UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

PH.D. THESIS

IN

INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

SAFEGUARDING PRIVACY THROUGH DEEP LEARNING TECHNIQUES

ROSARIO CATELLI

TUTOR: PROF. VALENTINA CASOLA

CO-TUTOR: DR. MASSIMO ESPOSITO

COORDINATOR: PROF. DANIELE RICCIO

XXXIII CICLO

SCUOLA POLITECNICA E DELLE SCIENZE DI BASE
DIPARTIMENTO DI INGEGNERIA ELETTRICA E TECNOLOGIE DELL'INFORMAZIONE

To Whom It May Concern

Acknowledgements

Writing thank yous has never been my strong point, but that does not mean there are not people who deserve them, in fact I might forget someone and that would be a problem.

Starting with the fundamentals, I definitely have to thank my tutor, Prof. Valentina Casola, who led me down this bumpy road despite my hot head. As well as my co-tutor, Dr. Massimo Esposito, who helped me get my car back on the road after the wheels got flat.

And could not miss the thanks to my parents, Alfredo and Silvia, and my brother Christian, even though we had to endure more than usual due to smartworking and lockdowns caused by the COVID-19 pandemic: what a great accomplishment not to have resorted to firearms! All kidding aside, thank you for giving me the opportunity to get this far, as far as you knew I wanted to go.

Thanks to friends, even work colleagues (the list would be long), who have often changed my day and brightened the road with a single word, and especially dear Genny who at the intersection told me where to go.

A special thanks also goes to my in-laws, Donato and Tina, who have often taken care of many things that I, also for time reasons, could not and, overall, for supporting me in my choices despite uncertainty. And also that cousin-in-law I have, Salvatore, for being there and the stupid things we did together.

I thank my grandmother Maria and my grandmother-in-law Filomena, who will never read this page, for their popular proverbs of arcane memory and dubious understanding, unfailing in the topical moments of life.

Last but most important, my sweetheart Roberta: she always broke my back, to push me and make me do the things (according to her right!) that I had to do and, I must admit, in the end she was always right. Probably because she knows me better than I know

myself and makes crystal clear the real me that sometimes lies to himself.

And finally, forgive me, a moment of narcissism: I would like to thank myself. It is true that when you hit rock bottom you can always start scratching, but you tell yourself that a lot because you know where it comes from: stagger on, but never give up.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last few years, there has been a growing need to meet minimum security and privacy requirements. Both public and private companies have had to comply with increasingly stringent standards, such as the ISO 27000 family of standards, or the various laws governing the management of personal data. The huge amount of data to be managed has required a huge effort from the employees who, in the absence of automatic techniques, have had to work tirelessly to achieve the certification objectives. Unfortunately, due to the delicate information contained in the documentation relating to these problems, it is difficult if not impossible to obtain material for research and study purposes on which to experiment new ideas and techniques aimed at automating processes, perhaps exploiting what is in ferment in the scientific community and linked to the fields of ontologies and artificial intelligence for data management. In order to bypass this problem, it was decided to examine data related to the medical world, which, especially for important reasons related to the health of individuals, have gradually become more and more freely accessible over time, without affecting the generality of the proposed methods, which can be reapplied to the most diverse fields in which there is a need to manage privacy-sensitive information. In particular, in order to better circumscribe the problems addressed, in Section 1.1 the context and motivations behind the research carried out are introduced, then in Section 1.2 the open challenges faced in the field of interest and finally in Section 1.3 the contributions made to scientific research, providing a detailed list of publications made during this doctoral programme in Section 1.3.1.

1.1 Context and motivation

In recent years, the availability of textual clinical data in electronic form, known as Electronic Health Records (EHRs) and from which further information can be extracted to manage various critical health situations, has grown significantly. However, in order to be able to use such data, it is necessary to respect the restrictions on the privacy of individual patients, as outlined by the relevant legislation and imposed by both national and supranational privacy authorities: in the United States the current law in force is the Health Insurance Portability and Accountability Act (HIPAA)[1] while in the European Union (EU) there are both the General Data Protection Regulation (GDPR)[2] and several national legislations generally more restrictive but also less precise in indicating the exact procedures to follow.

In detail, a fundamental step to allow the sharing and publication of health data is the so called *de-identification*, widely used in the medical area and termed as *clinical de-identification*, which aims to avoid the disclosure of personal identities. But, in order to exploit health information for research purposes, it is necessary to aim at generalisation through surrogated terms rather than the deletion of privacy-sensitive information contained in medical records, safeguarding in this way also the readability of the documentation (Vincze and Farkas 2014). After proper de-identification, hence anonymisation of the data, it is possible to release and share them publicly.

According to HIPAA[3], there are two possible methods of de-identification:

- Expert Determination. It requires the employment of a human domain expert and it is a manual and really work intensive task.

- Safe Harbor. It can be automated since it defines 18 relevant identifiers, listed in Table 1.1, that must be removed and/or replaced with plausible and realistic surrogates. These identifiers are also called *Protected Health Information* (PHI) identifiers.

[1] https://www.hhs.gov/hipaa

[2] https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu

[3] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

Table 1.1: Excerpt from 45 CFR §164.514

#	**PHI Identifiers**
(A)	Names;
(B)	All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
(C)	All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
(D)	Telephone numbers;
(E)	Fax numbers;
(F)	Electronic mail addresses;
(G)	Social security numbers;
(H)	Medical record numbers;
(I)	Health plan beneficiary numbers;
(J)	Account numbers;
(K)	Certificate/license numbers;
(L)	Vehicle identifiers and serial numbers, including license plate numbers;
(M)	Device identifiers and serial numbers;
(N)	Web Universal Resource Locators (URLs);
(O)	Internet Protocol (IP) address numbers;
(P)	Biometric identifiers, including finger and voice prints;
(Q)	Full face photographic images and any comparable images; and
(R)	Any other unique identifying number, characteristic, or code [...]

For the sake of completeness, it must be said that the EU GDPR utilises the wider concept of *personal data* which includes PHI plus other sensitive data such as racial or ethnic origin and religion, proving to be more restricting in EU citizens' data collection and its usage, but it does not provide equally clear guidance on processes for removing such information.

Anyway, the de-identification process slows down the spread of publicly available health data sets. Consequently, researchers are committed to improving de-identification methods also to help the world of medical research. For instance, the COrona VIrus Disease 19 (COVID-19) is still a global threat opposed by experts, politicians and researchers from around the world (Hernandez-Matamoros et al. 2020). In particular, everyone is rushing to keep pace with the influx of potentially relevant studies related to COVID-19 in order to gain timely knowledge to manage the current pandemic (Røst et al. 2020) and the availability of these new studies has led to an exponential increase in the amount of textual clinical data to be analysed: unfortunately, this data cannot be used directly for medical investigations, due to the privacy restrictions as mentioned earlier.

1.2 Open challenges

In recent years, several communities have pushed towards progress with the organisation of challenges (e.g. i2b2[4] and ShARe/CLEF eHealth Evaluation Lab[5]) in the de-identification field, encouraging the development of systems to automate the task, whose progress has gone from rudimentary rule-based techniques to techniques based on machine learning first and deep learning later. In this way the problem of de-identification has benefited from the use of Natural Language Processing (NLP) techniques such as Named Entity Recognition (NER), a task that aims to identify certain entities within texts: in this specific case, entities have been assimilated to PHI identifiers, which are handled as the entities to be de-identified and then made anonymous through appropriate surrogates.

