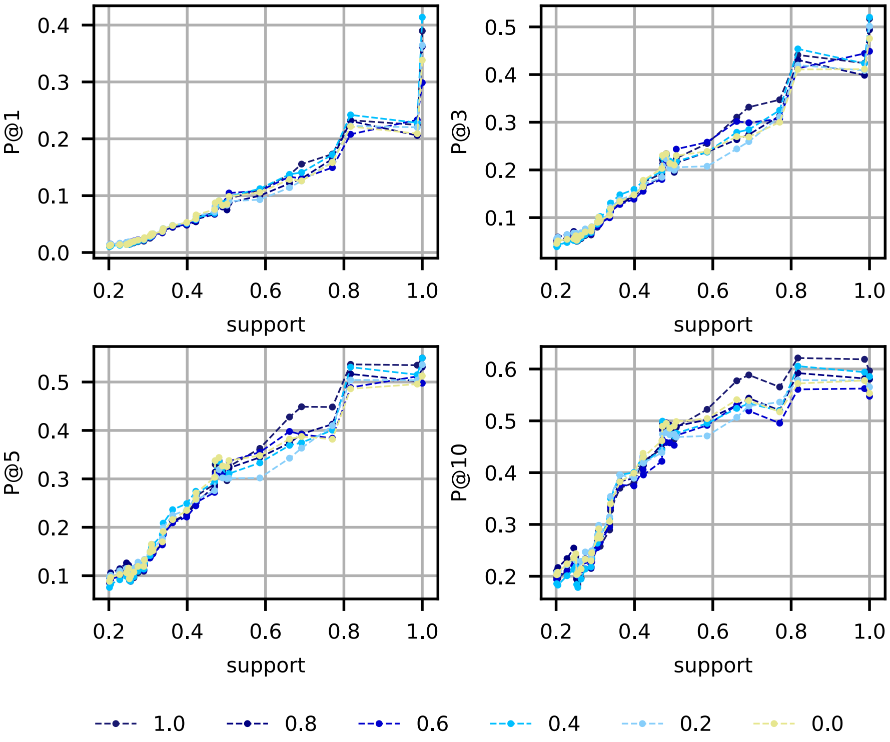


Figure 10.6. Prevalence of concepts in test data (support) vs precision (P@k) of the model in identifying them. We plot only concepts appearing in more than the 20% of patients.

Performance vs medical history length

In Figure 10.7, a nuanced interplay between the length of medical histories and model performance is depicted. Contrary to a monotonic augmentation, a pronounced dip in performance is observed initially, succeeded by an ascent. This trajectory can be elucidated by the relative predictability of initial disorders, which is bolstered by the expansive dataset employed, the circumscribed set of concepts, and the substantial patient cohort exhibiting merely one or two events in their medical annals as illustrated in Figure 10.3. During the phase corresponding to the performance trough,

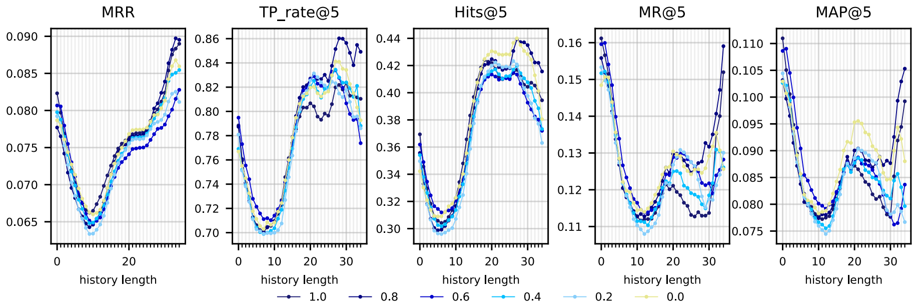


Figure 10.7. Performance trends of models with different weights attributed to the medical ontology constraint as the length of medical histories increases.

the data is scant and the concept spectrum is broad, rendering prediction more challenging. Conversely, in the subsequent phase, the enriched timeline furnishes ample data to facilitate diagnostic inference. A comparative examination of the curves reveals that the incorporation of a medical ontology can significantly enhance the predictive outcomes, particularly for patients with truncated medical histories. This enhancement is ascribed to the model's constrained access to temporal data, compelling it to effectively harness the knowledge encapsulated in the medical ontology for more accurate prognostications.

10.4 Conclusion & Future Work

In this chapter, a novel Temporal Knowledge Graph (TKG) framework, named MedTKG, has been articulated for the prognostication of forthcoming disorders, leveraging both dynamic and static data gleaned from Electronic Health Records (EHRs) and medical ontologies, correspondingly. The empirical evidence garnered suggests a promising avenue in augmenting the predictive prowess of the model vis-à-vis future disorders through the assimilation of medical ontologies. Moreover, a meticulous examination concerning the ramifications of varying parameters, which dictate the weight accorded to the medical ontology during the training phase, was conducted. Looking ahead, we envisage amplifying the breadth of our inquiry by encompassing a diverse array of datasets, inclusive of those articulated in assorted linguistic frameworks, and broadening our predictive horizon to encapsulate an extended gamut of medical events, such as medications and procedural interventions. In a subsequent stride, we are poised to scrutinize the clinical viability of the MedTKG framework through the orchestration of a clinical trial in collaboration with healthcare practitioners, aiming to evaluate its efficacy in enhancing patient-centric outcomes.

Part III

Case study on Italian data

Introduction

In previous sections, the challenges faced during the construction and analysis of a biomedical knowledge graph were explored, especially when faced with limited data and the incorporation of less-resourced languages. This foundation sets the stage for the subsequent chapters that delve into an analysis of individuals accessing healthcare services at a cardiology department in an Italian medical institution, specifically at the hospital of Naples Federico II.

The structure of this case study unfolds as follows: in the upcoming Chapter 11, the dataset will be introduced and the methodology for extracting essential medical events from raw patient histories will be elucidated. Subsequently, an in-depth analysis of the data will be presented in Chapter 12, with an emphasis on predicting potential adverse events.

Chapter 11

Information extraction

11.1 Raw data from the "Campania Salute" network

The dataset employed in this study has been supplied by the Department of Advanced Biomedical Sciences at the University of Naples Federico II. The dataset has been exported from the *Campania Salute (CS)* network [229], an healthcare organization estabilished in 1998 that involves 23 outpatient hypertensive clinics distributed in different community hospitals of the metropolitan area of Naples, 60 randomly selected GPs homogeneously distributed in the same area, and the Hypertension Clinic of the University of Naples Federico II as the co-ordinating centre.

The CS network facilitates the transmission of clinical information acquired during each patient visit between peripheral healthcare facilities, such as general practitioners' offices and community hospitals' hypertension clinics, tasked with overseeing low-risk hypertensive patients, and the coordinating center, primarily responsible for managing high-risk hypertensive patients. The coordinating center collaborates closely with the peripheral healthcare units in the therapeutic interventions and continuous monitoring of all hypertensive patients. This collaborative effort encompasses the evaluation of target organ damage (TOD) and associated diseases

Patient information is shared via online access to a remote database, which is seamlessly integrated through the utilization of smartcards. The central database leverages the *Wincare* software developed by TSD Projects in Milan, Italy. This software comprises distinct sections dedicated to medical history, physical examinations, biochemical profiles, electrocardiographic data, ultrasonography records, additional imaging assessments, and ambulatory blood pressure monitoring. The clinical and personal data of the enrolled patients are subjected to robust digital security measures to ensure confidentiality and data integrity at every access point within the network.

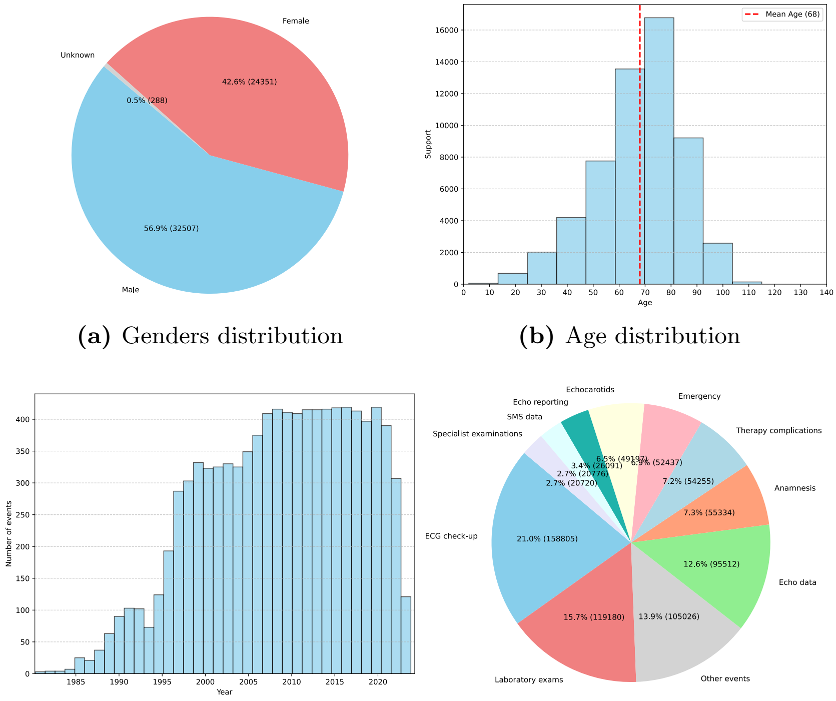
The dataset employed in this study was exported from the CS network in May 2023. It comprises data pertaining to 57,147 individuals for whom events have been documented within the CS network, with records spanning from the year 1980 onward. Figure 11.1a illustrates that 56.9% of the patient cohort comprises males, while 42.6% are females, with the remaining 0.5% gender being unknown. The age distribution of patients is depicted in Figure 11.1b, while Figures 11.1c and 11.1d show the distribution of events over time and the most occurring event types, respectively.

In the remainder of this chapter, the Information Extraction process employed to discern and subsequently scrutinize medical conditions, therapeutic interventions, and diagnostic assessments that constitute an individual's medical record will be described.

11.2 Extraction of medical problems from structured fields

With the collaboration of domain experts affiliated with the University of Naples Federico II (i.e. physicians, medical researchers), who have been utilizing the Wincare database since the establishment of the CS network, we have formulated a set of rules designed to discern the occurrence of medical conditions, therapeutic interventions, or diagnostic procedures within a particular temporal event. In this section, we will furnish comprehensive details concerning the concepts of interest and elucidate our methodology for ascertaining their manifestation in a patient's record at a specific timestamp.

Diabetes is a chronic medical condition characterized by elevated levels of glucose (sugar) in the blood, resulting from the inability of the body



(a) Genders distribution

(b) Age distribution

(c) Events distribution

(d) Most occurring event types

Figure 11.1. Dataset statistics

to effectively produce or utilize insulin. Insulin is a hormone produced by the pancreas that helps regulate blood sugar levels by facilitating the uptake of glucose into cells for energy. There are several types of diabetes, including Type 1 diabetes, Type 2 diabetes, and gestational diabetes, each with distinct causes and characteristics. Wincare contains some boolean fields that can be filled by phsycians by means of checkboxes in their user interface, that allow us also to identify insuline-treated patients and patients who are undergoing oral terapy. Additionally, we associate diabetes with patients exhibiting glycemic levels exceeding the threshold of 126.

Dyslipidemia is a medical condition characterized by abnormal levels of lipids (fats) in the blood. This typically includes elevated levels of cholesterol or triglycerides, or an imbalance in the various types of cholesterol, such as high levels of low-density lipoprotein (LDL) cholesterol, often referred to as "bad" cholesterol, and low levels of high-density lipoprotein (HDL) cholesterol, often called "good" cholesterol. Dyslipidemia is a significant risk factor for cardiovascular diseases, including atherosclerosis, heart attacks, and strokes, as it can lead to the accumulation of fatty deposits in the arteries, narrowing them and impairing blood flow. We extract the following types of dyslipidemia from categorical fields: statins, fibers, hypercholesterolemia, hypertriglyceridemia, mixed dyslipidemia.

Smoking habits refer to the routine of an individual of smoking tobacco products, such as cigarettes, cigars, or pipes. Smoking habits can range from occasional or social smoking to heavy and habitual use, and they can have a significant impact on an individual's health and well-being, as smoking is a major risk factor for various diseases, including lung cancer, cardiovascular diseases, and respiratory conditions. We extract information about smoking habits from a semi-structured field that allows us to distinguish current-smokers, ex-smokers and non-smokers.

Hypertensive disease also known as hypertension or high blood pressure, is a medical condition characterized by consistently elevated blood pressure in the arteries. Blood pressure is the force of blood against the walls of the arteries as the heart pumps it throughout the body. When this pressure remains consistently high, it can put added strain on the heart and blood vessels, potentially leading to serious health problems. Hypertensive disease is a significant risk factor for various cardiovascular problems, including heart disease, stroke, and kidney disease. We detect the presence of hypertensive disease by checking the entrance gate in the Wincare database.

Obesity is a medical condition characterized by an excessive and unhealthy accumulation of body fat to the extent that it can have a negative impact on a person's health. Obesity is associated with an increased risk of various health problems, including heart disease, type 2 diabetes, certain types of cancer, sleep apnea, and joint disorders, among others. It often results from a combination of genetic, environmental, and lifestyle factors, such as poor diet and lack of physical activity. We extract information about obesity by leveraging the body mass index (BMI), which is a measure of weight in relation to height. A BMI of 30 or higher is considered indicative of obesity.

Chronic Kidney Disease (CKD) is a long-term medical condition characterized by the gradual and progressive loss of kidney function over time. The kidneys play a crucial role in filtering waste products and excess fluids from the blood, regulating electrolyte balance, and maintaining overall bodily homeostasis. CKD is typically categorized into stages based on the level of kidney function, with Stage 1 being the mildest and Stage 5 being the most severe. Common causes of CKD include diabetes, high blood pressure, glomerulonephritis, and polycystic kidney disease, among others. As CKD progresses, it can lead to various complications, such as electrolyte imbalances, anemia, bone health problems, and an increased risk of cardiovascular disease. To identify the CKD stage, we estimate the glomerular filtrate rate (GFR) from serum creatinine and other readily available clinical parameters by means of the CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) equation [96] shown as follows:

$$GFR = 141 \cdot \left(\frac{scr}{\kappa}\right)^{\alpha} \cdot \left(\frac{scr}{\kappa}\right)^{-1.209} \cdot 0.993^{Age} \cdot 1.018\,[\text{if female}] \cdot 1.159\,[\text{if black}] \tag{11.1}$$

where scr is serum creatinine (mg/dL), κ is 0.7 for females and 0.9 for males, α is -0.329 for females and -0.411 for males.

Coronary Artery Disease (CAD) is a medical condition characterized by the narrowing or blockage of the coronary arteries, which are the blood vessels responsible for supplying oxygen and nutrients to the heart muscle. This narrowing occurs due to the buildup of fatty deposits and plaque on the artery walls, a process known as atherosclerosis. Coronary Artery Disease can lead to various cardiac events, including acute myocardial infarction, commonly referred to as a heart attack. It occurs when a coronary artery becomes severely blocked, leading to a lack of blood flow to a part of the heart muscle, which can result in tissue damage or cell death due to insufficient oxygen. Treatment options for CAD include:

- *PCI (Percutaneous Coronary Intervention).* This is a minimally invasive procedure in which a catheter with a balloon at its tip is used to open the blocked coronary artery. Often, a stent (a tiny mesh tube) is placed at the site of the blockage to help keep the artery open and improve blood flow.

- *CABG (Coronary Artery Bypass Grafting).* CABG is a surgical procedure in which a surgeon creates bypasses or detours around the blocked coronary arteries using blood vessels from other parts of the body. This allows blood to bypass the blockages and reach the heart muscle, restoring blood supply.

We are able to identify the presence of CAD, previous acute myocardial infarction or PCI and CAD by means of boolean values in the Wincare database.

CAD family history plays a significant role due to genetic factors, shared lifestyle and shared environmental factors. Knowing the family history of CAD can be crucial for early detection and prevention. We are able to identify this information in boolean fields of the database.

Atrial fibrillation (AFib) is a common heart rhythm disorder characterized by irregular and often rapid electrical signals in the upper chambers (atria) of the heart. Instead of contracting regularly and effectively to pump blood into the lower chambers (ventricles), the atria in AFib quiver or fibrillate, leading to an irregular heartbeat. This irregular heart rhythm can disrupt the normal flow of blood in the heart, potentially causing blood to pool and form clots. If these clots travel to other parts of the body, they can block blood vessels and lead to serious complications, such as strokes or heart attacks. Atrial fibrillation can occur for various reasons, including underlying heart conditions, high blood pressure, obesity, and other medical factors. We extract information about AFib in semi-structured fields of the database.

Stroke is a sudden and often serious medical condition that occurs when there is a disruption in the blood supply to a part of the brain. This disruption can be caused by a blocked blood vessel (ischemic stroke) or the rupture of a blood vessel (hemorrhagic stroke). Ischemic strokes are more common and result from the blockage of an artery in the brain, typically due to a blood clot or plaque buildup. Hemorrhagic strokes occur when a blood vessel in the brain ruptures, leading to bleeding within or around the brain. We identify the presence of a previous stroke by means of a boolean value in the database.

Heart Failure is a medical condition in which the heart is unable to pump blood efficiently to meet the body's needs. This occurs when the heart's ability to contract and/or relax is impaired, preventing it from effectively circulating oxygen and nutrient-rich blood to the body's organs and tissues. Heart failure can result from various underlying causes, including coronary artery disease, high blood pressure, heart valve disorders, and cardiomyopathy. It is characterized by symptoms such as shortness of breath, fatigue, fluid retention (edema), and reduced exercise tolerance. These symptoms can range from mild to severe and can significantly impact an individual's quality of life. To identify heart failure within our database, we consider the ejection fraction (EF) measure, which is used to assess the efficiency of the heart's pumping function. It represents the percentage of blood that is pumped out of the left ventricle of the heart (the main pumping chamber) with each heartbeat. In the context of heart failure, an ejection fraction value below 40% is used as a diagnostic criterion.