In detail, such techniques were firstly based on handcrafted rules

[4]Informatics for Integrating Biology and the Bedside (i2b2) then National NLP Clinical Challenges (n2c2) at https://portal.dbmi.hms.harvard.edu/

[5]ShARe/CLEF eHealth Evaluation Lab then CLEF eHealth Evaluation Lab at https://clefehealth.imag.fr/

to identify PHI entities, resulting simple to implement but less flexible with regard to both context and language changes, then based on machine learning to train a classifier to recognise PHI entities and their different types, but requiring large labelled data sets and time to carry out feature engineering (Stephane M Meystre et al. 2010; Stubbs, Kotfila, and Uzuner 2015).

Recently, deep learning-based systems have been used to perform sequence labeling, hence identify and classify PHI entities, leveraging large data sets to learn both the right features to be used and the best word representation in a numerical space (Yadav and Bethard 2018). However, these systems suffer from some limitations from a word representation perspective. In detail, they represent words as numerical vectors, statically pre-trained on large corpora and able to capture hidden information about a language, like word analogies or semantics. They treat words as atomic and rely on the distributional hypothesis (i.e. words with similar contexts have similar meanings), with the side-effect of worsening the quality in word representations for rarely observed or out-of-vocabulary words as well as for morpho-syntactic variations typical of handwritten text. In addition, these systems are not able to handle the polysemous and context-dependent nature of words, and treat each sentence as a single instance, reducing the representative power given by contextual information within the whole clinical text.

In this respect, it is important to make a clarification. The identification and removal of a PHI may seem to be the main tasks of the de-identification process but this is not actually the case, because it is also necessary to classify PHI correctly: after de-identification, anonymisation is often necessary and this process benefits from the substitution of personal data rather than its deletion for two main reasons, as stated by Vincze and Farkas 2014. The first lies in the readability of the text, which is preserved using appropriate surrogates, and the second in the best result obtained with the same imperfect de-identification: if some data will not be de-identified, it will be more difficult to distinguish real pieces of PHI from surrogates. Therefore, the ability of a de-identification system to analyse the context plays an important role: the better the classification of entities the better the result of anonymisation.

Furthermore, the language domain of interest, i.e. English due to a greater worldwide availability of Electronic Health Records

(EHRs), was taken for granted and de-identification challenges were organised by i2b2 group, founder in English. Unfortunately, experiences in languages other than English remained confined to a few sporadic cases, such as ShARe/CLEF eHealth Evaluation Lab and IberLEF 2019[6] with specific traces also in French (some Information Extraction tasks) and Spanish (the Medical Document Anonymisation track, also known as MEDDOCAN track), as well as a few case studies in other languages. Hence, outside the Anglo-Saxon-speaking countries, the use of the best performing deep learning methods is severely limited both by the lack of resources suitable for their exploitation, i.e. large data sets, and by poor experimentation on such languages, which are consequently defined as low-resource languages.

1.3 Thesis contributions

With this doctoral thesis work, an effort has been made to respond to the various open challenges outlined in the previous section.

Firstly, the best performing sequence labeling architecture has been used, i.e. a Bi-LSTM + CRF network, which exploits both the sequence modeling capacities of the Bidirectional Long Short-Term Memory (Bi-LSTM) network and sequence labeling abilities of the Conditional Random Field (CRF) to predict the target PHI entities. From this point of view, the main contributions can be summarised as follows:

- the Flair contextualised and character-level language model (Akbik, Blythe, and Vollgraf 2018) has been employed in order to represent input words and respectively (1) capture the meanings associated to the same word in various contexts of use, i.e. the polysemy of the word, and (2) better grasp, interpret and manage both morpho-syntactic variations, i.e. the structures of words, such as endings and prefixes, and misspelled and/or rare words, to which a collection of handwritten notes is subject, so as better classify entities;

- the enhancement of the Flair representation power concatenating and then stacking its embeddings with a classic word representation able to better capture the latent syntactic and se-

[6] https://sites.google.com/view/iberlef-2019

mantic similarities. In particular, the GloVe word embeddings (Pennington, Socher, and Manning 2014) have been chosen for being used in this work;

- the grouping of more sentences together and their usage as input instances for the network to broaden the context of representation and, thus, improve the learning capabilities, leveraging at its best the memory capacity of the Bi-LSTM + CRF architecture.

Several tests were conducted whose experimental results were analysed: the distribution of errors showed performances comparable or superior to the state of the art without the need to engineer rules. Moreover, through ablation analysis it was possible to further confirm the validity of the proposed solution for the English language.

Secondly, a scientific contribution regarding the positioning of the Italian language in the clinical de-identification scenario has been made and, to reach this goal, three objectives have been pursued:

- the first objective consisted in the creation of a new data set for clinical de-identification in Italian proposed for the first time to the scientific community in this work: starting from the COVID-19 medical records made available to the public in pdf format by the Italian Society of Radiology (SIRM)[7], the data were manually annotated according to i2b2 criteria (Stubbs and Uzuner 2015);

- the second objective consisted in the construction, on the top of the best performing sequence labelling architecture recognised by scientific literature, i.e. a Bidirectional Long Short-Term Memory (Bi-LSTM) plus Conditional Random Field (CRF) model (Huang, W. Xu, and Yu 2015), of a stacked form of word representation, not yet experimented for the clinical de-identification scenario in Italian, exploiting:

  - the Flair contextualised and character-level language model (Akbik, Blythe, and Vollgraf 2018) to represent input words;

[7] https://www.sirm.org/

  - FastText sub-word embeddings (Bojanowski et al. 2017) in order to better capture both the latent syntactic and semantic similarities;

- the third objective consisted in the execution of several experiments to verify the performance of the models previously described in comparison with BERT (Devlin et al. 2019), a Transformer (Vaswani et al. 2017) based architecture, which is considered the state-of-the-art language model for many NLP general tasks and also the NER one (Li et al. 2020), which includes the particular case of de-identification.

These tests have verified the effectiveness of different ways of functioning, for example statically or contextually and at character, sub-word or word level, on the Italian language which, even with an alphabet similar to the English one, presents a wide syntactic and morphological variety. In detail, the stacked embedding consisting of FastText and Flair has reached the best performance for the Italian de-identification scenario: the combined ability of handling context, polysemy and morpho-syntactic variations given by Flair and analysis at sub-word level given by FastText has surpassed the other models tested.

Thirdly, an improvement of some aspects of the scientific literature has been attempted, experimenting the new data set based on COVID-19 medical records for a low resource language like Italian. The aim is to investigate the ability of cross-linguistic methods to transfer knowledge between different languages while retaining the features necessary to correctly perform NER, which is the basis of de-identification and anonymisation systems. As far as is known, there are no multilingual approaches specifically designed for this task, nor any knowledge of the performance of existing systems with respect to the Italian language, which is the subject of study. Two different system architectures have been tested that showed state of the art performance. To this end, the i2b2 2014 training data set in English was used and, in accordance with the i2b2 annotation guidelines (Stubbs and Uzuner 2015), it was extended with the Italian data set created from the COVID-19 medical records provided by the Italian Society of Radiology[8] (SIRM). Different training approaches have been

[8]https://www.sirm.org/

tested, both monolingual in English with zero-shot test on Italian, and cross-language with mixed language training and test on Italian. The results were promising and allowed to identify the best architectural solution for low-resource Italian language cases for the clinical de-identification task. The application of the method described here would allow a better de-identification of Italian COVID-19 medical records, speeding up their public dissemination: more accurate anonymisation of privacy-sensitive information reduces the distrust of institutions to release data.

1.3.1 Publications

In different ways, each of these works contributed to outline my doctoral path, answering a few questions and asking new ones. But, above all, these works suggested the direction to take from time to time, in some cases similar to the previous one and in others completely different, encouraging a continuous tension while looking for solutions to problems.

The papers published in support of this doctoral thesis are listed below, divided into journal or conference articles, listed in order of acceptance from the most recent (top) to the least recent (bottom).

Journal Papers

- Rosario Catelli, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito (2021). "A Novel COVID-19 Data Set and an Effective Deep Learning Approach for the De-Identification of Italian Medical Records". In: *IEEE Access* 9, pp. 19097–19110. DOI: 10.1109/ACCESS.2021.3054479. URL: https://doi.org/10.1109/ACCESS.2021.3054479

- Marco Pota, Mirko Ventura, Rosario Catelli, and Massimo Esposito (2021). "An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian". In: *Sensors* 21.1, p. 133. DOI: 10.3390/s21010133. URL: https://doi.org/10.3390/s21010133

- Rosario Catelli, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito (2021). "Combining

contextualized word representation and sub-document level analysis through Bi-LSTM+CRF architecture for clinical de-identification". In: *Knowl. Based Syst.* 213, p. 106649. DOI: 10.1016/j.knosys.2020.106649. URL: https://doi.org/10.1016/j.knosys.2020.106649

- Rosario Catelli, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito (2020). "Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set". In: *Appl. Soft Comput.* 97.Part, p. 106779. DOI: 10.1016/j.asoc.2020.106779. URL: https://doi.org/10.1016/j.asoc.2020.106779

- Hassan Mokalled, Rosario Catelli, Valentina Casola, Daniele Debertol, Ermete Meda, and Rodolfo Zunino (2020). "The Guidelines to Adopt an Applicable SIEM Solution". In: *Journal of Information Security* 11.01, pp. 46–70. DOI: 10.4236/jis.2020.111003. URL: https://doi.org/10.4236%2Fjis.2020.111003

Conference Papers

- Valentina Casola and Rosario Catelli (Nov. 2020). "Semantic Management of Enterprise Information Systems through Ontologies". In: *Computer Science & Information Technology (CS & IT)*. AIRCC Publishing Corporation. DOI: 10.5121/csit.2020.101403. URL: https://doi.org/10.5121%2Fcsit.2020.101403

- Valentina Casola, Rosario Catelli, and Alessandra De Benedictis (2019). "A First Step Towards an ISO-Based Information Security Domain Ontology". In: *28th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2019, Naples, Italy, June 12-14, 2019*. Ed. by Sumitra Reddy. IEEE, pp. 334–339. DOI: 10.1109/WETICE.2019.00075. URL: https://doi.org/10.1109/WETICE.2019.00075

- Hassan Mokalled, Rosario Catelli, Valentina Casola, Daniele Debertol, Ermete Meda, and Rodolfo Zunino (2019). "The Ap-

plicability of a SIEM Solution: Requirements and Evaluation". In: *28th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2019, Naples, Italy, June 12-14, 2019*. Ed. by Sumitra Reddy. IEEE, pp. 132–137. DOI: 10.1109/WETICE.2019.00036. URL: https://doi.org/10.1109/WETICE.2019.00036

Chapter 2

Background

This chapter reviews all the scientific literature supporting this work. Section 2.1 provides an overview of the automated systems used in clinical de-identification, then an in-depth examination of systems based on rules and machine learning is reported in Section 2.1.1 while the more recent ones based on deep learning techniques are described in section 2.1.2. In Section 2.2 the most recent conceptual developments in language modeling are illustrated, in Section 2.3 approaches for languages other than English are described. Finally, Section 2.4 introduces cross-lingual transfer learning approaches, dividing between non BERT-based multilingual techniques in Section 2.4.1 and BERT-based multilingual techniques in Section 2.4.2.

2.1 Clinical de-identification through automated systems

A PHI, as a unit of information, is similar to what the literature indicates as a *named entity* (Nadeau and Sekine 2009). In the NLP field, the NER is the process by which such entities are recognized, specifically *clinical NER* when working on clinical notes (i.e. unstructured text within EHRs): the information contained therein may be critical for medical investigations, but it is necessary to de-identify it before it can be used, even if for research purposes. The subsequent process of replacing identified PHI with valid surrogates is called *anonymization*: it is no longer sufficient to identify an entity but the system must also be able to classify it in order to replace it correctly, so it is more correct to talk about Named Entity Recognition and

Classification (NERC) (Nadeau and Sekine 2009).

De-identification process can be manual or automated. De-identification is manual when human annotators are required to label PHI. This approach, as reported by Dernoncourt, J. Y. Lee, Uzuner, et al. 2017, has three main weaknesses:

1. crowd-sourcing of the activity is not possible because only a small group of people have access to the identified patient notes;

2. humans can make mistakes;

3. humans are expensive.

To try to solve these problems, automated systems have been developed and, from a historical and literature point of view, two major phases can be distinguished from the beginning to the present, whose watershed lies in the use of deep learning techniques, so there are systems and techniques based on deep learning and systems and techniques that do not use it, the latter extensively described by Stephane M Meystre et al. 2010 and Stubbs, Kotfila, and Uzuner 2015. More recently, the promising deep learning systems have been started being applied also to other languages different from English.

2.1.1 Pre-deep learning systems

Early NER systems used rule-based handcrafted algorithms, like Sweeney 1996. Leveraging not only rules but also specialized semantic dictionaries, gazetteers and patterns, these systems tried to identify PHI instances in EHRs like discharge summaries and laboratory reports (Friedlin and McDonald 2008) or X-ray reports (Neamatullah et al. 2008), or also pathologies (Thomas et al. 2002; Dilip Gupta, Saul, and Gilbertson 2004; Beckwith et al. 2006). They were easy to implement and did not require labeled data (except for system evaluation) or publicly available annotated data sets: this facilitated their dissemination. But they were not without flaws: fine-tuning was necessary for every change in the data set and both language changes (e.g. typing errors, abbreviations, variations) and word context (e.g. ”Mr. Parkinson’s” is PHI, ”Parkinson’s disease” is not) were not taken into account. For more details see Stephane M Meystre et al. 2010.

Different systems with pros and cons were proposed in NLP challenges, but from the outset it became clear that there was a strong limiting aspect associated with languages other than English. In fact, several automation tools were created (Neamatullah et al. 2008), used successfully (Tu et al. 2010) but hardly adaptable to languages other than English (Velupillai et al. 2009; Grouin, Rosier, et al. 2009).