Aortic disease refers to a group of medical conditions that affect the aorta, the largest and main artery in the human body that carries oxygen-rich blood from the heart to the rest of the body. We are interested in identifying the presence of *aortic dilation*, *aortic aneurysm* and *abdominal aortic aneurysm* in patients from our database. Details are provided as follows:

- *Aortic dilation*: it refers to the abnormal widening or enlargement of the aorta. This can occur in different segments of the aorta, including the ascending aorta (the portion leaving the heart) and the descending aorta (the portion extending down the chest and abdomen). Aortic dilation is often a precursor to more serious conditions like aortic aneurysms or dissections. Based on findings from Evangelista et al.

[57], we identify aorta dilation when the diameter of the aorta is greater than 40 mm in male adults and 34 mm in female adults.

- *Aortic aneurysm*: it is characterized by a localized and abnormal enlargement or bulging of the aorta, which is the main artery in the body that carries oxygen-rich blood from the heart to the rest of the body. This enlargement typically occurs in a weakened area of the aortic wall, causing the aorta to balloon outward. Based on findings from Evangelista et al. [57], we identify the presence of aortic aneurysm when the diameter of the ascending aorta is greater than 45 mm.

- *Abdominal aortic aneurysm (AAA)*: it is a specific type of aortic disease that involves the formation of a weakened and bulging area in the abdominal segment of the aorta. Based on the definition from Erbel et al. [162], we identify AAA when the diameter of the abdominal aorta is greater than 30 mm.

11.3 Dealing with free-text

Electronic Health Records typically comprise a combination of structured and free-text fields. While structured fields are designed to capture specific information such as patient demographics, lab results, and medical diagnoses, free-text fields allow healthcare providers to input information in a more flexible manner. This flexibility, however, can lead to inconsistencies and challenges in extracting relevant data. As a matter of fact, physicians, depending on their habits and preferences, might choose to enter critical patient characteristics in free-text notes even when some structured fields had been designed to contain such kind of information. This variability is a significant hurdle in data extraction as it necessitates methods to harmonize and reconcile disparate data sources.

To address this challenge, we engaged in collaborative efforts with medical experts to establish a comprehensive nomenclature system encompassing a set of aliases for each concept of interest. Subsequently, we employed string-matching algorithms in tandem with the creation of meticulously annotated datasets designed for Named Entity Recognition, Assertion Classification, and Entity Linking. We further harnessed the capabilities of transformer-based models through extensive training to proficiently tackle these multifaceted tasks. Comprehensive details about the extraction process from unstructured fields will be provided in the remainder of this section.

11.3.1 Medical Terminology

With the help of physicians who have been using the Wincare database since the establishment of the Campania Salute network, we have designed a medical terminology that provides a set of aliases for each concept healthcare providers are interesting to when providing their diagnoses or analyses. The terminology is reported in Table 11.1.

When analyzing a clinical note, the medical terminology is utilized within a string-matching algorithm to detect potential medical problems associated with the patient. It is crucial to emphasize that the mere identification of a pattern within the clinical note does not inherently indicate the presence of a medical problem. For instance, it is conceivable that the physician has simply documented that the patient negates to being a smoker, which, in fact, signifies the absence of the problem rather than its presence. Hence, these candidate mentions necessitate additional scrutiny through an *assertion classifier*, a NLP system designed to differentiate between present and absent medical problems. To develop such a classifier, we have meticulously annotated an extensive dataset, which will be described in the subsequent sections.

11.3.2 Datasets annotation

Aiming to extract further information from clinical narratives, our objective is twofold: firstly, to discern the definitive existence of medical conditions within a patient's records, and secondly, to establish a unique identifier with a biomedical ontology, such as the Unified Medical Language System (UMLS). To achieve these aims, we have curated annotated datasets tailored for the tasks of Named Entity Recognition, Assertion Classification, and Entity Linking, respectively.

Table 11.1. Medical terminology. A comprehensive list of aliases is associated with each specific concept of interest. Note that the aliases are presented in the Italian language due to the linguistic context of our clinical notes.

Medical concept	**Aliases**
Diabetes	Diabete, diabete mellito, Diabetico, Diabetica, DM, IDDM
Dyslipidemia	Iperlipidemia, dislipidemia, ipercolesterolemia familiare
Dyslipidemia (hypercholesterolemia)	ipercolesterolemia, colesterolo elevato
Dyslipidemia (hypertriglyceridemia)	ipertrigliceridemia
Dyslipidemia (mixed)	dislipidemia mista
Smoker	Fumatrice, fumatore, Fumo di sigaretta, Abitudine tabagica, tabagista, tabagismo
Ex smoker	ex Fumatrice, ex fumatore, ex-fumatore, ex-fumatrice, pregressa abitudine tabagica, ex tabagista, ex-tabagista, ex tabagismo, pregresso tabagismo
Hypertensive disease	ipertensione, Ipertensione arteriosa, Iperteso, ipertesa, ipertensione in terapia, elevati valori pressori
Obesity	Obesità, obeso, obesa, sovrappeso, BMI elevato
Chronic Kidney Disease	IRC, CKD, insufficienza renale, alterata funzionalità renale, insufficienza renale cronica
CKD (stage 5)	dialisi, terapia dialitica
Coronary Artery Disease (CAD)	Malattia aterosclerotica coronarica, CAD, angina stabile, malattia coronarica, malattia multivasale, Diffusa malattia aterosclerotica coronarica
CAD family history	Familiarità per CAD, Fratello deceduto con CAD, Sorella deceduta con CAD, Madre deceduta per CAD, Padre deceduto per CAD, familiarità per IMA, familiarità per infarto, Famigliarità per CAD, Anamnesi positiva per famigliarità, famigliarità per IMA, famigliarità per infarto, Anamnesi familiare positiva per cad
Chronic Obstructive Pulmonary Disease (COPD)	BPCO, insufficienza respiratoria, COPD, bronchite cronica, enfisema polmonare, ossigenoterapia
Atrial fibrillation	fibrillazione, FA, Fibrillazione atriale, fibr.atriale paros., F.A. PAROSSISTICA, Aritmia (FA), fibrillazione atriale, FA permanente, FA parossistica, F.A., Parossismi di fibrillazione atriale, AFib, FibA
Previous acute myocardial infarction	SCA, IMA, SCA STEMI, IMA non Q, STEMI inferiore, NSTEMI, SCA NSTEMI, IM grave, STEMI inferior, STEMI anteriore, IMA NSTEMI, INFARTO MIOCARDICO, NSTEMI, NSTE-ACS, Sindrome coronarica acuta, infarto miocardico acuto, infarto del miocardio
Previous percutaneous coronary intervention (PCI)	PCI, Angioplastica coronarica, Rivascolarizzazione miocardica, PTCA + STENT SU CX, ptca + stent IVA, PTCA, PTCA IVA e CX, PTCA + STENT, PTCA (IVA), Coronarografia + PTCA, PTCA con stent su IVA, CABG, PTCA+stent, Rivascolarizzazione chirurgica, STENTING C. SX, PTCA per restenosi, impianto di stent coronarico, impianto di DES, impianto di BMS, Rivascolarizzazione coronarica
Previous coronary artery bypass graft (CABG) surgery	CABG, TRIPLICE BYPASS, BY-PASS aorto coronarico, rivascolarizzazione chirurgica, BY pass, AMIS
Stroke	TIA, ICTUS emorragico, emorragia subaracnoidea, emorragia cerebrale, emorragia subaracnoi, Ictus, ICTUS EMORRAGICO, Emorragia subdurale, ischemia cerebrale, ICTUS, Ictus cerebri, TIA/Ictus, Ictus Cerebrale emorragico, Ictus emorragico, Stroke, Accidente cerebrovascolare, CVA, ictus embolico
Peripheral revascularization	Endoarterectomia, TEA car. Dx, stent carotideo, endoarteriectomia, TEA carotide dx, STENT Carotideo, PTCA carotide in. DX, PTCA carotide, PTCA CAR. INT. SX, PTA ICA DX, TEA RICA, TEA, PTA Carotide SX, TEA carotide, interv carotid dx, TEA sx, tromboendoarterectomia carotidea sx, TEA su ICA, endoarteriectomia ICA sn, safenectomia dx, safenectomia, embolectomia art inf dx, STENT AA ILIACA, PTCA FEMORALE SUP DX, safenectomia sn, PTCA + stent arteria iliaca com.dx, PTCA + STent iliaca destra, Angioplastica Iliaca, PTA femorale superficiale sin e dx, Endoprotesi per aneurisma dell' arteria iliaca, PTA periferica, rivascolarizzazione periferica, PTA BTK, Bypass arto-femorale, bypass aorto-bifemorale, PTA renale, stenting renale, stent renale, angioplastica renale
Aortic disease	AAA, aneurismectomia, aneurisma aorta addominale, endoprotesi AAA, Aneurisma aorta, Aneurisma AO addominale, Aneurisma Add, Endoprotesi AAA, TEVAR, EVAR, Bentall, Endoprotesi aortica, dissezione aortica, sindrome aortica acuta, rottura aneurisma aortico
Heart failure	sub edema polmonare, scompenso cardiaco, EPA, TEP, edema polmonare acuto, Scompenso, SubEPA, Edema polmonare, HFrEF, disfunzione ventricolare sinistra, scompenso destro
Peripheral artery disease	Arteriopatia obliterante arti inferiori, AOCP, PAD, arteriopatia arti inferiori, arto critico, arto ischemico, Fontaine, occlusione femorale, occlusione carotide, occlusione iliaco, occlusione iliaca, occlusione iliaco-femorale, occlusione popliteo, occlusione vertebrale, occlusione femoro-popliteo, stenosi femorale, stenosi carotide, stenosi iliaco, stenosi iliaca, stenosi iliaco-femorale, stenosi popliteo, stenosi vertebrale, stenosi femoro-popliteo, malattia femorale, malattia carotide, malattia iliaco, malattia iliaca, malattia iliaco-femorale, malattia popliteo, malattia vertebrale, malattia femoro-popliteo, ostruzione femorale, ostruzione carotide, ostruzione iliaco, ostruzione iliaca, ostruzione iliaco-femorale, ostruzione popliteo, ostruzione vertebrale, ostruzione femoro-popliteo, aterosclerosi femorale, aterosclerosi carotidea, aterosclerosi iliaco, aterosclerosi iliaca, aterosclerosi iliaco-femorale, aterosclerosi popliteo, aterosclerosi vertebrale, aterosclerosi femoro-popliteo
Aortic/mitral valve replacement	sost. valvola aortica, Sostituzione valvolare (mitro-aortica), Sost. Valv. AO, Sostituzione valvolare mitralica, applacazione di protesi aortica biologica, Sot. valv. Aortica, Sost. Valv, Sostituzione valvolare aortica, Sostituzione valvolare mitralica, Valvuloplastica mitralica, SVAo, Protesi valv.mitralica, TAVI, SAVR, Portatore di protesi meccanica, portatore di protesi biologica, portatrice di protesi meccanica, portatrice di protesi meccanica
Implantable cardioverter defibrillator (ICD)	ICD, impianto di ICD, Defibrillatore, impianto di defibrillatore, icd
Pace-maker (PM)	Impianto Pace-Maker, PMK, Pacemaker, impianto PMK, pace-maker
CRT-d implant	CRT-D, CRT, impianto di CRT, impianto di CRTd, CRTd

Data preparation

We collected a total of 227,721 documents by extracting all the unstructured text written by physicians or nurses during hospital admissions. We retained documents with at least 10 tokens. After this filtering step, we were left with 175,891 documents.

With the aim to keep as much information as possible from the available data, the first set of documents to be annotated has been extracted by clustering all the documents in 2000 groups and retaining their centroids. Specifically, (1) we used a transformer-based model pre-trained on Italian corpora [215] to compute document embeddings, each consisting in 768 numerical attributes; (2) then, we reduced their dimensionality to 50 attributes by PCA, explaining the 92.8% of the variance in the original corpus; (3) finally, we applied KMeans clustering and retained the nearest documents to cluster centroids as the unannotated starting corpus.

Annotators and annotation process

Each of the selected documents has been manually annotated independently by two annotators for each of the downstream tasks. Each document has been annotated by two annotators picked at random from a team of twenty-eight annotators with backgrounds in biomedical informatics, and contrasting annotations were then discussed to obtain a consensus.

Due to the high number of documents and the practical infeasibility to handle the whole set, we have followed a strategy based on Active Learning (AL) [202]: starting from an initial set of 736 samples, a transformer-based model has been trained and iteratively used to choose the most informative samples to be annotated from the available pool of unlabeled documents according to a bayesian-active-learning uncertainty measure [66, 13]. We performed a total of 6 iterations, obtaining a final set of 1578 annotated documents. Active learning has been applied only for the annotations regarding the NER task: AC and EL samples were then selected from the same NER dataset.

Annotation guidelines

We applied the same annotation guidelines as Uzuner et al. [247][1]. Specifically, we annotated mentions referring to three categories:

- *Medical problems*: observations made about a patient's body or mind that are believed to be abnormal or the result of a sickness.
- *Treatments*: words used to describe actions taken, interventions made, and medications administered to a patient.
- *Tests*: phrases used to describe actions taken on a patient or a sample of body fluid or tissue in order to identify, exclude, or learn more about a medical problem.

After having identified candidate entity mentions, we are interested in the following categories of assertions:

- *Present* (default category): the medical problem is associated with the patient.
- *Absent*: the medical problem is not associated with the patient.
- *Possible*: the patient may have the medical problem, but the note expresses uncertainty.
- *Conditional*: The patient only experiences the medical problem under specific circumstances (e.g. allergies).
- *Hypothetical*: the patient may develop the medical problem.
- *Not associated with patient (N/A)*: the medical problem is connected to a person other than the patient (e.g. family history).

Note that the AC dataset annotated with these categories is highly imbalanced, thus leading to poor results. For this reason, the dataset has been reviewed with the help of medical experts and simplified for the classification of *Present*, *Absent* and *Famili History*-related mentions.

Finally, we use the Unified Medical Language System (UMLS) metathesaurus to link entity mentions to unique identifiers.

[1] Annotation guidelines for NER, AC and RE from the i2b2-2010 challenge: https://www.i2b2.org/NLP/Relations/Documentation.php

Table 11.2. NER dataset statistics. Number of mentions per entity type, total number of sentences per split.

Entity type	Training	Validation	Test
Medical problem	3818	766	820
Treatment	1188	211	228
Test	1491	333	304
Documents	1104	237	237

11.3.3 NLP models

We compare several transformer-based architectures on all the downstream tasks. Specifically, we the experimented models are listed below:

- *BERT multilingual (cased)* [52]: pretrained BERT model on the top 104 languages with a Wikipedia dump using a masked language modeling (MLM) objective. This model is case sensitive.

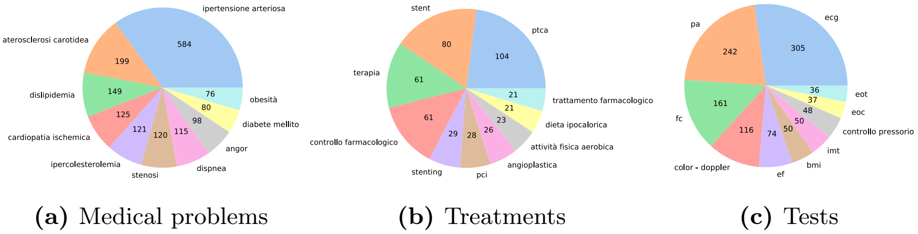
- *XLM-RoBERTa* [40]: XLM-RoBERTa model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages

- *BERT Italian (cased)*: pretrained BERT model on an Italian Wikipedia dump and various text from the OPUS [241] and OSCAR [231] corpora.

We train all our models for 30 `epochs` with a `learning rate` of $5 \cdot 10^{-5}$, an AdamW `optimizer` [154], a `batch size` of 8 and a `maximum sequence length` of 512. We evaluate the quality of methods in terms of precision, recall and f1 scores.

11.3.4 Named Entity Recognition

We divide NER training data in training, validation and test splits with a 70-15-15 ratio. Table 11.2 presents details about the resulting datasets, while Figure 11.2 shows the most occurring mentions for each entity type, separately.



(a) Medical problems **(b)** Treatments **(c)** Tests

Figure 11.2. Most occurring mentions of (a) medical problems, (b) treatments and (c) tests in our NER data.

Table 11.3. NER results

	MedicalProblems			*Treatments*			*Tests*		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT multilingual (cased)	0.859	0.896	0.877	0.724	0.663	0.692	0.790	0.778	0.784
XLM-RoBERTa	0.895	0.918	0.906	**0.802**	0.766	0.784	**0.823**	0.812	0.818
BERT Italian (cased)	**0.896**	**0.919**	**0.907**	0.790	**0.827**	**0.808**	0.811	**0.835**	**0.823**

Results The results presented in Table 11.3 depict the performance achieved by our baseline models when recognizing diverse entity types. The findings indicate a consistent high performance of the BERT model pretrained on Italian corpora, across all three entity categories. Conversely, the XLM-RoBERTa model exhibits superior precision in the identification of treatments and tests. This observation suggests that XLM-RoBERTa excels in recognizing mentions with reduced false positives.

11.3.5 Assertion Classification

The datasets annotated for Assertion Classification are presented in Tables 11.4 (original) and 11.5 (reviewed).

The original dataset is too imbalanced and did not reflect the needs of medical experts for the classification of assertions. For this reason, we revised the dataset to get sufficiently high performance while meeting physicians' needs. For the sake of completeness, we will provide results obtained both on the original and reviewed datasets.

Table 11.4. Assertion Classification dataset statistics (original). Number of samples per assertion class, total number of sentences per split.

Class	Training	Validation	Test
Present	3733	716	744
Absent	511	96	95
Possible	30	6	8
Conditional	67	16	18
Hypothetical	17	4	5
N/A	24	4	3
Sentences	994	213	214

Table 11.5. Assertion Classification dataset statistics (reviewed). Number of samples per assertion class, total number of sentences per split.