Nonetheless, rule-based systems (Guillen et al. 2006) could accurately recognize PHI instances based on repetitive formulas (e.g. fax or telephone numbers, e-mail), but their complexity increased significantly when such formulas began to be missing (e.g. names or places), reducing their effectiveness. Instead, methods based on machine learning (Szarvas, Farkas, and Busa-Fekete 2007; T. Chen, Cullen, and Godwin 2015; He et al. 2015) were able to achieve very good results as long as they had samples both in sufficient numbers and rich in features during training, so they were poor in complex and rare cases. Finally, hybrid systems (Wellner et al. 2007; Dehghan et al. 2015; Z. Liu, Y. Chen, et al. 2015; H. Yang and Garibaldi 2015) were able to achieve the best performance, detecting entities even in cases of scarcity of data and complex features, by combining the advantages of their predecessors, provided they also took on the disadvantages and time required to develop such a sophisticated system. Machine Learning (ML) algorithms used in these systems can be divided into two main categories. On the one hand, the decision tree (Freund and Schapire 1995) and the Support Vector Machine (SVM) (Hearst 1998), which modeled de-identification as a classification problem. On the other hand, the conditional random field (CRF) (Lafferty, McCallum, and Pereira 2001), the Hidden Markov Model (HMM) (Eddy 1996) and the structural SVM (SSVM) (Xue, S. Chen, and Q. Yang 2008) that modeled de-identification as a sequence labeling problem: these systems were more efficient than the former due to the exploitation of dependencies between neighboring labels and, among these, the CRF was the best, also in the de-identification field (for instance, see He et al. 2015). At the heart of machine learning there were hand-crafted features engineering, e.g. dictionary features (Dehghan et al. 2015; Z. Liu, Y. Chen, et al. 2015; H. Yang and Garibaldi 2015), N-grams, part-of-speech (POS), word vector features (Z. Liu, Y. Chen, et al. 2015; Tang, Cao, et al. 2013), and data set management in pre-processing (Z. Liu, Y. Chen,

et al. 2015) and post-processing (H. Yang and Garibaldi 2015). For a complete overview see Stubbs, Kotfila, and Uzuner 2015.

2.1.2 Deep learning systems

To go beyond machine learning-based systems limitations, researchers have moved forward deep learning area. Deep learning-based NER algorithms and systems (Huang, W. Xu, and Yu 2015; Chiu and Nichols 2016; Lample, Ballesteros, et al. 2016; Ma and E. H. Hovy 2016) have been produced and improved studying their behavior, then they have been applied to clinical NER systems (Dernoncourt, J. Y. Lee, Uzuner, et al. 2017; Z. Liu, Tang, et al. 2017).

These NER systems are based on two main building blocks: one is the embedding (Mikolov et al. 2013) that is the numerical method to represent words (or characters) and give a manageable input to systems, and one other is the neural network structure itself that leverages recurrent neural networks (RNNs) (Elman 1990; Goller and Küchler 1996) and their evolution, Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997), to improve sequence-based input representation and long-term dependency issue.

In problems such as the NER, which can be treated as sequence labeling problems, the Bi-LSTM architecture (Huang, W. Xu, and Yu 2015; Ma and E. H. Hovy 2016) and character-level representations (Ma and E. H. Hovy 2016; Chiu and Nichols 2016) have made great improvements that have immediately spilled over into the clinical domain (Y. Wu, M. Jiang, Lei, et al. 2015; Y. Wu, J. Xu, et al. 2015; Y. Wu, M. Jiang, J. Xu, et al. 2017; Y. Wu, Xi Yang, et al. 2018) then into the de-identification problem (Dernoncourt, J. Y. Lee, and Szolovits 2017; Z. Liu, Tang, et al. 2017).

Y.-S. Zhao et al. 2018 have experimented with a new method that better incorporates the special context of EHRs, improving the performance of de-identification systems, but at the expense of the time required to perform a rather complex feature engineering. Y. Kim, Heider, and Stéphane M. Meystre 2018 have shown that a stacked learning ensemble is more effective than other ensemble methods, but also this time at the cost of expensive feature engineering that required, for instance, the additional use of a SVM classifier in combination with deep learning techniques.

The latest transformer-based architectures (Vaswani et al. 2017) have proven to be superior to other architectures in various tasks, but not always in NER tasks like BERT (Devlin et al. 2019). Alsentzer et al. 2019 have shown that their Clinical BERT embeddings are superior in the general domain (not in de-identification field), and BioBERT by J. Lee et al. 2020 has also shown higher performance only for non de-identification tasks. Even in the most recent *MEDDOCAN: Medical Document Anonymization Track* in Spanish, Bi-LSTM + CRF-based systems have proven to be competitive with BERT-based systems in NER specific tasks, especially for low-resource languages, such as de-identification in Spanish (Marimon et al. 2019). Moreover Tang, D. Jiang, et al. 2019, by combining the Bi-LSTM + CRF architecture with the BERT neural language model and exploiting its embeddings, have achieved cutting-edge results in English.

In a completely transversal way to the task of clinical de-identification, important results have been obtained from Giorgi and Bader 2020 and Mehrabi et al. 2020: the former have identified strategies to improve the NER in biomedical field by increasing the capacity of generalization of the CRF component, while the latter have provided an interesting reference point for gender assessment in the systems of named entities recognition, observing a lower recognition of female names as "Person" type entities.

In the meantime, there have been several attempts to use attention mechanisms (Vaswani et al. 2017): for instance L. Luo et al. 2018 have used them for the chemical NER at the document level together with the Bi-LSTM + CRF architecture, while Y. Luo, Xiao, and H. Zhao 2020 for generic NER, trying to merge the information of the representations at sentence level with the information at document level. Hu et al. 2020 proposed a fusion attention mechanism to augment reliability of the context information of multi-token entities in document-level NER for news articles, then Gui et al. 2020 a novel approach to manage document-level label consistency in an effective way. But as far as is known so far, only C. Liu et al. 2019 have shown interest in the possibility of document-level analysis for de-identification, proposing a Bi-Capsule-LSTM-CRF architecture that has exhibited better results than Bi-LSTM + CRF on the i2b2 2014 data set but far from the state of the art that also makes use of rule-based methods. Bearing in mind the limitations of both the

document level analysis and the Bi-LSTM + CRF architecture, a new approach at the sub-document level has been proposed and its experimental results confirm performances similar or superior to the state of the art without any feature engineering.

2.2 Language models and embeddings

Embeddings are defined as vector representations of discrete variables such as words, characters or, even sentences. It is possible to obtain ready-to-use pre-trained embedding using large corpora, instead of training them alongside the model on what is frequently a small data set.

Such numerical representations were initially static with respect to context, e.g. assigned a numerical value to a word this would not change as the surrounding words varied, such as for Word2Vec by Mikolov et al. 2013 and GloVe by Pennington, Socher, and Manning 2014. Bojanowski et al. 2017 have tried to change the way embedding works with interesting results: instead of associating embedding to words, FastText embeddings break them into sub-words, i.e. a set of characters that make up n grams, in order to reconstruct the embedding associated with a single word by looking at the various sub-word components identified. Then such numerical representations started to be contextual, through the use of two identical architectural components: the first will work on the sequence while the second will work on the reverse sequence, capturing also the relations among words within paragraphs and developing the so-called Statistical Language Models or, shortly, Language Models (LM) like ELMo (Peters, Neumann, Iyyer, et al. 2018), Flair (Akbik, Blythe, and Vollgraf 2018), BERT (Devlin et al. 2019), GPT (Radford, J. Wu, et al. 2019) and so on.