Class	Training	Validation	Test
Present	5674	702	709
Absent	657	81	82
Family History	151	19	19
Sentences	994	213	214

Results Table 11.6 reports results obtained on the original AC dataset. While the performance obtained on the *Present* label is very high, it is not sufficient on all the other labels. While we could have dealt with this through imbalance learning techniques, since the medical experts we have been collaborating were extremely insterested just in recognizing the presence, absence or the family history of a medical problem, we have decided to rather revise the dataset to include only these three labels. Results referring to the revised dataset are reported in Table 11.7.

11.3.6 Entity Linking

Every term annotated for NER has been further processed to link mentions to the UMLS terminology. We provide a chord diagram of co-occurring entities in the same sentences in Figure 11.3.

All the entities reported in our training data form a Lexicon that is used whenever we need to link a mention to UMLS. Specifically, we re-

Table 11.6. AC (original) results

Model	**Present**	**Absent**	**Possible**	**Conditional**	**Hypothetical**	**N/A**
			Precision			
BERT multilingual (cased)	0.925	0.606	**0.250**	0.105	0.000	0.250
XLM-RoBERTa	0.926	**0.690**	0.000	**0.167**	0.000	**0.500**
BERT Italian (cased)	**0.933**	0.630	0.200	0.158	0.000	0.333
			Recall			
BERT multilingual (cased)	0.935	0.600	**0.125**	0.111	0.000	**0.333**
XLM-RoBERTa	**0.957**	0.632	0.000	0.111	0.000	**0.333**
BERT Italian (cased)	0.935	**0.663**	**0.125**	**0.167**	0.000	**0.333**
			F1			
BERT multilingual (cased)	0.930	0.603	**0.167**	0.108	0.000	0.286
XLM-RoBERTa	**0.941**	**0.659**	0.000	0.133	0.000	**0.400**
BERT Italian (cased)	0.934	0.646	0.154	**0.162**	0.000	0.333

Table 11.7. AC (reviewed) results

	Present			*Absent*			*Family History*		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT multilingual (cased)	0.983	**0.978**	0.980	0.835	0.865	0.850	0.894	0.894	0.894
XLM-RoBERTa	0.974	0.980	0.977	0.800	0.829	0.814	0.916	0.578	0.709
BERT Italian (cased)	**0.985**	0.977	**0.981**	0.829	0.890	**0.858**	0.947	0.947	**0.947**

quire a numerical representation $V_{concept}$ by using a pre-trained language model [52, 24] as a feature extractor. This process involves feeding each phrase that contains a specific mention from the Lexicon into the feature extractor.

The feature extractor takes these input phrases and maps each token within them to its corresponding word embedding, which we denote as $V_{context}$. This word embedding consists of an array of numerical features that represent the token within the context in which it is found. When mentions consist of multiple tokens, the $V_{context}$ is computed by averaging the word embeddings of all the tokens in the mention.

After retrieving a $V_{context}$, the overall numerical representation of the concept $V_{concept}$ is updated as follows:

$$V_{concept} = V_{concept} + lr \cdot (1 - sim) \cdot V_{context}, \tag{11.2}$$

where lr is a regularization term defined by the reciprocal of the number of times a mention appears in the whole dataset and sim is the cosine similarity between $V_{concept}$ and $V_{context}$:

$$lr = \frac{1}{C_{concept}} \tag{11.3}$$

$$sim(V_{concept}, V_{context}) = \max(0, \frac{V_{concept}}{||V_{concept}||} \cdot \frac{V_{context}}{||V_{context}||}) \tag{11.4}$$

$V_{concept}$ is initialized to the $V_{context}$ value of the first sentence where the mention appears.

Results We have applied the methodology described above to link entities while using the transformer architectures presented in Section 11.3.3. Results have been evaluated in terms of accuracy and we present them in Table 11.8.

Although the utilization of transformer architectures pretrained on Italian corpora has demonstrated its utility also in this context, the results are worse w.r.t. NER and AC tasks. This diminished performance can be attributed to the inherent intricacies associated with mapping and aligning the linguistic structures to biomedical ontologies.

Table 11.8. EL results

Class	**MedicalProblem**	**Treatment**	**Test**
BERT multilingual (cased)	0.594	0.500	0.724
XLM-RoBERTa	0.548	0.400	0.557
BERT Italian (cased)	**0.651**	**0.600**	**0.729**

To delve deeper into the performance of the EL model, we have reported its accuracy in associating with each unique test concept identifier (CUI), depicted in a descending order in Figure 11.4. Across all entity types, a consistent pattern emerges: a subset of concepts, approximately half, is amenable to accurate linking by the model, while another subset consistently eludes correct association.

Several instances of errors have been documented in Table 11.9. These instances serve as illustrative demonstrations, highlighting that, notwithstanding the presence of errors, the model's predictions exhibit a degree of semantic relevance to the input entity mention. The applicability of this

Table 11.9. EL examples. English translations for the input samples are reported to ease the understanding of international readers.

Sentence	Ground truth	Prediction
Il pz riferisce dolore toracico qualche giorno fa, per il ripresnetarsi della **sintomatologia anginosa**. (*The patient reports chest pain a few days ago due to resurface of* ***angina symptoms****.*)	C0002962 (Angina Pectoris)	C0008031 (Chest Pain)
Cardiopatia ischemica con indici di funzione sistolica ventricolare sinistra ridotti a riposo (EF circa 35%). (***Ischemic heart disease*** *with reduced left ventricular systolic function indices at rest (EF about 35%)*)	C0151744 (Myocardial Ischemia)	C1869045 (Ischaemic heart disease)
Test sottomassimale a medio carico eseguito in terapia farmacologica . Assenza di sintomi, aritmie e **anomalie** del tratto ST-T. Buono il profilo pressorio (*Mid-load submaximal test performed under drug therapy. Absence of symptoms, arrhythmias and ST-T tract* ***abnormalities****. Good pressor profile.*)	C0000768 (Congenital Abnormality)	C0919620 (Electrocardiogram ST-T change)
Ipertensione Arteriosa Grado III STADIO III TIA/**ICTUS**. Rischio cardiovascolare aggiuntivo molto elevato. Controllo farmacologico insoddisfacente. (*Hypertension Grade III STAGE III TIA/****STROKE****. Very high additional cardiovascular risk. Unsatisfactory pharmacological control.*)	C0038454 (Cerebrovascular accident)	C0007787 (Transient Ischemic Attack)
Sospendere gradualmente Atenololo una settimana prima del test ergometrico (1/2 cp per i primi 2 giorni, 1/4 cp per i successivi 2 poi sospende) e sostituirlo con **Norvasc** 5. (*Gradually discontinue Atenolol one week before the exercise test (half a tablet for the first 2 days, a quarter tablet for the next 2 and then discontinue) and replace it with* ***Norvasc*** *5.*)	C1606917 (amlodipine 5 MG [Norvasc])	C0162712 (Norvasc)

model is contingent upon the inherent criticality of the specific application context.

The accuracy of predictions is inherently dependent on the level of the similarity value generated by the system. Elevated similarity scores correlate positively with heightened prediction accuracy. However, as demonstrated in Figure 11.5, there exists a discernible inverse relationship between accuracy and support, i.e. the number of test samples for which the system can furnish a given similarity threshold. In light of these results, to guarantee high accuracy while maintaining a good support, we link instances with our model when the output similarity score provided by the model is higher than 0.8, otherwise we use UMLS query search APIs to get the most syntactically relevant link to the input mention.

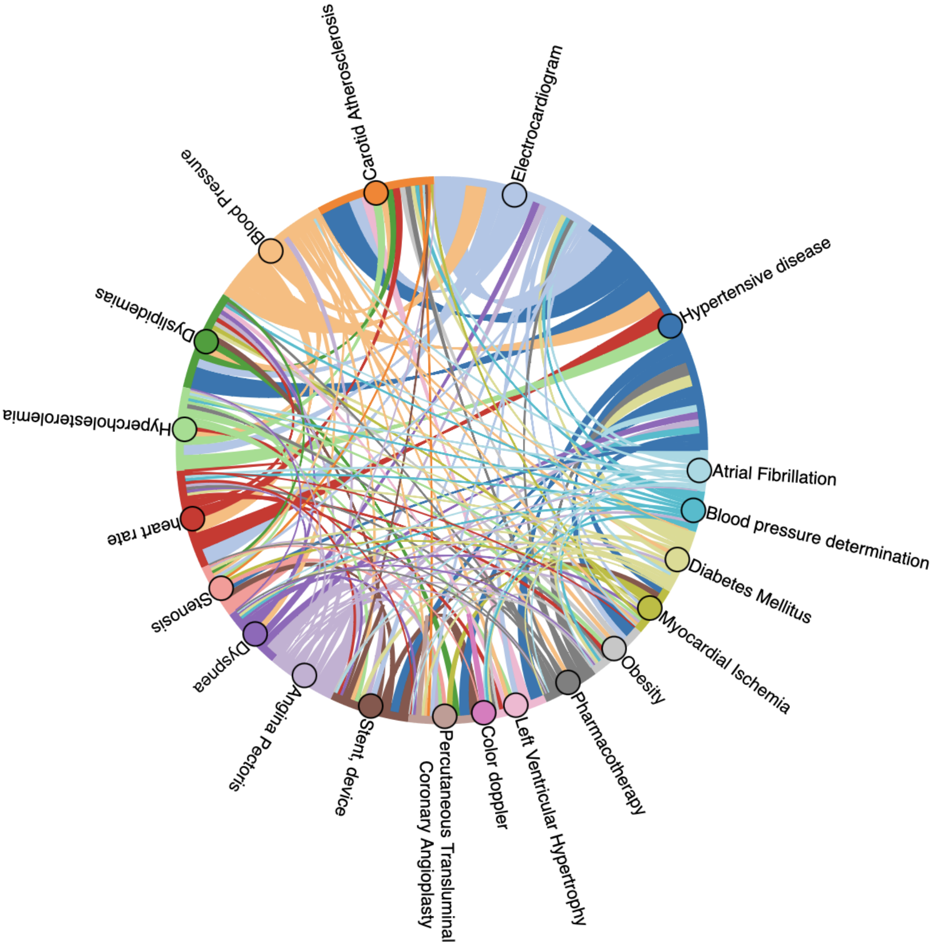


Figure 11.3. Chord diagram of the Entity Linking corpus. Here we show the co-mentions of the top 20 UMLS codes. Instead of the concept unique identifiers (CUIs), we show the name of the terms to ease the understanding of the plot.

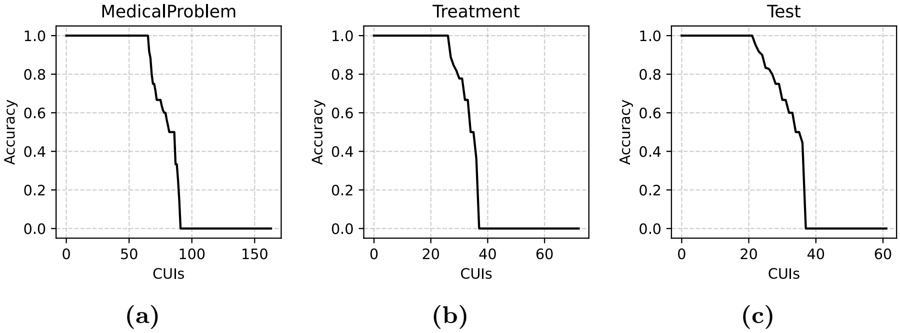


Figure 11.4. Accuracy of the proposed EL method on the test CUIs, reported separately for each entity type and in descending order.

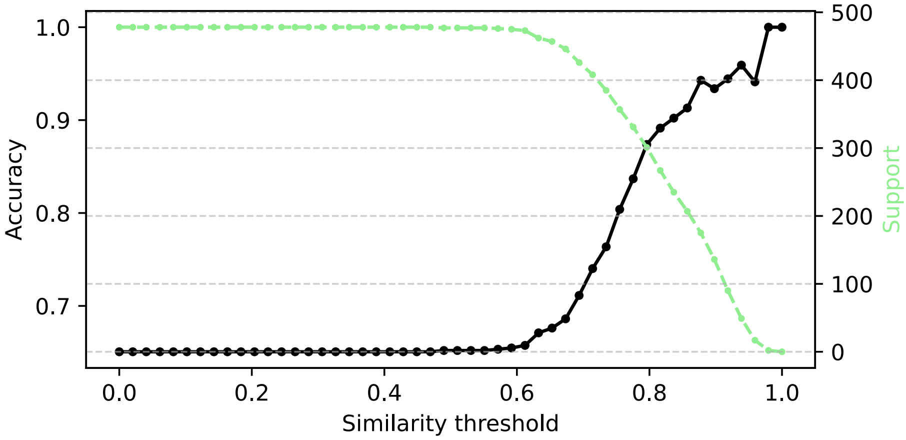


Figure 11.5. Trade-off between accuracy and support at different similarity threshold levels.

Chapter 12

Analytics

The dataset from the Campania Salute (CS) network, referenced in Section 11.1, stands as a pivotal repository of clinical information, documenting a plethora of medical events for a heterogeneous patient population. Using the Information Extraction techniques outlined in Chapter 11, a Temporal Knowledge Graph (TKG) can be constructed for each patient. These TKGs not only facilitate the structured representation of patients' medical histories for domain specialists but also optimize them for advanced processing through graph-based embedding techniques. In particular, this chapter explores the application of MedTKG, the methodology described in Chapter 10.

In contrast to the strategy in Chapter 10 that employs the SNOMED-CT ontology, the UMLS ontology is adopted here. This choice broadens the spectrum of relationships between biomedical concepts, potentially enhancing predictive capabilities regarding future adverse events for patients. The intricate web of relationships within UMLS, when juxtaposed with SNOMED-CT, may yield superior result quality.

This chapter presents the analysis results. Section 12.1 provides a detailed overview of the methods used to shape TKGs that depict the medical histories of patients, combined with the Medical Ontology Graph—a collection of relations extracted from the UMLS Metathesaurus. Following this, Section 12.2 offers a comprehensive examination of the outcomes, based on the methodology described in Chapter 10. Illustrative cases processed using this approach and reviewed by domain specialists are also included.

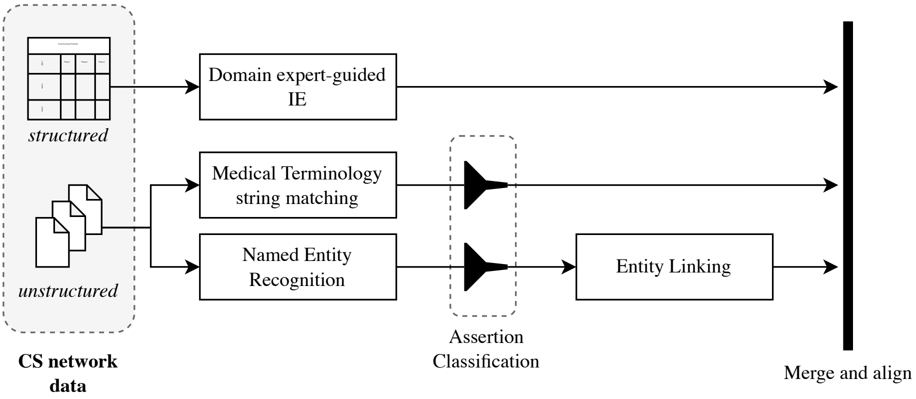


Figure 12.1. Overview of the TKG extraction flow

12.1 Materials & Methods

12.1.1 Medical Histories: patients TKGs

Figure 12.1 provides a schematic representation of the workflow employed for constructing a TKG that encapsulates the medical history of a patient.

This process is initiated by aggregating comprehensive patient information from both structured and unstructured sources within the Campania Salute network. Subsequently, these data are subjected to a Information Extraction pipeline with methods detailed in Chapter 11.

To be more specific, the proposed methodology extracts a range of adverse events from structured data using a specific procedure, which is detailed in Section 11.2. Furthermore, numerous medical concepts are derived from unstructured clinical notes. This extraction uses two methods: (1) a string-matching algorithm based on medical terminology, and (2) a transformer-based Named Entity Recognition (NER) model tailored for Italian medical text. For more information on these methods, see Section 11.3.

Concepts extracted from free-text notes is further processed with an Assertion Classification model (refer to Section 11.3.5 for details) that filters out all the non-relevant events (i.e. concepts that are absent, hypothetical or conditional). Then, while concepts extracted under domain experts

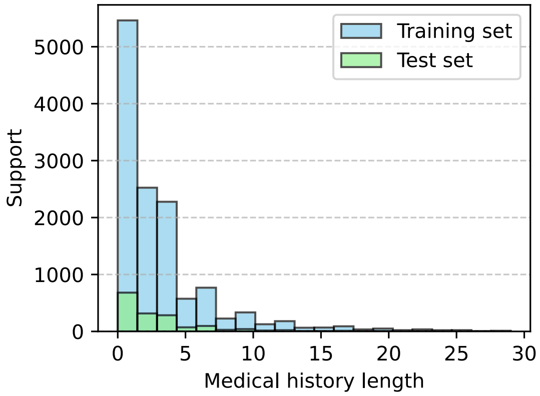


Figure 12.2. Medical history lengths distributions across training and test data. We do not report the distribution for validation data since it overlaps with test data due to the stratified nature of our splitting strategy.

guidance affer to pre-determined categories which are already linked to UMLS concept identifiers, an Entity Linking step is required for concepts extracted with the NER model (refer to Section 11.3.6 for details). Information extracted with the above-described modalities is then merged and aligned so as to constitute the medical history of a patient as a TKG, as defined in Definition 3.3. Events occurring during the same day have been grouped in the same timestamp. The resulting dataset comprises 16,100 patients.

To empirically evaluate the efficacy of the methods, a systematic partitioning of the dataset was performed. Specifically, 80% of the patient cohort (12,896 patients) was allocated to the training dataset, while the remaining 20% was evenly distributed between the validation and test datasets (1,594 and 1,610 patients, respectively). The partitioning was conducted in a stratified manner, using the lengths of patients' medical histories as the stratification criterion. This strategy ensures an equivalent distribution of patients across various historical record lengths in both the training and test datasets. Figure 12.2 displays medical history lengths distributions across training and test data, similar to those observed in the MIMIC-III data as shown in Figure 10.3.

A distribution of the number of extracted concepts for each patient is

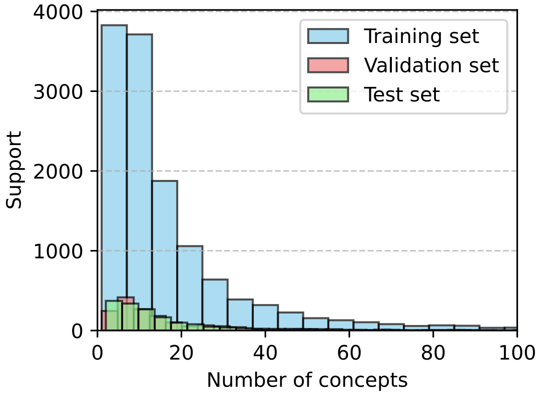


Figure 12.3. Distribution of the number of extracted concepts for each patient. We limit the x-axis between 0 and 100, though some outliers exceed this threshold, the maximum value being 301.

shown in Figure 12.3. The majority of concepts extracted refer to disorders and findings, as summarized in Table 12.1.

Medical Ontology Graph

We use the UMLS query APIs to get all the relationships between the medical concepts stored in our training data. We take the following relationship types into consideration:

- *RN*: the two concepts have a narrow relationship
- *RB*: the two concepts have a broad relationship
- *RO*: the two concepts have a relationship other than narrower or broader.

Based on the 6,812 concepts (CUIs) in the dataset, a total of 1,092 narrow relationships (RN), 475 broad relationships (RB), and 1,728 other (RO) have been extracted. Figure 12.4 illustrates the interconnections among the top-5 concept types. The illustration highlights that the primary associations are between disorders and findings, pathologic functions and disorders, and within therapeutic procedures.

Table 12.1. Number of concepts in the dataset splits. We show the occurrences of the most relevant concepts separately, and the total number of concepts in the last row.

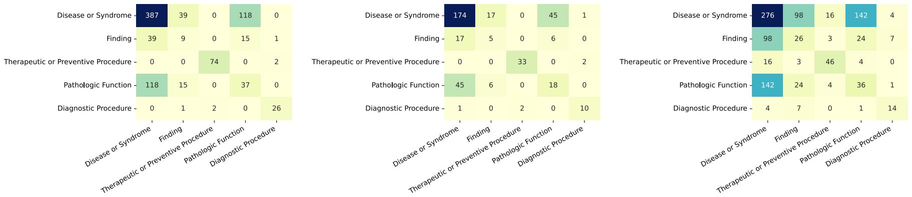
Concept type	**Training**	**Validation**	**Test**
Disease or Syndrome	82,028	10,135	10,232
Finding	24,029	2,919	2,959
Therapeutic or Preventive Procedure	17,215	2,109	2,169
Pathologic Function	15,259	1,867	1,925
Diagnostic Procedure	15,138	2,023	1,854
Congenital abnormality	8,738	1,084	1,036
Medical Device	6,176	685	689
Functional Concept	6,116	762	835
Health Care Activity	4,807	619	626
Organism Function	4,366	538	590
Total	221,798	27,451	27,810

12.2 Adverse events prediction

The previously described TKGs and the medical ontology graph will be harnessed to construct a system capable of identifying potential disorders that may be linked to patients in the future. This effort will be guided by the methodology outlined in Chapter 10. It is recommended to refer to the techniques detailed in Chapter 10 for insights into the prediction of future disorders. The findings will be presented and analyzed in this section from various viewpoints.

12.2.1 Results

The experimental results of the proposed methodology methodology are presented in Table 12.2. The findings demonstrate that the proposed model exhibits a notable capability to forecast a forthcoming disorder, finding a true positive at the top-ranking position in approximately 67% of test samples. As k increases, the TP Rate also increases significantly, reaching 93.04% when considering the top 10 predictions (k=10). This indicates that when the system considers more predictions, it becomes more effective at identifying adverse events correctly. This is a positive trend, suggesting that the predictions become more reliable as more options



(a) Narrow **(b)** Broad **(c)** Others

Figure 12.4. Relationships between top-5 occurring concept types. Results are shown for each relationship type, separately.

Table 12.2. Performance of the system for adverse events prediction. k indicates the number of top predictions considered for the computation of the metric (except for MRR, that does not depend from k).

Metric	**k=1**	**k=3**	**k=5**	**k=10**
TP Rate	67.06	82.06	86.9	93.04
Hits	67.06	42.84	31.74	20.42
MR	20.27	33.93	39.76	48.29
MAP	20.27	29.88	32.47	34.96
MRR	42.99			

are considered. However, at $k = 1$, Hits is equal to TP Rate, but as k grows, Hits drops more rapidly. This suggests that the system may make some accurate predictions within the top few predictions but becomes less accurate as more predictions are included.

Ensuring a robustly high recall rate in the prognostication of prospective medical conditions poses a challenge that stems from the propensity of the model to generate predictions for disorders that, while theoretically plausible, do not manifest in reality. Consequently, this discrepancy gives rise to negative predictions, exerting a detrimental influence on the mean recall (MR) metric.

While there has been prior discussion on the general performance of the system, the focus will now shift to the top 20 most-occurring concepts, where an enhanced performance is anticipated. Specifically, Table 12.3 showcases precision scores (P@k) for various concepts associated with dis-

Table 12.3. Precision (P@k) on the 20 most-occurring concepts. Results in the first quartile are highlighted in green, while those in the last quartile are highlighted in red.

Disorder or Syndrome	P@1	P@3	P@5	P@10
Chronic Kidney Insufficiency (C0403447)	68.11	84.28	92.24	98.02
Mitral Valve Insufficiency (C0026266)	37.11	85.55	96.42	100.00
Myocardial Ischemia (C0151744)	48.89	74.45	86.28	94.28
Chronic myocardial ischemia (C0264694)	43.98	56.68	69.59	83.82
Hypertensive disease (C0020538)	12.38	43.89	51.16	73.67
Aortic Diseases (C0003493)	37.85	57.46	80.48	94.80
Dyslipidemias (C0242339)	20.99	34.52	49.18	87.77
Aortic Valve Insufficiency (C0003504)	9.79	27.78	39.95	80.25
Tricuspid Valve Insufficiency (C0040961)	10.63	41.95	57.85	94.83
Diabetes (C0011847)	45.10	75.07	85.86	98.12
Cardiomyopathies (C0878544)	64.78	69.80	70.33	71.61
Heart failure (C0018801)	37.86	63.13	66.34	78.15
Sclerotic aortic valve (C4015488)	13.28	41.28	46.22	53.78
Left Ventricular Hypertrophy (C0149721)	23.68	57.20	63.16	79.09
Obesity (C0028754)	21.01	53.70	57.99	71.89
Left cardiac ventricular dilatation (C0344911)	10.57	21.59	29.56	37.52
Atrial Fibrillation (C0004238)	12.70	47.58	56.35	61.72
Aortic sclerosis (C1331537)	0.72	17.95	38.42	50.81
Peripheral Arterial Diseases (C1704436)	11.97	44.22	45.03	64.10
Myocardial Infarction (C0027051)	9.16	47.45	72.30	72.30

orders and syndromes. These precision scores represent the percentage of accurate predictions made by a machine learning model when evaluating the top k predicted concepts. Additionally, the table emphasizes precision scores across different quartiles for a comprehensive understanding. Precision scores in the first quartile are accentuated in green, denoting superior performance for the respective concept. Conversely, scores in the last quartile are marked in red, signifying a comparatively lower performance.

Analyzing the table, it can be observed that: *Chronic Kidney Insufficiency* shows consistently high precision scores across all prediction scenarios, with the highest precision at P@10 (98.02%). This indicates that the machine learning model has a high accuracy in predicting this particular concept; *Mitral Valve Insufficiency* also has high precision scores, with perfect precision at P@10 (100.00%); *Myocardial Ischemia* exhibits relatively high precision scores, particularly at P@1 (48.89%) and P@3 (74.45%). However, precision drops as the number of predictions increases, relatively

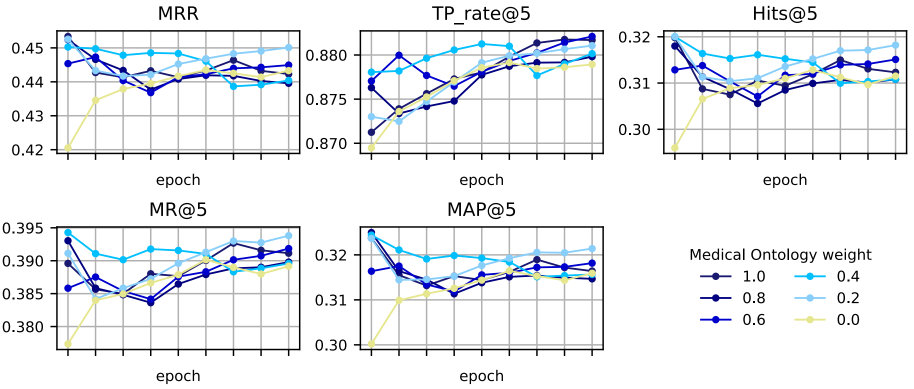


Figure 12.5. Trends of performance with different Medical Ontology weights on validation data over training epochs

to other concepts; *Hypertensive disease* has a relatively low precision score at P@1 (12.38%), indicating that it is more challenging for the model to predict this concept accurately in the top position. However, the precision improves as the number of predictions increases. It is worth to note that *Tricuspid Valve Insufficiency* is rarely predicted at the top position, but it is one of the most accurately predicted concepts among the first ten predictions: this means that the model may correctly predict this disorder, but might not be very confident in doing so. If the model was trained on data that reflects the clinical decision-making process, it may have learned to prioritize certain conditions over others based on historical medical practices.

12.2.2 Training results

In this section, we summarize training results that highlight the impact of the medical ontology and medical histories.

Trends of performance during the training of the model are shown in Figure 12.5. Our findings highlight the impact of using the external medical ontology as a source for medical information, that guarantees higher performance from the very beginning of the learning process.

Figure 12.6 illustrates the trends of concepts ever predicted (CEP) with different Medical Ontology weights, showing that the use of external

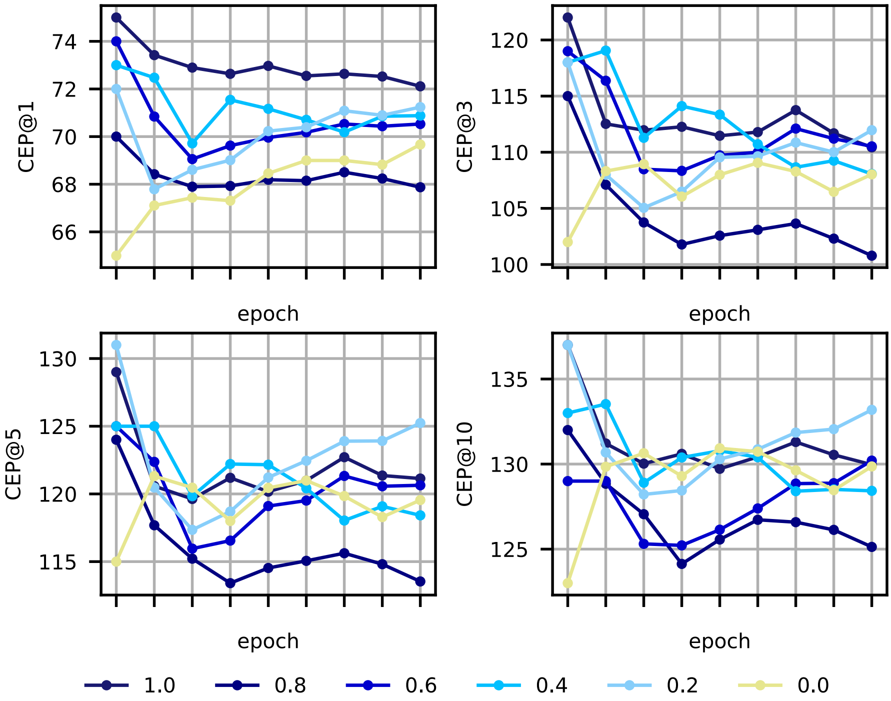


Figure 12.6. Trends of concepts ever predicted (CEP) with different Medical Ontology weights on validation data over training epochs.

knowledge guarantees a higher coverage of medical concepts in model predictions, especially in the top-scoring ones (represented by CEP@1 scores).

Finally, Figure 12.7 reports results showing the impact of leveraging longer medical histories. Thanks to the inner characteristics of the Italian dataset, we have observed better results w.r.t. MIMIC-III data (ref. to Figure 10.7). The metrics that benefit the most from longer medical histories are MRR, TP Rate and Hits, meaning that the model improves its ability to distinguish between true and false positives.

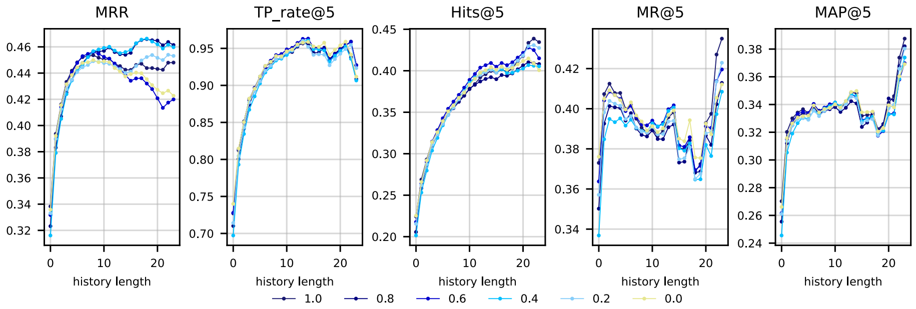


Figure 12.7. Performance trends of models with different weights attributed to the medical ontology constraint as the length of medical histories increases.

12.2.3 Examples

In this section, predictions from the proposed model for a subset of five test patients, which were not part of the training set, are presented. Discrepancies between these predictions and the established ground truth are highlighted. Additionally, all predictions have been analyzed and discussed with the help of domain experts.

Patient 1 (Figure 12.8) This patient has a complex medical history with various interconnected disorders and conditions. The ground truth shows that the patient will be related to Chronic Kidney Insufficiency at the next time step t_6, which is something we would expect. Predictions provided by MedTKG highlight further medical problems that might be related to the patient's history and could possibly manifest in the future. In the following, we detail every prediction:

- *Chronic Kidney Insufficiency (CKI).* The patient already has CKI from t_3 to t_6. Diabetes and hypertension (from t_0) are well-known risk factors for CKI. Over time, high blood sugar levels from diabetes and high blood pressure can damage the kidneys' glomeruli, leading to CKI.

- *Mitral Valve Insufficiency.* Both hypertension and CKI can contribute to left ventricular hypertrophy and atrial fibrillation, which in turn can lead to mitral valve insufficiency.

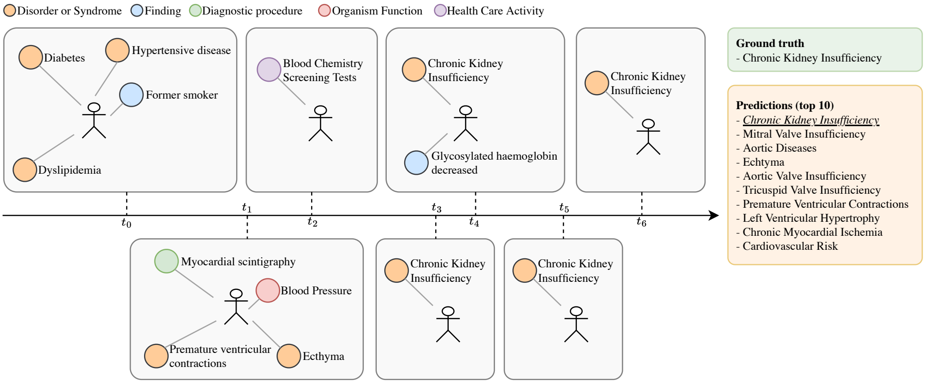

Disorder or Syndrome
Finding
Diagnostic procedure
Organism Function
Health Care Activity
Diabetes
Hypertensive disease
Former smoker
Dyslipidemia
Blood Chemistry Screening Tests
Chronic Kidney Insufficiency
Glycosylated haemoglobin decreased
Chronic Kidney Insufficiency
t_0
t_1
t_2
t_3
t_4
t_5
t_6
Myocardial scintigraphy
Blood Pressure
Premature ventricular contractions
Ecthyma
Chronic Kidney Insufficiency
Chronic Kidney Insufficiency
Ground truth
- Chronic Kidney Insufficiency
Predictions (top 10)
- Chronic Kidney Insufficiency
- Mitral Valve Insufficiency
- Aortic Diseases
- Echtyma
- Aortic Valve Insufficiency
- Tricuspid Valve Insufficiency
- Premature Ventricular Contractions
- Left Ventricular Hypertrophy
- Chronic Myocardial Ischemia
- Cardiovascular Risk

Figure 12.8. Example of prediction from a given medical history (n. 1)

- *Aortic Diseases.* Hypertension (from t_0) is a significant risk factor for aortic diseases as it can cause damage to the aorta's wall.
- *Ecthyma.* The patient has already had ecthyma at t_1. Diabetes (from t_0) can impair immune system function and wound healing, making the patient susceptible to skin infections like ecthyma.
- *Aortic Valve Insufficiency.* Hypertension and left ventricular hypertrophy can strain the aortic valve, leading to aortic valve insufficiency.
- *Tricuspid Valve Insufficiency.* Elevated blood pressure, especially pulmonary hypertension, can lead to right heart strain and subsequently tricuspid valve insufficiency.
- *Premature Ventricular Contractions (PVCs).* The patient has already had PVCs at t_1. Hypertension, left ventricular hypertrophy, and myocardial ischemia can contribute to the occurrence of PVCs.
- *Left Ventricular Hypertrophy (LVH).* Hypertension (from t_0) is a well-known cause of LVH as it makes the heart work harder to pump blood, causing the left ventricle to thicken or stiffen.
- *Chronic Myocardial Ischemia.* The presence of diabetes, hypertension, and dyslipidemia (from t_0) significantly increases the risk of

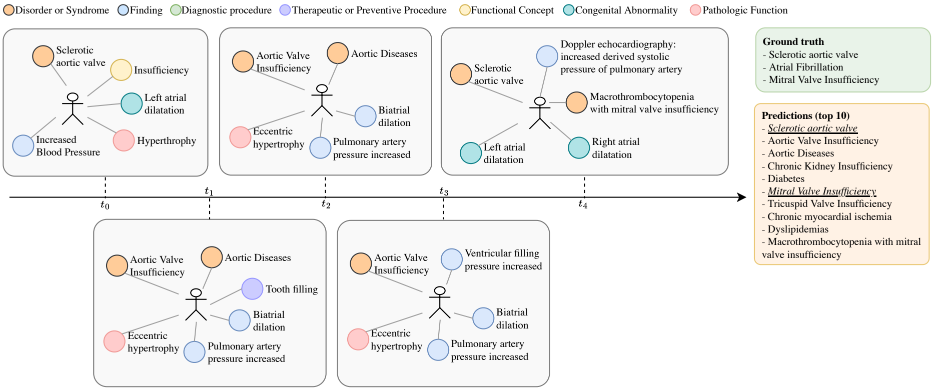

Disorder or Syndrome
Finding
Diagnostic procedure
Therapeutic or Preventive Procedure
Functional Concept
Congenital Abnormality
Pathologic Function
Sclerotic aortic valve
Insufficiency
Left atrial dilatation
Increased Blood Pressure
Hyperthrophy
Aortic Valve Insufficiency
Aortic Diseases
Biatrial dilation
Eccentric hypertrophy
Pulmonary artery pressure increased
Sclerotic aortic valve
Doppler echocardiography: increased derived systolic pressure of pulmonary artery
Macrothrombocytopenia with mitral valve insufficiency
Left atrial dilatation
Right atrial dilatation
Ground truth
- Sclerotic aortic valve
- Atrial Fibrillation
- Mitral Valve Insufficiency
Predictions (top 10)
- Sclerotic aortic valve
- Aortic Valve Insufficiency
- Aortic Diseases
- Chronic Kidney Insufficiency
- Diabetes
- Mitral Valve Insufficiency
- Tricuspid Valve Insufficiency
- Chronic myocardial ischemia
- Dyslipidemias
- Macrothrombocytopenia with mitral valve insufficiency
t_0
t_1
t_2
t_3
t_4
Aortic Valve Insufficiency
Aortic Diseases
Tooth filling
Biatrial dilation
Eccentric hypertrophy
Pulmonary artery pressure increased
Aortic Valve Insufficiency
Ventricular filling pressure increased
Biatrial dilation
Eccentric hypertrophy
Pulmonary artery pressure increased

Figure 12.9. Example of prediction from a given medical history (n. 2)

coronary artery disease, which can lead to chronic myocardial ischemia.