An approach similar to FastText embeddings has been used by BERT, whose tokenizer is based on WordPieceModel segmenter M. Schuster and Nakajima 2012 which always works on a sub-word level.

Flair Akbik, Blythe, and Vollgraf 2018 have instead descended to the atomic level of text, seeing it not as a sequence of words or sub-words, but as a sequence of characters and adding to this contextual capability. This has resulted in state-of-the-art results in several NLP tasks. Unlike ELMo, Flair embeddings are based on a slightly

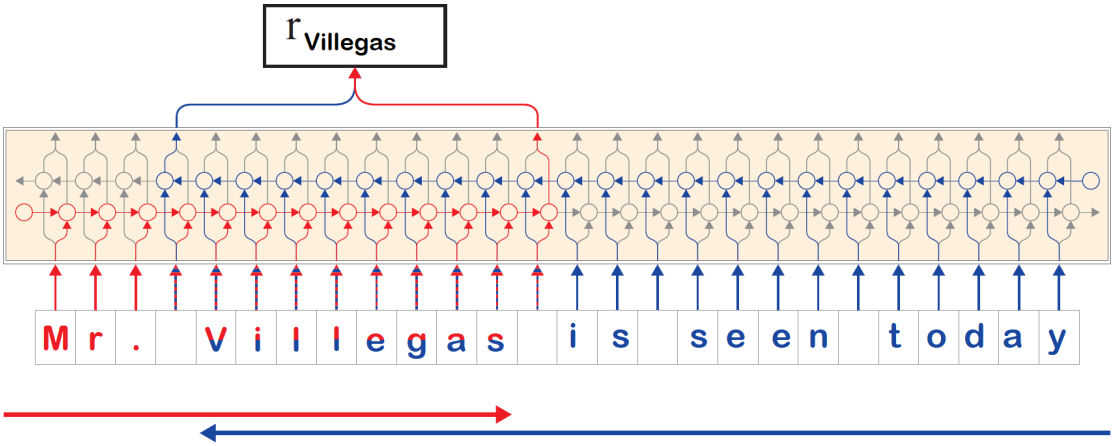


Figure 2.1: Extraction of a contextual string embedding for the word *Villegas*. To form the final embedding, the output hidden states from both forward and backward language models are concatenated. The first (shown in red) will contain information propagated from the beginning of the sentence up to the last character in the word, the second (shown in blue) will contain information propagated from the end of the sentence up to the first character in the word.

different mechanism: its bidirectional language model (biLM) runs on a characters sequence instead of a tokens sequence hence called *character level LM*. In details, looking at its architecture in Figure 2.1, it shows that the extraction of the *Villegas* word embedding is dynamic and goes at the same pace as the surrounding context: it happens through the composition of pre-trained character embeddings from the neighboring text.

Far from being language modeling a solved problem, Peters, Neumann, Zettlemoyer, et al. 2018 showed how the depth of the network conditions the level of information that can be learned, from local information to long-range dependent information, through the representation of contextual words derived from pre-trained biLMs, while Khandelwal et al. 2018 indicated room for further improvement by identifying the limitations of LSTM-based architectures, unable to take into account more than 200 tokens of context on average.

2.3 Clinical de-identification for specific languages

Automatic de-identification and anonymisation systems in languages other than English, although lacking in language resources, have seen greater development in recent years. For example, in Danish, Pantazos, Lauesen, and Lippert 2017 have tried to balance the system in a way that both preserves readability and does not degrade the confidentiality of the large public EHR data set available. Also in Dutch, there have been developments: Scheurwegs et al. 2013 were the first to test machine learning techniques, whereas Trienes et al. 2020 proceeded to compare even the most modern deep learning systems. In both cases it was necessary to request EHRs from Dutch institutes, which are often not publicly available. In French, both Grouin and Névéol 2014 and Gaudet-Blavignac et al. 2018 explored the possibilities of rule-based and CRF-based systems on data sets built by retrieving EHRs from French hospitals. In German, first Tomanek et al. 2012 and then Richter-Pechanski, Riezler, and Dieterich 2018 developed rule-based techniques and machine learning, but they remain proof-of-concepts due to the lack of extensive data training. In Norwegian, a rule-based method was developed by Tveit et al. 2004. Also in Polish language there has been the development of some rules-based system Marciniak, Mykowiecka, and Rychlik 2010; Borowik et al. 2019. Rules-based systems in Portuguese Mamede, Baptista, and Dias 2016 or machine learning in Swedish Alfalahi, Brissman, and Dalianis 2012 were developed. Finally, in Spanish language there was the only other challenge organized besides the English language ones: the most recent MEDDOCAN: Medical Document Anonymization Track Marimon et al. 2019 within IberLEF 2019[1]. As far as we know, there is no research on the subject in Italian to date.

[1] https://sites.google.com/view/iberlef-2019

2.4 Cross-lingual transfer learning approaches

This section examines how low-resource languages have been managed over time. In detail the techniques that preceded BERT are described in Section 2.4.1, while BERT and its multilingual version are illustrated in Section 2.4.2.

2.4.1 Non BERT-based multilingual techniques

In the field of transfer learning, a branch of particular interest applied to the NLP domain is that of cross-lingual transfer learning which, as stated by Pan and Q. Yang 2010 is a type of transductive transfer learning where the source and target domain are different, i.e. training and prediction take place on corpora in different languages, and cross-linguistic transfer occurs through the use of a single cross-linguistic representation space. Initially, task-specific models were popular, based on a coarse-grained representation such as Part of Speech (PoS) tags, and then exploited a delexicalized parser (Zeman and Resnik 2008).

Recently, cross-lingual word embeddings have started to be used in combination with specific neural architectures, obtaining interesting results in various tasks, such as PoS tagging (J.-K. Kim et al. 2017), NER (Xie et al. 2018) and dependency parsing (Ahmad et al. 2019). In addition, several studies have been carried out analyzing the effects of different ways of constructing cross-lingual space: for example, Ruder, Vulic, and Søgaard 2019 analysed methods for learning cross-lingual embedding through both joint training and post-training mapping of monolingual embeddings, whereas Lample, Conneau, et al. 2018 and Artetxe, Labaka, and Agirre 2018 demonstrated that with the alignment of two monolingual word embedding spaces in unsupervised ways it is possible to get better results.

In detail, Lample, Conneau, et al. 2018 introduced Multilingual Unsupervised and Supervised Embeddings[2] (MUSE), created by aligning the embedding spaces of monolingual word embeddings, without using parallel corpora, in an unsupervised way. Then a supervised version of MUSE, crosslingual fastText-based embeddings

[2]https://github.com/facebookresearch/MUSE

(Bojanowski et al. 2017), was also released. These embeddings are generated by aligning the monolingual fastText embeddings in a common space using bilingual dictionaries as ground-truth. Only static embedding vectors have been released and, without the model, it is not possible to generate embeddings for Out-Of-Vocabulary (OOV) words.

Instead, Heinzerling and Strube 2018 proposed Byte-pair Embeddings (BPEmb) to tackle the Out-Of-Vocabulary (OOV) problem. Based on Byte-Pair Encoding (BPE) by Sennrich, Haddow, and Birch 2016, BPEmb create each word embedding by composing the necessary sub-word embeddings. In particular, Heinzerling and Strube 2018 found that BPEmb offer nearly the same accuracy as word embeddings, but at a fraction of the model size, a valuable choice to train small models. BPEmb were released in 275 languages and were successfully used in several cross-lingual scenarios (Bingel and Bjerva 2018; Yimam et al. 2018; M. Zhao and Schütze 2019). Moreover, Zhu, Vulic, and Korhonen 2019 showed the importance of sub-word segmentation, due to the absence of any "one-size-fits-all" configuration, because performance is both task- and language-dependent. In addition, Sahin et al. 2020 noted that sub-word based models perform better than word-based models, like MUSE and Word2Vec (Mikolov et al. 2013), in several low-resources languages scenarios.

Contextual language models, such as Embeddings from Language Models (ELMo) (Peters, Neumann, Iyyer, et al. 2018) and Flair (Akbik, Bergmann, Blythe, et al. 2019), proved to be superior to static models, like Word2Vec by Mikolov et al. 2013 and Global Vectors for Word Representation (GloVe) by Pennington, Socher, and Manning 2014, thanks to the ability to analyze the context, and this further improved performance in cross-lingual scenarios. Flair embeddings (Akbik, Blythe, and Vollgraf 2018; Akbik, Bergmann, and Vollgraf 2019), which constitute a character-based contextual language model on which the Flair NLP framework (Akbik, Bergmann, Blythe, et al. 2019) is based, prompted Johnson et al. 2019 to test a novel methodology for cross-lingual transfer learning for Japanese NER, based on a Bi-LSTM architecture and embeddings at both word and character level as input.

Furthermore, Howard and Ruder 2018 proposed the Universal Language Model Fine-tuning (ULMFiT), an effective transfer learn-

ing method based on an appropriate fine-tuning strategy to improve language models performance, while Radford, Narasimhan, et al. 2018 proposed generative pre-training techniques which led to the Generative Pre-trained Transformer (GPT) language model, which uses an encoder based on transformers (Vaswani et al. 2017). Then Conneau and Lample 2019 extended generative pre-training to cross-lingual models and obtained state of the art results, while T. Schuster et al. 2019 tested cross-lingual alignment with ELMo embeddings overcoming the state of the art for zero-shot dependency parsing.

Additionally, Mulcaire, Kasai, and Smith 2019 experimented a polyglot system based on ELMo, showing relevant results. Indeed, to create a multilingual system, there are two possible alternatives:

1. train a specific model for each language;

2. train only one model for all languages.

In particular, Mulcaire, Kasai, and Smith 2019 have shown how the second choice provides better results especially in the case of low resources languages thanks to the enrichment of the model with the data of languages that, although different, can be linked together on different aspects of the language (e.g. semantics, morphology, syntax, and so on). Starting both from this principle and encouraging results obtained on Slavic languages by Arkhipov et al. 2019, it was decided to consider pre-trained multilingual models on large corpora and fine-tune them on the target language, Italian, which is a low resources language. This way the extremely computationally expensive training procedure can be totally avoided, initializing the model with the multilingual one.

While the world of research has made an effort to organize knowledge in order to better use it against the COVID-19 (Vaishya et al. 2020; Mohamadou, Halidou, and Kapen 2020; Santos et al. 2020; Suri et al. 2020; Coombs 2020; Shakil et al. 2020; Hernandez-Matamoros et al. 2020; Hazarika and Deepak Gupta 2020; Marques, Agarwal, and Torre Díez 2020), on the other hand a series of research with pandemic focus has followed.

For instance, Arora et al. 2020 released, during the COVID-19 global pandemic, a multilingual data set containing more than 5 thousand statements in English, Spanish, French and Spanglish (Spanish + English). This data set was used to study some cross

lingual transfer learning techniques related to the Intent Detection task, observing performance improvement in most models with cross lingual training compared to models with mono lingual training. Based on this assumption, both zero shot and cross lingual training approaches were tested.

Finally, Kırbaş et al. 2020 have used a LSTM model for COVID-19 prediction, detailing evaluation criteria of the models under analysis and providing a prospective estimate of the total number of cases.

2.4.2 BERT-based multilingual techniques

BERT is a deep contextual language model, based on transformers (Vaswani et al. 2017). Unlike ELMo and GPT, BERT is trained by Cloze Task (Taylor 1953), commonly known as masked language modeling, which is different from classic right-to-left or left-to-right language modeling, allowing it to encode information from both directions in each level freely. Furthermore, BERT also optimizes a target for the classification of the next sentence, so that the paired sentences during training are half consecutive pairs and half random pairs. Lastly, BERT uses a sub-word vocabulary based on the WordPieceModel segmenter (M. Schuster and Nakajima 2012), a data-based approach to break down a word into sub-words that is more effective than operating at the word level. As demonstrated by Devlin et al. 2019, BERT is able to achieve high performance in several sentence classification tasks thanks to the fine tuning of the transformer encoder followed by a softmax classification layer fine-tuned for 2-3-4 epochs with a learning rate in the order of e-5: in the case of NER, a sequence of shared softmax classifications produces sequence tagging patterns.

The multilingual BERT (mBERT[3]), differs exclusively for the different training data set consisting of Wikipedia data in 104 languages provided as they are, without the typical links of cross-lingual methods, but appropriately scaled. Leveraging WordPiece, mBERT thus generated is a model in which common sub-words are shared between languages even far apart in the form of a standalone lexicon. Many have recently started investigating the performance of

[3]https://github.com/google-research/bert/blob/master/multilingual.md

mBERT. Among them, Pires, Schlinger, and Garrette 2019 have carried out a series of experiments showing how the transfer also happens in languages in different scripts, although it works better with typologically similar languages. Instead, S. Wu and Dredze 2019 consider a broader spectrum of NLP tasks, comparing mBERT with different methods of zero-shot cross-lingual transfer and experimenting with different strategies to improve generalization capabilities. Moreover, K et al. 2020 studied the contribution of the different components of mBERT to its cross-lingual skills, stressing that the depth of the network is more relevant than the lexical overlap between languages.

Furthermore, Heinzerling and Strube 2019 found that although mBERT performs well in scenarios with medium and high language resources, non-contextual embedding working at the sub-word level, such as BPEmb, outperforms mBERT in low-resources scenarios. Finally, Hvingelby et al. 2020 explored cross-lingual transfer for Danish using several architectures for supervised NER, including Flair, fastText, BPE and both monolingual (Danish) and multilingual BERT, on a modestly-sized training set, testing different training and fine-tuning approaches.

Additionally, Neuraz et al. 2020 used both BERT and Bi-LSTM + CRF architectures to create a drug extraction model to study the ability to respond quickly to emerging diseases such as COVID-19.

Finally, Mohammad et al. 2020 proposed a new *Artificial Intelligence and NLP based Islamic FinTech Model* (based on several NLP techniques, from rules to deep learning) to analyze the impact of the COVID-19 pandemic on the poor and small and medium enterprises, predicting possible future scenarios by leveraging the use of specific taxes of Islamic countries to deal with them.