- *Cardiovascular Risk.* The comorbidity of diabetes, hypertension, and dyslipidemia (from t_0), along with CKI and the cardiac issues noted, significantly heightens the patient's overall cardiovascular risk.

The interplay of these conditions and their progression over time can lead to a compound effect on the patient's health, often exacerbating one or more of the other conditions in a vicious cycle. Each condition needs to be managed appropriately to mitigate the risk of further complications.

Patient 2 (Figure 12.9) The medical history illustrates a case of a patient grappling with escalating cardiac complications, prominently affecting the aortic valve and related cardiac anatomy. MedTKG proficiently pinpoints *Sclerotic Aortic Valve* and *Mitral Valve Insufficiency* as plausible medical concerns for the patient, aligning with expectations since the medical history mentions *Sclerotic Aortic Valve* and *Macrothrombocytopenia with Mitral Valve Insufficiency*. Yet, it overlooks *Atrial Fibrillation (AFib)*, a potentially significant issue warranting consideration in a thorough clinical evaluation and management strategy, given the progressive nature of the patient's cardiac ailments.

An analysis of how the predicted future medical problems could be related to the historical medical issues of the patient is provided as follows:

- *Sclerotic Aortic Valve.* Initially mentioned at t_0 and t_4, sclerotic aortic valve can cause stenosis, restricting blood flow from the heart to the aorta. This condition can progress over time, leading to symptoms like chest pain, fatigue, and sometimes heart failure.

- *Aortic Valve Insufficiency.* This condition is recurrently mentioned from t_1 to t_3. It refers to the inability of the aortic valve to close tightly, allowing blood to flow backward into the heart. This can be a progression or a consequence of the sclerotic aortic valve.

- *Aortic Diseases* Mentioned at t_1 and t_2, aortic diseases could encompass a range of disorders affecting the aorta, potentially stemming from the sclerotic aortic valve or valve insufficiency.

- *Chronic Kidney Insufficiency.* Cardiac and renal functions are closely interrelated. Aortic diseases and hypertension (noted at t_1) could lead to renal artery stenosis, impeding blood flow to the kidneys and resulting in chronic kidney insufficiency.

- *Diabetes.* While not directly mentioned in the history, hypertension and aortic diseases can be associated with metabolic disorders like diabetes. Additionally, diabetes is a risk factor for atherosclerosis, which may worsen aortic and valvular conditions.

- *Mitral Valve Insufficiency.* Mentioned at t_4 with macrothrombocytopenia. Mitral valve insufficiency could result from the dilation of cardiac chambers and altered hemodynamics noted in earlier timestamps.

- *Tricuspid Valve Insufficiency.* The dilation of cardiac chambers and increased pulmonary artery pressures mentioned might affect the function of the tricuspid valve over time.

- *Chronic Myocardial Ischemia.* Aortic valve disorders and hypertension can lead to increased cardiac workload, potentially resulting in myocardial ischemia over time due to inadequate blood supply.

- *Dyslipidemias.* Dyslipidemia is a common comorbidity with hypertension and aortic diseases. It might exacerbate atherosclerotic changes in the aortic valve and vessels.

- *Macrothrombocytopenia with Mitral Valve Insufficiency.* Mentioned at t_4, it is a rare disorder that presents with both cardiac and hematologic abnormalities. The cardiac alterations observed in earlier timestamps might have predisposed or coexisted with this condition.

The progressive cardiac issues of this patient, along with associated findings like increased blood pressure and cardiac chamber dilations, create a complex clinical picture with intertwined cardiac, renal, and potentially metabolic disorders.

Patient 3 (Figure 12.10) The patient has a complex cardiovascular medical history characterized by multiple valvular insufficiencies (mitral and aortic), aortic diseases, and systemic conditions like obesity, hypertension, diabetes, and dyslipidemia, which are well-known risk factors for cardiovascular disease. They have also experienced a myocardial infarction, further complicating their cardiac status. Various diagnostic and therapeutic interventions have been employed, including the insertion of a cardioverter-defibrillator, radio-frequency ablation, angiogram, cardiac electrophysiologic studies, and dilate procedures. Despite these interventions, persistent issues like left atrial dilation, elevated pulmonary artery pressure, and eccentric hypertrophy indicate ongoing cardiac remodeling and dysfunction. MedTKG correctly identifies *Myocardial Ischemia* as a potential risk for the patient. However, it is worthy to note that *Chronic Myocardial Ischemia* — which implies a long-standing or recurring issue — is ranked higher among the predictions, possibly because the extensive cardiac history including myocardial infarction at t_1, the ongoing nature of their heart problems suggests a more chronic course of myocardial ischemia.

In the following, we evaluate how each of the predicted future medical problems could be related to the patient:

- *Mitral Valve Insufficiency.* This condition has already been mentioned at multiple timestamps (t_0, t_2, t_3). Given the persistence of

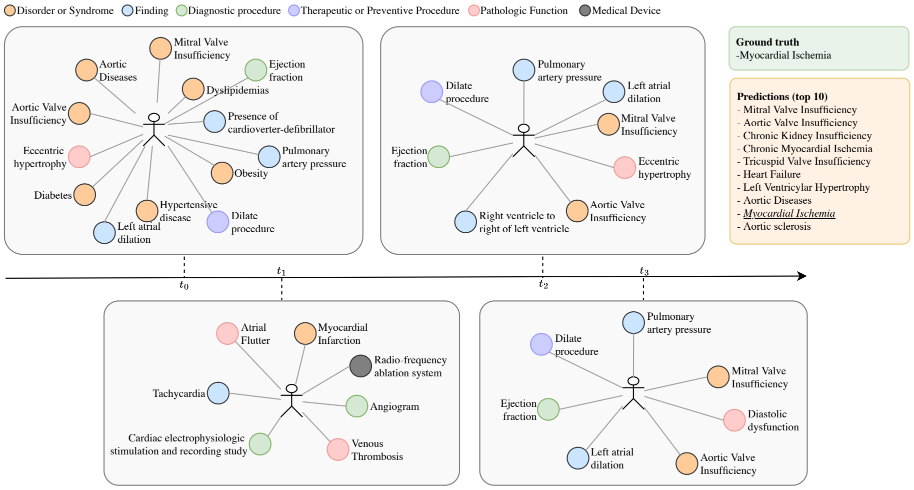

Disorder or Syndrome
Finding
Diagnostic procedure
Therapeutic or Preventive Procedure
Pathologic Function
Medical Device
Mitral Valve Insufficiency
Aortic Diseases
Ejection fraction
Dyslipidemias
Aortic Valve Insufficiency
Presence of cardioverter-defibrillator
Eccentric hypertrophy
Pulmonary artery pressure
Obesity
Diabetes
Hypertensive disease
Dilate procedure
Left atrial dilation
Pulmonary artery pressure
Dilate procedure
Left atrial dilation
Mitral Valve Insufficiency
Ejection fraction
Eccentric hypertrophy
Right ventricle to right of left ventricle
Aortic Valve Insufficiency
Ground truth
-Myocardial Ischemia
Predictions (top 10)
- Mitral Valve Insufficiency
- Aortic Valve Insufficiency
- Chronic Kidney Insufficiency
- Chronic Myocardial Ischemia
- Tricuspid Valve Insufficiency
- Heart Failure
- Left Ventricylar Hypertrophy
- Aortic Diseases
- Myocardial Ischemia
- Aortic sclerosis
t_0
t_1
t_2
t_3
Atrial Flutter
Myocardial Infarction
Radio-frequency ablation system
Tachycardia
Angiogram
Cardiac electrophysiologic stimulation and recording study
Venous Thrombosis
Pulmonary artery pressure
Dilate procedure
Mitral Valve Insufficiency
Ejection fraction
Diastolic dysfunction
Left atrial dilation
Aortic Valve Insufficiency

Figure 12.10. Example of prediction from a given medical history (n. 3)

this issue, it may continue to affect the patient or even worsen over time.

- *Aortic Valve Insufficiency.* Similar to mitral valve insufficiency, aortic valve insufficiency is also mentioned repeatedly across the timestamps (t_0, t_2, t_3). The ongoing presence of this condition indicates a likelihood of its continuation or progression.

- *Chronic Kidney Insufficiency.* Hypertensive disease and diabetes (t_0) are well-known risk factors for chronic kidney disease. The heart and kidney have a complex interplay; issues like heart failure or valvular diseases can exacerbate kidney dysfunction and vice versa.

- *Chronic Myocardial Ischemia.* The patient has a history of myocardial infarction (t_1) and other cardiac-related issues, which may predispose them to chronic myocardial ischemia due to a possible underlying atherosclerotic disease.

- *Tricuspid Valve Insufficiency.* The persistent left atrial dilation, increased pulmonary artery pressure (t_0, t_2, t_3), and other cardiac abnormalities may place additional strain on the tricuspid valve, potentially leading to tricuspid valve insufficiency.

- *Heart Failure.* The patient has numerous risk factors for heart failure including hypertension, obesity, diabetes (t_0), and myocardial infarction (t_1). Additionally, valvular diseases and reduced ejection fraction (t_0, t_2) are direct contributors to heart failure.
- *Left Ventricular Hypertrophy (LVH).* Hypertension (t_0) and aortic valve diseases can cause increased pressure load on the left ventricle, possibly leading to LVH. The eccentric hypertrophy mentioned (t_0, t_2, t_3) also points towards an ongoing process of cardiac remodeling.
- *Aortic Diseases.* Aortic diseases were already present at t_0. The patient's hypertension and aortic valve insufficiency could contribute to the progression of these aortic diseases.
- *Myocardial Ischemia.* Similar to chronic myocardial ischemia, the past myocardial infarction (t_1) and potential underlying atherosclerotic disease may predispose the patient to episodes of myocardial ischemia.
- *Aortic Sclerosis.* The presence of aortic diseases and aortic valve insufficiency (t_0, t_2, t_3), along with dyslipidemias (t_0), may contribute to a process of aortic sclerosis.

Patient 4 (Figure 12.11) The medical history of this patient depicts a complex and interrelated array of cardiovascular and renal disorders, alongside interventions and various findings. The potential future medical issues listed are deeply interconnected with the past medical conditions and may arise due to the progression of these conditions or as complications thereof. In the following, we delve into how each of these future medical problems might be related to the medical history:

- *Pericardial Effusion.* This condition, which involves the accumulation of fluid in the pericardial cavity, persists in the patient medical history from timestamp t_1 and could be a result of several existing

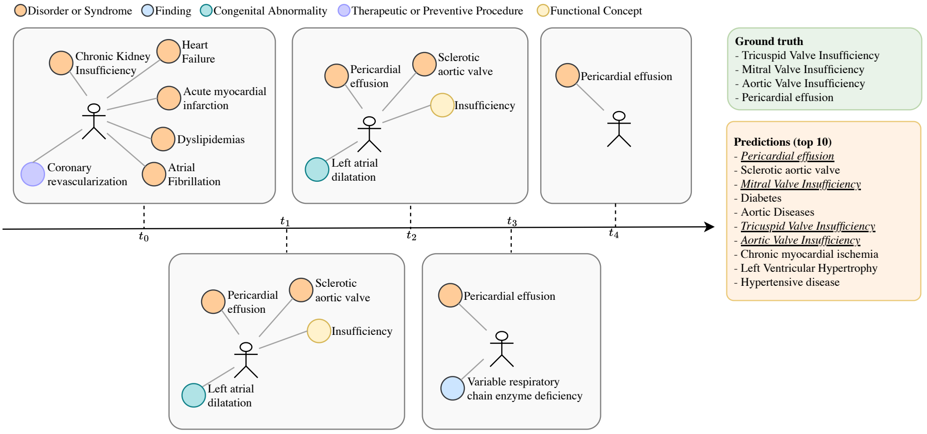


Figure 12.11. Example of prediction from a given medical history (n. 4)

conditions such as heart failure, myocardial infarction, and atrial fibrillation. These conditions can lead to inflammation or irritation of the pericardium, contributing to pericardial effusion.

- *Sclerotic Aortic Valve.* A sclerotic aortic valve is observed at t_1 and t_2, potentially arising from the patient's dyslipidemias, which could cause lipid deposition and calcification in the aortic valve, leading to sclerosis over time.

- *Mitral Valve Insufficiency.* This might develop due to the left atrial dilatation noted at t_1 and t_2, which could stretch and deform the mitral valve annulus, leading to mitral valve insufficiency. Additionally, the chronic myocardial ischemia and left ventricular hypertrophy could further impair mitral valve function.

- *Diabetes.* Diabetes could potentially develop due to the interplay between dyslipidemias, chronic kidney insufficiency, and potentially a sedentary lifestyle or other risk factors not specified in the patient's history.

- *Aortic Diseases.* Aortic diseases might manifest from the existing sclerotic aortic valve condition, dyslipidemias, and possibly hypertension (if present). These conditions can cause structural and functional changes in the aorta.

- *Tricuspid Valve Insufficiency.* This could arise due to the atrial fibrillation and right-sided heart changes secondary to left-sided heart conditions like heart failure and mitral valve insufficiency.

- *Aortic Valve Insufficiency.* This could result from the progression of the sclerotic aortic valve condition and possibly exacerbated by chronic myocardial ischemia.

- *Chronic Myocardial Ischemia.* The patient's history of acute myocardial infarction, dyslipidemias, and coronary revascularization suggests a background of coronary artery disease, which could progress to chronic myocardial ischemia.

- *Left Ventricular Hypertrophy (LVH).* LVH might develop as a compensatory response to the increased workload from aortic or mitral valve disorders, hypertension, or the progression of heart failure.

- *Hypertensive Disease.* Hypertension might either pre-exist or develop secondary to chronic kidney insufficiency and heart disorders. The interplay between renal and cardiac function, alongside the dyslipidemias, creates a fertile ground for hypertensive disease.

Each of these future medical problems reflects the complex interactions and the cumulative burden of cardiovascular and renal conditions in the patient's medical history. Early identification, monitoring, and management of these potential future issues are crucial for improving the patient's prognosis and quality of life.

Patient 5 This patient has a complex medical history involving cardiac issues, which may predispose them to a variety of other health conditions in the future. Our model correctly identifies CKI within the top-3 predictions as one of the possible future disorders. Below is a step-by-step explanation of how each medical problem listed might be related to the patient in the future based on their past medical history:

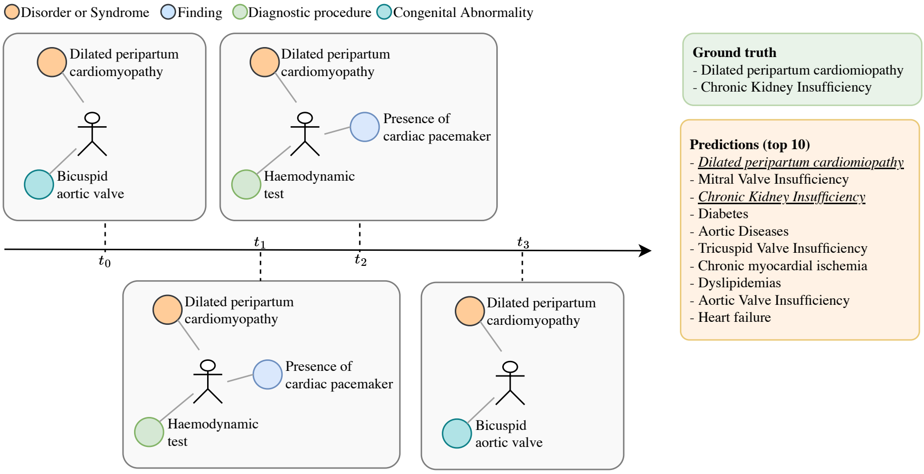

Disorder or Syndrome
Finding
Diagnostic procedure
Congenital Abnormality
Dilated peripartum cardiomyopathy
Bicuspid aortic valve
Dilated peripartum cardiomyopathy
Presence of cardiac pacemaker
Haemodynamic test
t_0
t_1
t_2
t_3
Dilated peripartum cardiomyopathy
Presence of cardiac pacemaker
Haemodynamic test
Dilated peripartum cardiomyopathy
Bicuspid aortic valve
Ground truth
- Dilated peripartum cardiomiopathy
- Chronic Kidney Insufficiency
Predictions (top 10)
- *Dilated peripartum cardiomiopathy*
- Mitral Valve Insufficiency
- *Chronic Kidney Insufficiency*
- Diabetes
- Aortic Diseases
- Tricuspid Valve Insufficiency
- Chronic myocardial ischemia
- Dyslipidemias
- Aortic Valve Insufficiency
- Heart failure

Figure 12.12. Example of prediction from a given medical history (n. 5)

- *Dilated Peripartum Cardiomyopathy (DPCM).* This condition is already part of the patient's medical history. It's a form of heart failure that occurs during pregnancy or in the postpartum period, leading to dilation and poor contraction of the heart chambers.