Chapter 3

Materials and Methods

This chapter describes materials and methods used within the research work. Section 3.1 provides an overview of data sets employed for clinical de-identification, while Section 3.2 shows in detail the system architectures implemented. Moreover, Section 3.3 describes use cases and related experimental setups. Finally, Section 3.4 explains how evaluation metrics work.

3.1 Data sets

This section explains which data sets have been used. In detail, the i2b2/UTHealth 2014 de-identification corpus is illustrated in Section 3.1.1 and the related pre-processing procedure is described in Section 3.1.1, while the SIRM COVID-19 de-identification corpus is illustrated in Section 3.1.2, the related annotation and pre-processing procedures in Sections 3.1.2 and 3.1.2 respectively. Finally, Section 3.1.3 provides more detail on how the data sets were, in some cases, merged or used simultaneously.

3.1.1 The i2b2/UTHealth 2014 de-identification corpus

The i2b2/UTHealth 2014 de-identification corpus was used. It was released by Stubbs, Kotfila, and Uzuner 2015, from the i2b2 National Center for Biomedical Computing for the NLP Shared Tasks Challenges, whose de-identification guidelines reported by Stubbs and Uzuner 2015 conform to the HIPAA Safe Harbor criteria, adding

doctor and hospital name and all ages to the list of identifiers to be removed. Finally the data set was hand-labeled and surrogated before the release. This data set consists of 1304 longitudinal medical records of 296 patients with 2-5 records selected per patient and is officially divided into training and testing set, where the training set contains 790 documents (including 269 for validation) while the testing set contains 514 documents. Each document is a medical record in xml format and the named entities within the documents are annotated as text spans with corresponding entity types. For the purposes of i2b2 annotation project, the 18 categories of PHI identifiers have been expanded to include more specific identifiers, therefore regrouped into 7 main categories and several sub-categories. Table 3.1 presents an exhaustive list of PHI distributions in the i2b2/UTHealth 2014 de-identification corpus. Moreover in Table 3.1, a further detail has been added regarding the division of entities, separating the pure training entities from the validation entities as in the original data set.

Pre-processing of the i2b2/UTHealth 2014 de-identification corpus

At this stage the chosen data set is pre-processed to perform error correction and format conversion.

First, some errors related to entity information within the dedicated spans are detected in the raw i2b2 xml files. In detail, an error corrector is introduced to resolve the misalignment between the initial and final offset values of the entity characters, checking and adjusting them based on the correct position of the entity within the text. In addition, some errors related to the separation of entities from other text have been observed: in particular, there is an absence of the space character before, inside or after the entity in the text with respect to the related information in the dedicated span. For example, the entity *Gambia* was tied to the word *Home*, so *GambiaHome* could not be properly tokenized as it was. To solve this problem, an entity spacer was introduced that allowed to insert a space when needed, parsing the entire document and recalculating the offset values of the initial and final characters of all entities within the document. In addition, any absence of the space character within the entity was conveniently taken into account or not in

Table 3.1: PHI distributions in the i2b2/UTHealth 2014 de-identification corpus

PHI category: subcategory	**TR**	**VD**	**TS**	**Total**
AGE	810	423	764	1997
CONTACT: EMAIL	3	1	1	5
CONTACT: FAX	5	3	2	10
CONTACT: IPADDRESS	0	0	0	0
CONTACT: PHONE	229	80	215	524
CONTACT: URL	2	0	0	2
DATE	5254	2248	4980	12482
ID: ACCOUNT	0	0	0	0
ID: BIO ID	1	0	0	1
ID: DEVICE	7	0	8	15
ID: HEALTH PLAN	1	0	0	1
ID: ID NUMBER	171	90	195	456
ID: LICENSE	0	0	0	0
ID: MEDICAL RECORD	398	213	422	1033
ID: SSN	0	0	0	0
ID: VEHICLE	0	0	0	0
LOCATION: CITY	259	135	260	654
LOCATION: COUNTRY	53	13	117	183
LOCATION: HOSPITAL	928	509	875	2312
LOCATION: ORGANIZATION	85	39	82	206
LOCATION: OTHER	4	0	13	17
LOCATION: STATE	221	93	190	504
LOCATION: STREET	144	72	136	352
LOCATION: ZIP CODE	139	73	140	352
NAME: DOCTOR	1932	953	1912	4797
NAME: PATIENT	879	437	879	2195
NAME: USERNAME	219	45	92	356
PROFESSION	149	85	179	413
Total # of entities	11893	5512	11462	28867

accordance with the information span of the entity.

Secondly, raw i2b2 xml files are converted to brat standoff format[1], then from brat standoff format to delimiter-separated values (DSV) format. As a basis for the conversion scripts, the publicly available NeuroNER tool[2] (Dernoncourt, J. Y. Lee, and Szolovits 2017) was used, exploiting spacy as tokenizer and *en core web lg* as its language model[3]. Some conversion errors were noticed, so several changes to the original python scripts were made before using the pre-processed text as input for the proposed NER system. In converted files, all the labels of the entities are attached to the tokens according to the IOB tagging format (Ramshaw and Marcus 1995) where B-tag represents the beginning of the label, I-tag is attributed to all the following tokens that still belong to the same named entity and finally O represents all the tokens not labeled.

3.1.2 The SIRM COVID-19 de-identification corpus

The Italian SIRM COVID-19 data set, based on a collection of 115 unannotated medical records in pdf format released by SIRM[4], was developed. In order to proceed with the annotations, the guidelines adopted by Stubbs and Uzuner 2015 for the 2014 i2b2/UTHealth de-identification track were followed.

Finally, the SIRM COVID-19 data set was split: 65 medical records were used for training and 50 medical records for testing. To this end, the Table 3.2 presents an exhaustive list of PHI distributions in the SIRM COVID-19 de-identification corpus, with further details on training and testing entities. In the first column C:Subcategory, C: stands for the category to which the entities belong if present, in particular C, I, L, N stand for Contact, ID, Location and Name respectively. In detail, named entities are annotated by using subcategories as labels. Subcategories are then grouped into the appropriate categories as outlined by Stubbs and Uzuner 2015. Some statistical data concerning the SIRM COVID-19 data set have been reported in Table 3.3.

[1] https://brat.nlplab.org/standoff.html

[2] http://neuroner.com/

[3] https://spacy.io/models/en#en_core_web_lg

[4] https://www.sirm.org/category/senza-categoria/covid-19/

Table 3.2: PHI entity distributions in the SIRM COVID-19 de-identification corpus. TR stands for Training data set and TS stands for test data set.

PHI C:Subcategory	**TR**	**TS**	**Total**
AGE	63	55	118
C:PHONE	3	7	10
C:URL	66	76	142
DATE	64	90	154
I:ID NUMBER	137	129	266
L:CITY	38	63	101
L:COUNTRY	1	5	6
L:HOSPITAL	134	132	266
L:ORGANIZATION	4	9	13
L:OTHER	3	6	9
N:DOCTOR	303	430	733
N:PATIENT	3	0	3
PROFESSION	38	27	65
PHI Category	**TR**	**TS**	**Total**
AGE	63	55	118
CONTACT	69	83	152
DATE	64	90	154
ID	137	129	266
LOCATION	180	215	395
NAME	306	430	736
PROFESSION	38	27	65
Total # of entities	857	1029	1886

Table 3.3: Statistical data concerning the SIRM COVID-19 data set.