- *Mitral Valve Insufficiency.* The dilation and impaired contraction of the heart chambers in DPCM could adversely affect the function of the mitral valve, leading to mitral valve insufficiency, where the valve does not close properly and allows blood to flow backward into the heart.

- *Chronic Kidney Insufficiency.* Heart and kidney function are closely related. Poor cardiac function from DPCM or other cardiac issues can lead to decreased blood flow to the kidneys, potentially resulting in chronic kidney insufficiency.

- *Diabetes.* While not directly related to the cardiac issues, people with cardiovascular disease may have a higher risk of developing diabetes, possibly due to shared risk factors like obesity and hypertension.

- *Aortic Diseases.* The patient has a congenital abnormality, Bicuspid

Aortic Valve, which could predispose them to aortic diseases, like aortic stenosis or aortic aneurysm.

- *Tricuspid Valve Insufficiency.* Similar to mitral valve insufficiency, the dilation of heart chambers in DPCM could affect the tricuspid valve's function leading to tricuspid valve insufficiency.

- *Chronic Myocardial Ischemia.* DPCM and other structural heart issues could lead to inadequate blood supply to the heart muscle over time, resulting in chronic myocardial ischemia.

- *Dyslipidemias.* Dyslipidemias are often associated with cardiovascular diseases. Although not directly linked to the current conditions of the patient, the presence of heart disease might be associated with or exacerbated by lipid metabolism disorders.

- *Aortic Valve Insufficiency.* The Bicuspid Aortic Valve can lead to aortic valve insufficiency, where the valve does not close properly, allowing some blood to flow back into the heart.

- *Heart Failure.* The history of DPCM and other cardiac abnormalities significantly increase the risk of future heart failure, where the heart is unable to pump blood effectively to meet the needs of the body.

12.3 Conclusion

In this chapter, the intricate process of applying Information Extraction methods to the Campania Salute (CS) network Italian dataset was meticulously explored, culminating in the creation of Temporal Knowledge Graphs (TKGs), one for each patient in the database. These TKG serves as a pivotal tool, bridging the gap between vast datasets and actionable insights, particularly in the realm of patient medical histories. MedTKG, i.e. the TKG-based predictive method proposed in Chapter 10, has been applied and its predictive capabilities concerning potential adverse events for patients have been assessed. The experimental results, which underscore the proficiency model in forecasting forthcoming disorders, have proven the robustness of the approach and its potential implications in the healthcare domain.

Part IV

Conclusions

In the medical field, much of the data exists in written form, making it challenging to analyze. The aim of this thesis is to enable computer systems to interpret and utilize this data to assist healthcare professionals and patients. This thesis introduces novel methods that allow computers to process and comprehend medical texts, particularly in languages with limited digital resources. Once interpreted, the data integrates into knowledge graphs, that enable the design of systems to predict potential health issues a patient may encounter in the future.

In Chapter 2, the potential of Artificial Intelligence in revolutionizing precision medicine has been emphasized. The chapter introduced the importance of Knowledge Graphs in this realm, particularly when grappling with sparse data availability in niche areas like healthcare or in languages with constrained resources, exemplified by the Italian context. Chapter 3 set the groundwork for this thesis by clarifying the fundamental principles of Knowledge Graphs. As we explored detailed aspects like Named Entity Recognition, Entity Linking, Relation Extraction, and Adverse Events Prediction, the depth and breadth of this research began to unfold.

Few-shot learning, a critical theme for languages and domains with limited resources, took center stage in Chapter 4. The demand for effective Named Entity Recognition with minimal annotated data in complex areas such as healthcare was put forth, illuminating the challenges and the need for innovative solutions. The practical efficacy of pre-trained transformer models in the realm of Named Entity Recognition, particularly for Spanish data, was showcased in Chapter 5. This chapter underscored the achievements of transformer architectures in the BioASQ Disease Text Mining challenge, highlighting the power and potential of these pre-trained models in real-world applications.

Data augmentation, a cornerstone for bolstering model performance, was thoroughly investigated in Chapter 6. Through a novel methodology based on context similarity, this chapter accentuated the importance of generating plausible augmented samples, minimizing noise, and ensuring a consistent improvement over leading baselines. Building upon this, Chapter 7 introduced an advanced methodology for selecting the most pertinent samples for enhancing Named Entity Recognition. This policy-based active learning framework sought to prioritize the most impactful augmented samples, shedding light on the untapped potential of data augmentation.

While data augmentation methods have been designed and tested for few-shot scenarios, there exist many annotated datasets for the recognition of medical mentions from unstructured text. However, since they are very costly due to the time needed and the expert knowledge needed for the annotation efforts, they usually specialize for one single entity type. Chapter 8 addressed a pivotal challenge: the fusion of multiple single-entity biomedical datasets into a unified model. TaughtNet offers a solution, showcasing the strength of knowledge distillation and multi-task learning.

With the focus shifting to relation extraction in Chapter 9, a multi-task learning paradigm, grounded in transformer-based architectures, was articulated. This model harnessed the strengths of shared encoding layers, amplifying performance even in scenarios constrained by limited data.

Chapter 10 introduced the novel Temporal Knowledge Graph framework, MedTKG. By amalgamating dynamic medical history data from Electronic Health Records with static information from medical ontologies, MedTKG promised a powerful tool for predicting future disorders in patients. In this chapter, the framework has been tested on MIMIC-III (Medical Information Mart for Intensive Care III), which is a large, freely available database comprising de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset was developed by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT).

Chapters 11 and 12 brought the research home, focusing on the real-world application of the techniques developed. Using an extensive Italian dataset from the CampaniaSalute network provided by the Department of Advanced Biomedical Sciences at the University of Naples Federico II, the process of information extraction from clinical notes was elaborated upon. Then, MedTKG has been trained on the resulting temporal knowledge graphs representing patients medical histories. The results, validated with domain experts, revealed the efficacy of the methodologies in extracting relevant insights and predicting potential disorders.

While the contributions of this thesis are manifold and profound, future work might be devoted to handle several limitations.

The MedTKG models presented in this study possess limited generalizability as they were trained specifically on the MIMIC-III and CampaniaSalute datasets, which focus on critical care and hypertensive units respectively. Consequently, these models may not perform optimally outside their respective training domains. Future research should explore integrating datasets from various hospital departments to develop a more universally applicable model.

Furthermore, while we have shown that the utilization of static relationships between medical concepts produces an enhancement of model performance in predicting future disorders, the influence of such static information has the potential to be even greater. Notably, association networks between genes, disorders, symptoms, treatments, and other related factors are being developed by many researchers [156, 76, 70]. Integrating these networks into the system could further amplify its performance.

Another field of promising exploration is undoubtedly the extraction of valuable information from specialized healthcare unstructured data in low-resource languages, with a focus on employing Large Language Models (LLMs). Recent advancements have showcased the significant potential of LLMs in zero-shot contexts, which is particularly compelling for the healthcare sector. A notable instance is the utilization of LLMs like InstructGPT for zero- and few-shot information extraction from clinical text, despite the models not being specifically trained for the clinical domain [7].

Furthermore, in the realm of healthcare diagnostics, the integration of Explainable Artificial Intelligence (XAI) stands as a promising, and necessary, frontier. The complexity and critical nature of medical decision-making demand not only accuracy but also transparency in AI-driven systems. In light of this, being easy-visualizable and explorable, knowledge graphs offer a good potential for explainability, as proven by current literature [6]. Thus, future research could focus on developing XAI models that provide clinicians with comprehensible insights into the decision-making process of AI algorithms.

Additionally, the fusion of XAI with patient-specific data interpretation could pave the way for more personalized healthcare, where AI not only diagnoses but also explains variations in disease manifestation and treatment responses among individuals. This human-centric approach in AI could revolutionize the way healthcare providers interact with technology, making it a collaborative tool rather than a mysterious black box

In conclusion, this thesis demonstrates the significant potential of Artificial Intelligence (AI) in advancing the field of precision medicine. Despite existing challenges and unexplored areas, the forward trajectory, guided by the findings and methodologies of this research, is ripe with opportunities. The integration of AI and healthcare, as showcased in this work, extends beyond mere technological innovation; it signifies a substantial promise for improved healthcare delivery. This amalgamation holds the potential to usher in an era of personalized, accurate, and proactive healthcare solutions.

Bibliography

[1] David Ifeoluwa Adelani, Michael A. Hedderich, Dawei Zhu, Esther van den Berg, and Dietrich Klakow. Distant Supervision and Noisy Label Learning for Low Resource Named Entity Recognition: A Study on Hausa and Yor\‘ub\’a. *arXiv:2003.08370 [cs]*, March 2020. arXiv: 2003.08370.

[2] Alireza Afradi and Arash Ebrahimabadi. Comparison of artificial neural networks (ann), support vector machine (svm) and gene expression programming (gep) approaches for predicting tbm penetration rate. *SN Applied Sciences*, 2:1–16, 2020.

[3] Alireza Afradi and Arash Ebrahimabadi. Prediction of tbm penetration rate using the imperialist competitive algorithm (ica) and quantum fuzzy logic. *Innovative Infrastructure Solutions*, 6(2):103, 2021.

[4] Alireza Afradi, Arash Ebrahimabadi, and Tahereh Hallajian. Prediction of tunnel boring machine penetration rate using ant colony optimization, bee colony optimization and the particle swarm optimization, case study: Sabzkooh water conveyance tunnel. *Mining of Mineral Deposits*, 14(2):75–84, 2020.

[5] Alireza Afradi, Arash Ebrahimabadi, and Tahereh Hallajian. Prediction of tbm penetration rate using fuzzy logic, particle swarm optimization and harmony search algorithm. *Geotechnical and Geological Engineering*, pages 1–24, 2021.

[6] Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10, 2022.

[7] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David A. Sontag. Large language models are few-shot clinical information extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang,

editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1998–2022. Association for Computational Linguistics, 2022.

[8] Ilseyar Alimova and Elena Tutubalina. Multiple features for clinical relation extraction: A machine learning approach. *Journal of Biomedical Informatics*, 103:103382, 2020.

[9] Héctor Martínez Alonso and Barbara Plank. When is multitask learning effective? semantic sequence prediction under varying data conditions. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 44–53, Online, 2017. Association for Computational Linguistics.

[10] Emily Alsentzer, J. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings. *ArXiv*, abs/1904.03323, 2019.

[11] Nasser Alshammari and Saad Alanazi. The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal*, 22(3):295–302, 2021.

[12] Umar Asif, Jianbin Tang, and Stefan Harrer. Ensemble knowledge distillation for learning improved and efficient networks. In Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang, editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 953–960. IOS Press, 2020.

[13] Parmida Atighehchian, Frederic Branchaud-Charron, Jan Freyberg, Rafael Pardinas, Lorne Schell, and George Pearse. Baal, a bayesian active learning library. `https://github.com/baal-org/baal/`, 2022.

[14] P. Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63:1 – 27, 2017.

[15] Maria-Florina Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *J. Comput. Syst. Sci.*, 75:78–89, 2009.

[16] Ilaria Bartolini, Vincenzo Moscato, Marco Postiglione, Giancarlo Sperlì, and Andrea Vignali. COSINER: context similarity data augmentation for named entity recognition. In Tomás Skopal, Fabrizio Falchi, Jakub Lokoc, Maria Luisa Sapino, Ilaria Bartolini, and Marco Patella, editors, *Similarity Search and Applications - 15th International Conference, SISAP 2022, Bologna, Italy, October 5-7, 2022, Proceedings*, volume 13590 of *Lecture Notes in Computer Science*, pages 11–24. Springer, 2022.

[17] Ilaria Bartolini, Vincenzo Moscato, Marco Postiglione, Giancarlo Sperlì, and Andrea Vignali. Data augmentation via context similarity: An application to biomedical named entity recognition. *Information Systems*, 119:102291, 2023.

[18] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *EMNLP/IJCNLP*, 2019.

[19] Asma Ben Abacha and Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of biomedical semantics*, 2(5):1–11, 2011.

[20] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *ICML '09*, 2009.

[21] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270, 2004.

[22] I. Bondarenko, S. Berezin, A. Pauls, T. Batura, Y. Rubtsova, and B. Tuchinov. Using Few-Shot Learning Techniques for Named Entity Recognition and Relation Extraction. In *2020 Science and Artificial Intelligence conference (S.A.I.ence)*, pages 58–65, November 2020.

[23] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795, 2013.

[24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin,

Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[26] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *NeurIPS*, 2020.

[27] Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343, Online, July 2020. Association for Computational Linguistics.

[28] Yixin Cao, Zikun Hu, Tat-seng Chua, Zhiyuan Liu, and Heng Ji. Low-resource name tagging learned with weakly labeled data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 261–270, Hong Kong, China, November 2019. Association for Computational Linguistics.

[29] Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021.

[30] R. Caruana. Multitask learning. In *Encyclopedia of Machine Learning and Data Mining*, 1998.

[31] Yevgen Chebotar and Austin Waters. Distilling knowledge from ensembles of neural networks for speech recognition. In *INTERSPEECH*, 2016.

[32] Chenhua Chen and Yue Zhang. Learning how to self-learn: Enhancing self-training using neural reinforcement learning. *2018 International Conference on Asian Language Processing (IALP)*, pages 25–30, 2018.

[33] Miao Chen, Ganhui Lan, Fang Du, and Victor S. Lobanov. Joint learning with pre-trained transformer on named entity recognition and relation extraction tasks for clinical analytics. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 234–242, Online, 2020. Association for Computational Linguistics.

[34] Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. Data augmentation for cross-domain named entity recognition. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5346–5356. Association for Computational Linguistics, 2021.

[35] Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[36] Yen-Chun Chen, Zhe Gan, Yu Cheng, J. Liu, and Jing jing Liu. Distilling knowledge learned in bert for text generation. In *ACL*, 2020.

[37] Jason P. C. Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

[38] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 787–795. ACM, 2017.

[39] Nigel Collier and Jin-Dong Kim. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLP-BA/BioNLP)*, pages 73–78, Geneva, Switzerland, August 28th and 29th 2004. COLING.

[40] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics, 2020.

[41] Ryan Cotterell and Kevin Duh. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[42] Gamal K. O. Crichton, Sampo Pyysalo, Billy Chiu, and A. Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18, 2017.

[43] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online, August 2021. Association for Computational Linguistics.

[44] Wendy W Dai, Dmitriy Dligach, and Steven J Bethard. Synthetic patient generation using generative adversarial networks. In *Machine Learning for Healthcare Conference*, pages 157–168. PMLR, 2018.

[45] Xiang Dai and Heike Adel. An analysis of simple data augmentation for named entity recognition. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3861–3867. International Committee on Computational Linguistics, 2020.

[46] Xiang Dai and Heike Adel. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[47] Sandipan Dandapat and Andy Way. Improved Named Entity Recognition using Machine Translation-based Cross-lingual Information. *Computación y Sistemas*, 20(3):495–504, September 2016.

[48] S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *COLT*, 2005.

[49] Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha P. Talukdar. Hyte: Hyperplane-based temporally aware knowledge graph embedding. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2001–2011. Association for Computational Linguistics, 2018.

[50] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[51] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[53] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online, November 2020. Association for Computational Linguistics.

[54] Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Informatics*, 47:1–10, 2014.

[55] Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

[56] RI Doğan, R Leaman, and Z. Lu. NCBI disease corpus: a resource for disease name recognition and concept normalization. https://pubmed.ncbi.nlm.nih.gov/24393765/, 2014.

[57] Artur Evangelista, Marta Sitges, Guillaume Jondeau, Robin Nijveldt, Mauro Pepi, Hug Cuéllar, Gianluca Pontone, Eduardo Bossone, Maarten Groenink, Marc Richard Dweck, Jolien W. Roos-Hesselink, L. Mazzolai, Roland R J van Kimmenade, Victor Aboyans, and José F. Rodríguez-Palomares. Multimodality imaging in thoracic aortic diseases: a clinical consensus statement from the european association of cardiovascular imaging and the european society of cardiology working group on aorta and peripheral vascular diseases. *European heart journal. Cardiovascular Imaging*, 2023.

[58] Meng Fang, Yuan Li, and Trevor Cohn. Learning how to active learn: A deep reinforcement learning approach. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 595–605. Association for Computational Linguistics, 2017.

[59] Chelsea Finn, A. Rajeswaran, Sham M. Kakade, and S. Levine. Online meta-learning. In *ICML*, 2019.

[60] Joseph Fisher and Andreas Vlachos. Merge and label: A novel neural network architecture for nested NER. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5840–5850. Association for Computational Linguistics, 2019.

[61] Evan French and Bridget T. McInnes. An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics*, 137:104252, 2023.

[62] Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data. *arXiv:1704.06360 [cs]*, April 2017. arXiv: 1704.06360.

[63] Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 993–1000, Limassol Cyprus, April 2019. ACM.

[64] J. D. Frizzell, L. Liang, P. J. Schulte, C. W. Yancy, P. A. Heidenreich, A. F. Hernandez, and E. D. Peterson. Prediction of 30-day all-cause readmissions

in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA cardiology*, 2(2):204–209, 2017.

[65] T. Fukuda, Masayuki Suzuki, Gakuto Kurata, S. Thomas, Jia Cui, and B. Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *INTERSPEECH*, 2017.

[66] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR, 2017.

[67] Valerio La Gatta, Vincenzo Moscato, Marco Postiglione, and Giancarlo Sperlì. Few-shot named entity recognition with cloze questions, 2021.

[68] Farhad Soleimanian Gharehchopogh and ZA Khalifehlou. Study on information extraction methods from text mining and natural language processing perspectives. *AWER Procedia Information Technology & Computer Science*, 1:1321–1327, 2012. https://www.academia.edu/download/55319254/881-5088-2-PB.pdf.

[69] Rishab Goel, Seyed Mehran Kazemi, Marcus A. Brubaker, and Pascal Poupart. Diachronic embedding for temporal knowledge graph completion. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3988–3995. AAAI Press, 2020.

[70] Janet Piñero González, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I. Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, 48(Database-Issue):D845–D855, 2020.

[71] Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. Meta-learning for low-resource neural machine translation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3622–3631. Association for Computational Linguistics, 2018.

[72] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse

effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892, 2012.

[73] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, 2020.

[74] M. Habibi, Leon Weber, Mariana L. Neves, David Luis Wiegandt, and U. Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33:i37 – i48, 2017.

[75] Maryam Habibi, Leon Weber, Mariana Neves, David L. Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. In *Bioinformatics*, volume 33, pages i37–i48. Oxford Academic, 2017.

[76] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, 33(Database-Issue):514–517, 2005.

[77] Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. Revisiting self-training for neural sequence generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[78] Yongquan He, Peng Zhang, Luchen Liu, Qi Liang, Wenyuan Zhang, and Chuang Zhang. HIP network: Historical information passing network for extrapolation reasoning on temporal knowledge graph. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1915–1921. ijcai.org, 2021.

[79] Matthew Henderson and Ivan Vulić. ConVEx: Data-efficient and few-shot slot labeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3375–3389, Online, June 2021. Association for Computational Linguistics.

[80] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *Journal of biomedical informatics*, 46 5:914–20, 2013.

[81] Geoffrey E. Hinton, Oriol Vinyals, and J. Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

[82] Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo J. Nevado-Holgado. Few-shot learning for named entity recognition in medical text. *CoRR*, abs/1811.05468, 2018.

[83] Lixiang Hong, Jinjian Lin, Shuya Li, Fangping Wan, Hui Yang, Tao Jiang, Dan Zhao, and Jianyang Zeng. A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. *Nature Machine Intelligence*, 2(6):347–355, 2020.

[84] Lixiang Hong, Jinjian Lin, Shuya Li, Fangping Wan, Hui Yang, Tao Jiang, Dan Zhao, and Jianyang Zeng. A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. *Nat. Mach. Intell.*, 2(6):347–355, 2020.

[85] Ali Hosseinalipour, Farhad Soleimanian Gharehchopogh, Mohammad Masdari, and Ali Khademi. A novel binary farmland fertility algorithm for feature selection in analysis of the text psychology. *Appl. Intell.*, 51(7):4824–4859, 2021.

[86] Ali Hosseinalipour, Farhad Soleimanian Gharehchopogh, Mohammad Masdari, and Ali Khademi. Toward text psychology analysis using social spider optimization algorithm. *Concurrency and Computation: Practice and Experience*, 33(17):e6325, 2021.

[87] Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393, Online, July 2020. Association for Computational Linguistics.

[88] Yutai Hou, Cheng Chen, Xianzhen Luo, Bohan Li, and Wanxiang Che. Inverse is better! fast and accurate prompt for few-shot slot tagging. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 637–647, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[89] Yutai Hou, Zhihan Zhou, Yijia Liu, Ning Wang, Wanxiang Che, Han Liu, and Ting Liu. Few-Shot Sequence Labeling with Label Dependency Transfer and Pair-wise Embedding. *arXiv:1906.08711 [cs]*, September 2019. arXiv: 1906.08711.

[90] Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. Attention-guided answer distillation for machine reading comprehension. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October*

31 - November 4, 2018, pages 2077–2086. Association for Computational Linguistics, 2018.

[91] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. Few-shot named entity recognition: An empirical baseline study. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10408–10423. Association for Computational Linguistics, 2021.

[92] Kexin Huang, Jaan Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342, 2019.

[93] Ming-Siang Huang, Po-Ting Lai, Pei-Yen Lin, Yu-Ting You, Richard Tzong-Han Tsai, and Wen-Lian Hsu. Biomedical named entity recognition and linking datasets: survey and our recent development. *Briefings in Bioinformatics*, 21(6):2219–2238, 06 2020.

[94] Minlie Huang, Xiaoyan Zhu, and Ming Li. A hybrid method for relation extraction from biomedical literature. *International journal of medical informatics*, 75(6):443–455, 2006.

[95] Po-Sen Huang, Chenglong Wang, Rishabh Singh, Wen-tau Yih, and Xiaodong He. Natural language to structured query generation via meta-learning. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 732–738. Association for Computational Linguistics, 2018.

[96] Lesley A. Inker, Nwamaka D. Eneanya, Josef Coresh, Hocine Tighiouart, Dan Wang, Yingying Sang, Deidra C. Crews, Alessandro Doria, Michelle M. Estrella, Marc Froissart, Morgan E. Grams, Tom Greene, Anders Grubb, Vilmundur G. Gudnason, Orlando M. Gutiérrez, Roberto S. N. Kalil, Amy B. Karger, Michael Mauer, Gerjan J Navis, Robert G. Nelson, Emilio D. Poggio, Roger A. Rodby, Peter Rossing, Andrew D. Rule, Elizabeth Selvin, Jesse C. Seegmiller, Michael G. Shlipak, Vicente E. Torres, Wei-Chun Yang, Shoshana H. Ballew, Sara Couture, Neil R. Powe, and Andrew S. Levey. New creatinine- and cystatin c-based equations to estimate gfr without race. *The New England journal of medicine*, 2021.

[97] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4163–4174. Association for Computational Linguistics, 2020.

[98] Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. Recurrent event network: Autoregressive structure inferenceover temporal knowledge graphs. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6669–6683. Association for Computational Linguistics, 2020.

[99] Yufang Jin and Jing Luo. Acl anthology: A digital archive of research papers in computational linguistics. In *ACL*, 2019.

[100] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[101] Muhammad Raza Khan, M. Ziyadi, and M. Abdelhady. Mt-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers. *ArXiv*, abs/2001.08904, 2020.

[102] Hamed Khataei Maragheh, Farhad Soleimanian Gharehchopogh, Kambiz Majidzadeh, and Amin Babazadeh Sangar. A new hybrid based on long short-term memory network with spotted hyena optimization algorithm for multi-label text classification. *Mathematics*, 10(3):488, 2022.

[103] Munui Kim, Seung Han Baek, and Min Song. Relation extraction for biological pathway construction using node2vec. *BMC Bioinform.*, 19-S(8):75–84, 2018.

[104] Sungchul Kim, Kristina Toutanova, and H. Yu. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *ACL*, 2012.

[105] Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL, 2014.

[106] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1317–1327. The Association for Computational Linguistics, 2016.

[107] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[108] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[109] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille, 2015.

[110] Hyoun-Joong Kong. Managing unstructured big data in healthcare system. *Healthcare Informatics Research*, 25:1 – 2, 2019.

[111] Michal Konkol and Miloslav Konopík. Segment representations in named entity recognition. In Pavel Král and Václav Matousek, editors, *Text, Speech, and Dialogue - 18th International Conference, TSD 2015, Pilsen,Czech Republic, September 14-17, 2015, Proceedings*, volume 9302 of *Lecture Notes in Computer Science*, pages 61–70. Springer, 2015.

[112] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard JB Dobson. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit, 2020.

[113] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina M. Ibrahim, James T. Teo, and Richard J. B. Dobson. Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit. *Artif. Intell. Medicine*, 117:102083, 2021.

[114] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James T. Teo, and Richard J. B. Dobson. Medgpt: Medical concept prediction from clinical narratives. *CoRR*, abs/2107.03134, 2021.

[115] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Dong-Hong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, S. V. Ramanan, P. Senthil Nathan, Slavko Zitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber Ahmad Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin M. Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, K. E. Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, C Ata, Tolga Can, Anabel Usie, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzábal, and Alfonso Valencia. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2 – S2, 2015.

[116] Jens Kringelum, Sonny Kim Kjærulff, Søren Brunak, Ole Lund, Tudor I. Oprea, and Olivier Taboureau. Chemprot-3.0: a global chemical biology diseases mapping. *Database J. Biol. Databases Curation*, 2016, 2016.

[117] Jason Krone, Yi Zhang, and Mona Diab. Learning to classify intents and slot labels given a handful of examples. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 96–108, Online, July 2020. Association for Computational Linguistics.

[118] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[119] Timothée Lacroix, Guillaume Obozinski, and Nicolas Usunier. Tensor decompositions for temporal knowledge base completion. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[120] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, K. Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *HLT-NAACL*, 2016.

[121] Ouyu Lan, Xiao Huang, Bill Yuchen Lin, He Jiang, Liyuan Liu, and Xiang Ren. Learning to contextually aggregate multi-source supervision for sequence labeling. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2134–2146. Association for Computational Linguistics, 2020.

[122] Julien Leblay and Melisachew Wudage Chekol. Deriving validity time in knowledge graph. In Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1771–1776. ACM, 2018.

[123] Changki Lee, Yi-Gyu Hwang, and Myung-Gil Jang. Fine-grained named entity recognition and relation extraction for question answering. In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 799–800, online, 2007. ACM.

[124] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240, 2020.

[125] David D. Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. *SIGIR Forum*, 29(2):13–19, 1995.

[126] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *SIGIR '94*, 1994.

[127] P. Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *CLINICALNLP*, 2020.

[128] Chuanjiang Li, Shaobo Li, Huan Wang, Fengshou Gu, and Andrew D Ball. Attention-based deep meta-transfer learning for few-shot fine-grained fault diagnosis. *Knowledge-Based Systems*, page 110345, 2023.

[129] Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinform.*, 18(1):198:1–198:11, 2017.

[130] Guohao Li, Matthias Müller, Ali K. Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9266–9275. IEEE, 2019.

[131] J. Li, A. Sun, J. Han, and C. Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[132] J Li, Y Sun, RJ Johnson, D Sciaky, CH Wei, R Leaman, AP Davis, CJ Mattingly, TC Wiegers, and Z. Lu. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4860626/, 2016.

[133] J. Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, A. P. Davis, C. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016, 2016.

[134] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016, 2016.

[135] Jing Li, Shuo Shang, and Ling Shao. MetaNER: Named Entity Recognition with Meta-Learning. In *Proceedings of The Web Conference 2020*, pages 429–440, Taipei Taiwan, April 2020. ACM.

[136] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.*, 34(1):50–70, 2022.

[137] Qingqing Li, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, Jian Wang, Liang Yang, Kan Xu, and Yijia Zhang. A multi-task learning based approach to biomedical entity relation extraction. In Huiru Jane Zheng, Zoraida Callejas, David Griol, Haiying Wang, Xiaohua Hu, Harald H. H. W. Schmidt, Jan Baumbach, Julie Dickerson, and Le Zhang, editors, *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, Madrid, Spain, December 3-6, 2018*, pages 680–682, Online, 2018. IEEE Computer Society.

[138] Y. Li, Yongxin Yang, W. Zhou, and Timothy M. Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, 2019.

[139] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaïne, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi Khorshidi. BEHRT: transformer for electronic health records. *CoRR*, abs/1907.09538, 2019.

[140] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Networks Learn. Syst.*, 33(12):6999–7019, 2022.

[141] Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. Temporal knowledge graph reasoning based on evolutional representation learning. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 408–417. ACM, 2021.

[142] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. BOND: bert-assisted open-domain named entity recognition with distant supervision. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash, editors, *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1054–1064. ACM, 2020.

[143] Angli Liu, Jingfei Du, and Veselin Stoyanov. Knowledge-Augmented Language Model and Its Application to Unsupervised Named-Entity Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1142–1150, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[144] Bing Liu, Xuchu Yu, Anzhu Yu, Pengqiang Zhang, Gang Wan, and Ruirui Wang. Deep few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2290–2304, 2019.

[145] Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Fangzheng Xu, Huan Gui, Jian Peng, and Jiawei Han. Empower sequence labeling with task-aware neural language model. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5253–5260. AAAI Press, 2018.

[146] Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041, Online, November 2020. Association for Computational Linguistics.

[147] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.

[148] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Improving multi-task deep neural networks via knowledge distillation for natural language understanding, 2019.

[149] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding, July 2019.

[150] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics.

[151] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[152] Yuang Liu, W. Zhang, and Jijie Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020.

[153] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2395–2405, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[154] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[155] Y. Lou, T. Qian, F. Li, and D. Ji. A Graph Attention Model for Dictionary-Guided Named Entity Recognition. *IEEE Access*, 8:71584–71592, 2020. Conference Name: IEEE Access.

[156] Kezhi Lu, Kuo Yang, Hailong Sun, Qian Zhang, Qiguang Zheng, Kuan Xu, Jianxin Chen, and Xuezhong Zhou. Sympgan: A systematic knowledge integration system for symptom-gene associations network. *Knowl. Based Syst.*, 276:110752, 2023.

[157] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. KAME: knowledge-based attention model for diagnosis prediction in healthcare. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang, editors, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 743–752. ACM, 2018.

[158] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bidirectional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[159] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264, 2020.

[160] Elham Rahimzadeh Mahalleh and Farhad Soleimanian Gharehchopogh. An automatic text summarization based on valuable sentences selection. *International Journal of Information Technology*, 14(6):2963–2969, 2022.

[161] Stefano Marchesin and Gianmaria Silvello. TBGA: a large-scale gene-disease association dataset for biomedical relation extraction. *BMC Bioinform.*, 23(1):111, 2022.

[162] Authors/Task Force members, Raimund Erbel, Victor Aboyans, Catherine Boileau, Eduardo Bossone, Roberto Di Bartolomeo, Holger Eggebrecht, Arturo Evangelista, Volkmar Falk, Herbert Frank, Oliver Gaemperli, Martin Grabenwöger, Axel Haverich, Bernard Iung, Athanasios John Manolis, Folkert Meijboom, Christoph A. Nienaber, Marco Roffi, Hervé Rousseau, Udo Sechtem, Per Anton Sirnes, Regula S. von Allmen, Christiaan J.M. Vrints, ESC Committee for Practice Guidelines (CPG), Jose Luis Zamorano, Stephan Achenbach, Helmut Baumgartner, Jeroen J. Bax, Héctor Bueno, Veronica Dean, Christi Deaton, Çetin Erol, Robert Fagard, Roberto Ferrari, David Hasdai, Arno Hoes, Paulus Kirchhof, Juhani Knuuti, Philippe Kolh, Patrizio Lancellotti, Ales Linhart, Petros Nihoyannopoulos, Massimo F. Piepoli, Piotr Ponikowski, Per Anton Sirnes, Juan Luis Tamargo, Michal Tendera, Adam Torbicki, William Wijns, Stephan Windecker, Document reviewers, Petros Nihoyannopoulos, Michal Tendera, Martin Czerny, John Deanfield, Carlo Di Mario, Mauro Pepi, Maria Jesus Salvador Taboada, Marc R. van Sambeek, Charalambos Vlachopoulos, Jose Luis Zamorano, Michael Grimm, Oktay Musayev, Agnès Pasquet, Zumreta Kušljugić, Maja Cikes, Georgios P. Georghiou, Josef Stasek, Henning Molgaard, Sirje Kõvask;, Ville Kytö, Guillaume Jondeau, Zviad Bakhutashvili, Yskert von Kodolitsch, Costas Tsioufis, András Temesvári, Ronen Rubinshtein, Francesco Antonini-Canterin, Olga Lunegova, Peteris Stradins, Elie Chammas, Regina Jonkaitiene, Andrew Cassar, Knut Bjørnstad, Kazimierz Widenka, Miguel Sousa Uva, Daniel Lighezan, Jovan Perunicic, Juraj Madaric, Isidre Vilacosta, Magnus Bäck, Abdallah Mahdhaoui, Recep Demirbag, and Ivan Kravchenko. 2014 ESC Guidelines on the diagnosis and treatment of aortic diseases: Document covering

acute and chronic aortic diseases of the thoracic and abdominal aorta of the adultThe Task Force for the Diagnosis and Treatment of Aortic Diseases of the European Society of Cardiology (ESC). *European Heart Journal*, 35(41):2873–2926, 08 2014.

[163] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995.

[164] Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online, July 2020. Association for Computational Linguistics.

[165] Marvin Minsky. Computation: Finite and infinite machines. 1967.

[166] Mike D. Mintz, Steven Bills, R. Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL/IJCNLP*, 2009.

[167] Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast, editors, *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 179–203. CEUR-WS.org, 2022.

[168] Tom M. Mitchell. *Machine learning, International Edition.* McGraw-Hill Series in Computer Science. McGraw-Hill, 1997.

[169] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533, 2015.

[170] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning.* Adaptive computation and machine learning. MIT Press, 2012.

[171] Vincenzo Moscato, Giuseppe Napolano, Marco Postiglione, and Giancarlo Sperlì. Multi-task learning for few-shot biomedical relation extraction. *Artif. Intell. Rev.*, 56(11):13743–13763, 2023.

[172] Vincenzo Moscato, Marco Postiglione, Carlo Sansone, and Giancarlo Sperlí. Taughtnet: Learning multi-task biomedical named entity recognition from single-task teachers. *IEEE J. Biomed. Health Informatics*, 27(5):2512–2523, 2023.

[173] Vincenzo Moscato, Marco Postiglione, Guido Secondulfo, Giancarlo Sperlí, and Andrea Vignali. Learning how to augment data: An application to biomedical NER. In Zina M. Ibrahim, Honghan Wu, and Nirmalie Wiratunga, editors, *Proceedings of the 6th International Workshop on Knowledge Discovery from Healthcare Data co-located with 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), Macao, China, August 20, 2023*, volume 3479 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.

[174] Aldrian Obaja Muis and Wei Lu. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2608–2618. Association for Computational Linguistics, 2017.

[175] Subhabrata Mukherjee and Ahmed Hassan Awadallah. Uncertainty-aware self-training for text classification with few labels. *ArXiv*, abs/2006.15315, 2020.

[176] T. B. Murdoch and A. S. Detsky. The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352, 2013.

[177] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.*, 54(1), feb 2021.

[178] Kamel Nebhi. A rule-based relation extraction system using dbpedia and syntactic parsing. In Sebastian Hellmann, Agata Filipowska, Caroline Barrière, Pablo N. Mendes, and Dimitris Kontokostas, editors, *Proceedings of the NLP & DBpedia workshop co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 22, 2013*, volume 1064 of *CEUR Workshop Proceedings*, Online, 2013. CEUR-WS.org.

[179] Isar Nejadgholi, Kathleen C. Fraser, and Berry de Bruijn. Extensive error analysis and a learning-based evaluation of medical entity recognition systems to approximate user experience. In *BIONLP*, 2020.

[180] K. Y. Ngiam and I. W. Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273, 2019.

[181] Jian Ni, Georgiana Dinu, and Radu Florian. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1470–1480. Association for Computational Linguistics, 2017.

[182] A. Obamuyide and A. Vlachos. Model-agnostic meta-learning for relation classification with limited supervision. In *ACL*, 2019.

[183] F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51, 2003.

[184] Cennet Oguz and Ngoc Thang Vu. Few-shot learning for slot tagging with attentive relational network. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1566–1572, Online, April 2021. Association for Computational Linguistics.

[185] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[186] Namyong Park, Fuchen Liu, Purvanshi Mehta, Dana Cristofor, Christos Faloutsos, and Yuxiao Dong. Evokg: Jointly modeling event time and network structure for reasoning over temporal knowledge graphs. In K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang, editors, *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 794–803. ACM, 2022.

[187] Hima Patel, Shanmukha C. Guttula, Ruhi Sharma Mittal, Naresh Manwani, Laure Berti-Équille, and Abhijit Manatkar. Advances in exploratory data analysis, visualisation and quality for data centric AI systems. In Aidong Zhang and Huzefa Rangwala, editors, *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 4814–4815. ACM, 2022.

[188] Nanyun Peng and Mark Dredze. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[189] Yifan Peng, Qingyu Chen, and Zhiyong Lu. An empirical study of multi-task learning on BERT for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 205–214, Online, July 2020. Association for Computational Linguistics.

[190] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 58–65, Online, 2019. Association for Computational Linguistics.

[191] Roberto Poli, Maria Pia di Buono, and Carlo Aliprandi. Challenges and opportunities in biomedical text mining in italian. *Journal of Biomedical Informatics*, 2021.

[192] Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50 – 50, 2006.

[193] Kun Qian and Z. Yu. Domain adaptive dialog generation via meta learning. In *ACL*, 2019.

[194] Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy, July 2019. Association for Computational Linguistics.

[195] A. Rajeswaran, Chelsea Finn, Sham M. Kakade, and S. Levine. Meta-learning with implicit gradients. In *NeurIPS*, 2019.

[196] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[197] Lance A Ramshaw and Mitchell P Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer, 1999.

[198] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *CoRR*, abs/2005.12833, 2020.

[199] A. Ratnaparkhi and M. Marcus. Maximum entropy models for natural language ambiguity resolution. 1998.

[200] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990, Online, 2019. Association for Computational Linguistics.

[201] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP*, 2020.

[202] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9):180:1–180:40, 2022.

[203] Esteban Safranchik, Shiying Luo, and Stephen H. Bach. Weakly supervised sequence tagging from noisy rules. In *AAAI*, 2020.

[204] Sunil Sahu and Ashish Anand. Recurrent neural network models for disease name recognition using domain invariant features. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2216–2225, Berlin, Germany, August 2016. Association for Computational Linguistics.

[205] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL, 2003.

[206] Erik F. Tjong Kim Sang and Jorn Veenstra. Representing text chunks. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 173–179. The Association for Computer Linguistics, 1999.

[207] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[208] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269, Online, 2021. Association for Computational Linguistics.

[209] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, April 2021. Association for Computational Linguistics.

[210] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269, Online, 2021. Association for Computational Linguistics.

[211] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer, 2018.

[212] J. Schmidhuber. On learning how to learn learning strategies. 1994.

[213] Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, L. E. S. Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, E. Paraiso, D. Teodoro, and Claudia Maria Cabral Moro Barra. Biobertpt - a portuguese neural language model for clinical named entity recognition. In *ClinicalNLP@EMNLP*, 2020.

[214] N. Schork. Personalized medicine: Time for one-person trials. *Nature*, 520:609–611, 2015.

[215] Stefan Schweter. Italian bert and electra models, November 2020.

[216] H. J. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. Inf. Theory*, 11:363–371, 1965.

[217] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, page 1070–1079, USA, 2008. Association for Computational Linguistics.

[218] H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In David Haussler, editor, *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992, Pittsburgh, PA, USA, July 27-29, 1992*, pages 287–294. ACM, 1992.

[219] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-end structure-aware convolutional networks for knowledge base completion. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3060–3067. AAAI Press, 2019.

[220] Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. Learning named entity tagger using domain-specific dictionary. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2054–2064. Association for Computational Linguistics, 2018.

[221] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5953–5959. ijcai.org, 2019.

[222] Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[223] Jin Shin and Sang Joon Kim. A mathematical theory of communication. 2006.

[224] Amit Singhal. Introducing the knowledge graph: Things, not strings, 2012. Official Google Blog.

[225] L. Smith, L.K. Tanabe, and R.J.n. Ando et al. The BioCreative II - Critical Assessment for Information Extraction in Biology Challenge. https://doi.org/10.1186/gb-2008-9-s2-s2, 2008.

[226] Larry L. Smith, Lorraine K. Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, C. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence E. Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter W. Adriaans, Christian Blaschke, Rafael Torres, Mariana L. Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur.

Overview of biocreative ii gene mention recognition. *Genome Biology*, 9:S2 – S2, 2008.

[227] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[228] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087, 2017.

[229] Eugenio Stabile, Raffaele Izzo, Francesco Rozza, Maria Angela Losi, Nicola De Luca, and Bruno Trimarco. Hypertension survey in italy: Novel findings from the campania salute network. *High Blood Pressure & Cardiovascular Prevention*, 24:363–370, 2017.

[230] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9120–9132, Online, 2020. PMLR.

[231] Pedro Ortiz Suarez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. 2019.

[232] S. Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *EMNLP/IJCNLP*, 2019.

[233] Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485, Hong Kong, China, November 2019. Association for Computational Linguistics.

[234] Zhiqing Sun, H. Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *ACL*, 2020.

[235] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017,*

Sydney, NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

[236] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1199–1208, Online, 2018. Computer Vision Foundation / IEEE Computer Society.

[237] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1701–1708. IEEE Computer Society, 2014.

[238] Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[239] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. Blockmix: Meta regularization and self-calibrated inference for metric-based meta-learning. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 610–618, Online, 2020. ACM.

[240] S. Thrun and L. Y. Pratt. Learning to learn. 1998.

[241] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[242] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

[243] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, 2001.

[244] Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7:92 – 92, 2005.

[245] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX:433–460, 1950.

[246] Özlem Uzuner, B. South, Shuying Shen, and S. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18 5:552–6, 2011.

[247] Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J. Am. Medical Informatics Assoc.*, 18(5):552–556, 2011.

[248] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[249] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[250] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Michal Walczak, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.

[251] Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin. A neural transition-based model for nested mention recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1011–1017. Association for Computational Linguistics, 2018.

[252] M. Wang and Christopher D. Manning. Cross-lingual pseudo-projected expectation regularization for weakly supervised learning. *ArXiv*, abs/1310.1597, 2013.

[253] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.*, 29(12):2724–2743, 2017.

[254] Shanshan Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. Order-free medicine combination prediction with graph convolutional reinforcement learning. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1623–1632. ACM, 2019.

[255] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[256] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. KGAT: knowledge graph attention network for recommendation. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 950–958. ACM, 2019.

[257] Xuan Wang, Y. Zhang, Xiang Ren, Yuhao Zhang, M. Zitnik, Jingbo Shang, C. Langlotz, and Jiawei Han. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35 10:1745–1752, 2019.

[258] Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis P. Langlotz, and Jiawei Han. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinform.*, 35(10):1745–1752, 2019.

[259] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[260] Yu-Xiong Wang, Ross B. Girshick, M. Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7278–7286, 2018.

[261] Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1–15, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[262] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics.

[263] Patricia L. Whetzel, Natasha Noy, N. Shah, P. Alexander, Csongor Nyulas, T. Tudorache, and M. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39:W541 – W545, 2011.

[264] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[265] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[266] Joseph Worsham and Jugal Kalita. Multi-task learning for natural language processing in the 2020s: Where are we going? *Pattern Recognit. Lett.*, 136:120–126, 2020.

[267] Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[268] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015.

[269] Chengjin Xu, Mojtaba Nayyeri, Fouad Alkhoury, Hamed Shariat Yazdi, and Jens Lehmann. TeRo: A time-aware knowledge graph embedding via temporal rotation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1583–1593, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[270] Chenjin Xu, Mojtaba Nayyeri, Fouad Alkhoury, Hamed Shariat Yazdi, and Jens Lehmann. Temporal knowledge graph completion based on time series gaussian embedding. In Jeff Z. Pan, Valentina A. M. Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal, editors, *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part I*, volume 12506 of *Lecture Notes in Computer Science*, pages 654–671. Springer, 2020.

[271] Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[272] Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[273] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. Transfer learning for sequence tagging with hierarchical recurrent networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[274] David Yarowsky and G. Ngai. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL*, 2001.

[275] Rui Ye, Xin Li, Yujie Fang, Hongyu Zang, and Mingzhong Wang. A vectorized relational graph convolutional network for multi-relational network alignment. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4135–4141. ijcai.org, 2019.

[276] Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. Data augmentation for spoken language understanding via joint variational generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7402–7409, 2019.

[277] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics*, 20(S10), May 2019.

[278] Houjin Yu, Xian-Ling Mao, Zewen Chi, Wei Wei, and Heyan Huang. A robust and domain-adaptive approach for low-resource named entity recognition. In Enhong Chen and Grigoris Antoniou, editors, *2020 IEEE International Conference on Knowledge Graph, ICKG 2020, Online, August 9-11, 2020*, pages 297–304. IEEE, 2020.

[279] Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. Counterfactual Generator: A Weakly-Supervised Method for Named Entity Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online, November 2020. Association for Computational Linguistics.

[280] Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. Counterfactual generator: A weakly-supervised method for named entity recognition. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7270–7280. Association for Computational Linguistics, 2020.

[281] Ye Zhang, Matthew Lease, and Byron C. Wallace. Active discriminative text representation learning. In *AAAI*, 2017.

[282] Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Shaowu Zhang, Yuanyuan Sun, and Liang Yang. A hybrid model based on neural networks for biomedical relation extraction. *Journal of Biomedical Informatics*, 81:83–92, 2018.

[283] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.

[284] Deyu Zhou, Lei Miao, and Yulan He. Biomedical relation extraction: From binary to complex. *Computational and Mathematical Methods in Medicine*, 2019, 2019.

[285] Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. Melm: Data augmentation with masked entity language modeling for low-resource ner. In *Proceedings of the 60th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2251–2262, 2022.

[286] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Author's Publications

2023

C12 Cirillo, M., Moscato, V., Postiglione, M. (2023). PicusLab @ BC8 SympTEMIST track: Disambiguating Entity Linking Candidates with Question Answering. BioCreative VIII @ AMIA 2023.

J11 Bartolini, I., Moscato, V., Postiglione, M., Sperlì, G., Vignali, A. (2023). Data augmentation via context similarity: An application to biomedical Named Entity Recognition. Information Systems, 119, ISSN 0306-4379, https://doi.org/10.1016/j.is.2023.102291.

C11 Bosco, A., Capuozzo, S., Celano, B., Gravina, M., Marrone, S., Maurelli, M. P., Moscato, V., Pontillo, G., Postiglione, M., Rinaldi, A. M., Rinaldi, L., Russo, C., Sperlì, G., Tommasino, C., Cringoli, G., Sansone, C., AI in healthcare: Activities of the University of Naples Federico II node of the CINI-AIIS Lab, Ital-IA 2023. https://ceur-ws.org/Vol-3486/127.pdf

C10 Riccio, G., Romano, A., Korsun, A., Cirillo, M., Postiglione, M., La Gatta, V., Ferraro, A., Galli, A., Moscato, V., Healthcare Data Summarization via Medical Entity Recognition and Generative AI. The 2nd Italian Conference on Big Data and Data Science (ITADATA)

C9 Postiglione, M., Esposito, G., Izzo, R., La Gatta, V., Moscato, V., Piccolo, R. (2023). Harnessing multi-modality and expert knowledge for adverse events prediction in clinical notes. 1st International Workshop on Multi-Modal Medical Imaging Processing (M3IP) @ ICIAP.

C8 Moscato, V., Postiglione, M., Secondulfo, G., Sperlì, G., Vignali, A. (2023). Learning How To Augment Data: An Application To Biomedical NER. 6th International Workshop on Knowledge Discovery from Healthcare Data (KDH) @ IJKAI. https://ceur-ws.org/Vol-3479/paper6.pdf

J10 Moscato, V., Postiglione, M., & Sperlí, G. (2023). Few-shot Named Entity Recognition: definition, taxonomy and research directions. ACM Transactions on Intelligent Systems and Technology. https://doi.org/10.1145/3609483

J9 Ferraro, A., Galli, A., La Gatta, V., Postiglione, M. (2023). Benchmarking Open Source and Paid Services for Speech to Text: An Analysis of Quality and Input Variety. Frontiers in Big Data, 6. 10.3389/fdata.2023.1210559

J8 Moscato, V., Napolano, G., Postiglione, M., & Sperlí, G. (2023). Multitask learning for few-shot biomedical relation extraction. Artificial Intelligence Review, 1-21. https://doi.org/10.1007/s10462-023-10484-6

J7 La Gatta, V., Moscato, V., Postiglione, M. & Sperlí, G. (2023). COVID-19 Sentiment Analysis Based on Tweets. IEEE Intelligent Systems, 38, 51-55. https://doi.org/10.1109/MIS.2023.3239180

C7 Ferraro, A., Galli, A., La Gatta, V., Moscato, V., Postiglione, M., Sperlì, G., Amato, F. (2023). HEMR: Hypergraph Embeddings for Music Recommendation. 31st Symposium on Advanced Database Systems. https://ceur-ws.org/Vol-3478/paper46.pdf

C6 Ferraro, A., Galli, A., La Gatta, V., Moscato, V., Postiglione, M., Sperlì, G., Moscato, F. (2023). Unsupervised Anomaly Detection in Predictive Maintenance using Sound Data. 31st Symposium on Advanced Database Systems. https://ceur-ws.org/Vol-3478/paper53.pdf

J6 Moscato, V., Postiglione, M., Sansone, C., & Sperlí, G. (2023). TaughtNet: Learning Multi-Task Biomedical Named Entity Recognition From Single-Task Teachers. IEEE Journal of Biomedical and Health Informatics, 27, 2512-2523. https://doi.org/10.1109/JBHI.2023.3244044

2022

J5 D'Auria, D., Moscato, V., Postiglione, M., Romito, G., & Sperlí, G. (2022). Improving graph embeddings via entity linking: A case study on Italian clinical notes. Intell. Syst. Appl., 17, 200161. https://doi.org/10.1016/j.iswa.2022.200161

J4 La Gatta, V., Moscato, V., Pennone, M., Postiglione, M., & Sperlí, G. (2022). Music Recommendation via Hypergraph Embedding. IEEE transactions on neural networks and learning systems, PP. https://doi.org/10.1109/TNNLS.2022.3146968

C5 Moscato, V., Postiglione, M., & Sperlí, G. (2022). Biomedical Spanish Language Models for entity recognition and linking at BioASQ DisTEMIST. Conference and Labs of the Evaluation Forum. https://ceur-ws.org/Vol-3180/paper-22.pdf

C4 Bartolini, I., Moscato, V., Postiglione, M., Sperlí, G., & Vignali, A. (2022). COSINER: COntext SImilarity data augmentation for Named Entity Recognition. Similarity Search and Applications. https://doi.org/10.1007/978-3-031-17849-8_2

C3 Ferraro, A., Galli, A., Gatta, V.L., & Postiglione, M. (2022). A Deep Learning pipeline for Network Anomaly Detection based on Autoencoders. 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), 260-264. https://doi.org/10.1109/MetroXRAINE54828.2022.9967598

2021

J3 La Gatta, V., Moscato, V., Postiglione, M., & Sperlí, G. (2021). CASTLE: Cluster-aided space transformation for local explanations. Expert Syst. Appl., 179, 115045. https://doi.org/10.1016/j.eswa.2021.115045

J2 La Gatta, V., Moscato, V., Postiglione, M., & Sperlí, G. (2021). PASTLE: Pivot-aided space transformation for local explanations. Pattern Recognition Letters, 149, 67-74. https://doi.org/10.1016/j.patrec.2021.05.018

C2 Postiglione, M. (2021). Towards an Italian Healthcare Knowledge Graph. Similarity Search and Applications. https://doi.org/10.1007/978-3-030-89657-7_29

C1 Cinque, M., Moscato, V., Postiglione, M., Riccio, M. P. (2021). Diagnosing severity levels of Autism Spectrum Disorder with Machine Learning. Data-centric AI Workshop @ NeurIPS. https://datacentricai.org/neurips21/papers/69_CameraReady__DataCentricAI____Autism.pdf

2020

J1 La Gatta, V., Moscato, V., Postiglione, M., & Sperlí, G. (2020). An Epidemiological Neural Network Exploiting Dynamic Graph Structured Data Applied to the COVID-19 Outbreak. IEEE Transactions on Big Data, 7, 45-55. https://doi.org/10.1109/TBDATA.2020.3032755