SIRM COVID-19 data set stats	
Average tokens per document:	262.8
Average NEs per document:	16.4
Average tokens per NE:	2.2

Annotation procedure

The annotation procedure was carried out as described in the following. In detail, data were annotated according to criteria similar to the i2b2/UTHealth 2014 de-identification corpus, so as to maintain uniformity between the recognition categories. Where the appropriate subcategory was not available, it was decided to opt for the closest one semantically: for example, the Italian regions were annotated as LOCATION: OTHER, since they belonged neither to the subcategory LOCATION: COUNTRY nor to the subcategory LOCATION: STATE or the street names identifying the hospitals were aggregated with LOCATION: HOSPITAL entities. In particular, each document was labeled manually and independently by three Italian native speakers, who are researchers in the e-health domain, with the agreement among the annotators calculated by majority. The global agreement for the entire annotation procedure was measured using the Observed Agreement index Goodman and Kruskal 1979 which provides a good approximation in multi-annotator contexts, also offering robustness against imperfect (textual) data Bobicev and Sokolova 2017. In addition to the Observed Agreement index, in order to take into account the level of Inter Annotator Agreement (IAA) in terms of excess over the agreement obtained by chance, the Krippendorff coefficient α Krippendorff 1980 was also calculated. The latter expresses the IAA in terms of disagreement, observed (D_o) and due to chance (D_e): $\alpha = 1 - D_o/D_e$ and, not imposing a minimum number of items, mitigates the statistical effects of low sample data sets such as the one used. The Observed Agreement index value was 0.68, while the Krippendorff coefficient α value was 0.71: according to the grid for the interpretation of coefficients proposed by Landis and Koch 1977 the values obtained indicate a "substantial" agreement.

The disagreement among the annotators is generally motivated by the extreme difficulty, variety and uncertainty of natural language and, therefore, by a very diverse and often subjective linguistic understanding of the meaning of each category. In any case, disagreement is not strictly an indicator of low quality annotation, poor annotator training or insufficient guidelines, especially in semantic tasks Aroyo and Welty 2015, but can be used directly to improve the behavior of automatic systems Chklovski and Mihalcea 2003; Plank,

D. Hovy, and Søgaard 2014.

In Table 3.4 some sample sentences of disagreement among the three annotators have been reported. In particular in the *Sentence* column is reported the sentence under examination with the alternation of red and black colors to indicate different tokens, while in the macro column *Annotator sequence* are reported the annotation sequences for tokens of the first (*#1*), second (*#2*) and third (*#3*) annotator.

Table 3.4: Annotators' disagreement examples.

	Annotator sequence		
Sentence	**#1**	**#2**	**#3**
Uomo di 60 anni *(Man of 60 years)*	O O B-AGE O	O O B-AGE O	O O B-AGE I-AGE
ASL Latina Paziente Maschio *(ASL Latina Male Patient)*	B-HOSPITAL I-HOSPITAL O O	B-HOSPITAL I-HOSPITAL O O	B-HOSPITAL B-CITY O O
COVID-19: caso 98 *(COVID-19: case 98)*	O O O B-IDNUM	O O O B-IDNUM	O O O O
Giunge al PS *(Arrives to the ER)*	O O O	O O B-LOCATION_OTHER	O O B-HOSPITAL
Performance of radiologists *(This is a test sentence)*	O O O	O O B-DOCTOR	O O B-PROFESSION

Pre-processing of the SIRM COVID-19 de-identification corpus

The data set has been annotated manually, generating the annotations in brat standoff format. In addition, several python scripts have been written to convert the data. First the pdf files were transformed into text using the python library *pandas*, then a python script was used to convert the brat standoff format to the CONLL format more suitable for the framework used, using as basis the publicly available NeuroNER tool Dernoncourt, J. Y. Lee, and Szolovits 2017 with spacy as tokenizer and *it_core_news_sm* as language model. To improve tokenizer results, entities have been separated from the rest of the text when wrongly attached, inserting a space before and after when appropriate. As a consequence, the misalignment caused between the initial and final offset values of the characters of the entity has been verified and adjusted within the text.

In this case too, in the converted files, all entity labels are attached to the tokens according to the IOB tagging format (Ramshaw and Marcus 1995) where O represents all untagged tokens, B-tag represents the beginning of the label and finally I-tag is attributed to

all the following tokens that still belong to the same named entity.

3.1.3 Training strategies: mixing data sets

In order to analyze the crosslingual capabilities of the multilingual NER systems under examination, the English i2b2/UTHealth 2014 de-identification corpus and the Italian SIRM COVID-19 de-identification corpus (created ad hoc for the investigation of the performance of low language resource systems like the Italian one) were appropriately mixed.

Four methods were tested:

1. EN. It provides a training set exclusively in language with high resources.

2. IT. It provides a training set exclusively in language with low resources.

3. MIX. It provides a mixed training set, both in high resource language and low resource language.

4. EN-IT. It provides two separate training sets for two distinct training phases: the first with the high resource language data set, the second with the low resource language data set.

Table 3.5 presents an exhaustive list of entity distributions in the de-identification corpora used. In particular, with reference to the first column *C:Subcategory*, *C:* stands for the category to which the entities belong if divided into subcategories, and in detail it can be C, I, L, N which stand for CONTACT, ID, LOCATION and NAME respectively. Instead the TR_{i2b2}, TR_{SIRM} and TS_{SIRM} columns indicate the i2b2 training data set and the SIRM COVID-19 training and testing data sets respectively. Finally, the i2b2 guidelines (Stubbs and Uzuner 2015) provided the subcategories C:IPADDRESS, I:ACCOUNT, I:LICENSE, I:SSN and I:VEHICLE, but the same i2b2 data set has none, so it was preferred not to include them in the Table 3.5.

Finally, to represent at a glance the distribution of the entities within the data sets used for training and testing, a clustered column chart has been constructed, shown in Figure 3.1.

Table 3.5: PHI distributions in the i2b2/UTHealth 2014 training data set and in the SIRM COVID-19 de-identification corpus

C:Subcategory	**TR_{i2b2}**	**TR_{SIRM}**	**TS_{SIRM}**
AGE	810	63	55
C:EMAIL	3	0	0
C:FAX	5	0	0
C:PHONE	229	3	7
C:URL	2	66	76
DATE	5254	64	90
I:BIO ID	1	0	0
I:DEVICE	7	0	0
I:HEALTH PLAN	1	0	0
I:ID NUMBER	171	137	129
I:MEDICALRECORD	398	0	0
L:CITY	259	38	63
L:COUNTRY	53	1	5
L:HOSPITAL	928	134	131
L:ORGANIZATION	85	4	8
L:OTHER	4	3	6
L:STATE	221	0	0
L:STREET	144	0	0
L:ZIP CODE	139	0	0
N:DOCTOR	1932	302	425
N:PATIENT	879	3	0
N:USERNAME	219	0	0
PROFESSION	149	38	27
Category	**TR_{i2b2}**	**TR_{SIRM}**	**TS_{SIRM}**
AGE	810	63	55
CONTACT	239	69	83
DATE	5254	64	90
ID	578	137	129
LOCATION	1833	180	213
NAME	3030	305	425
PROFESSION	149	38	27
Total #	11893	856	1022

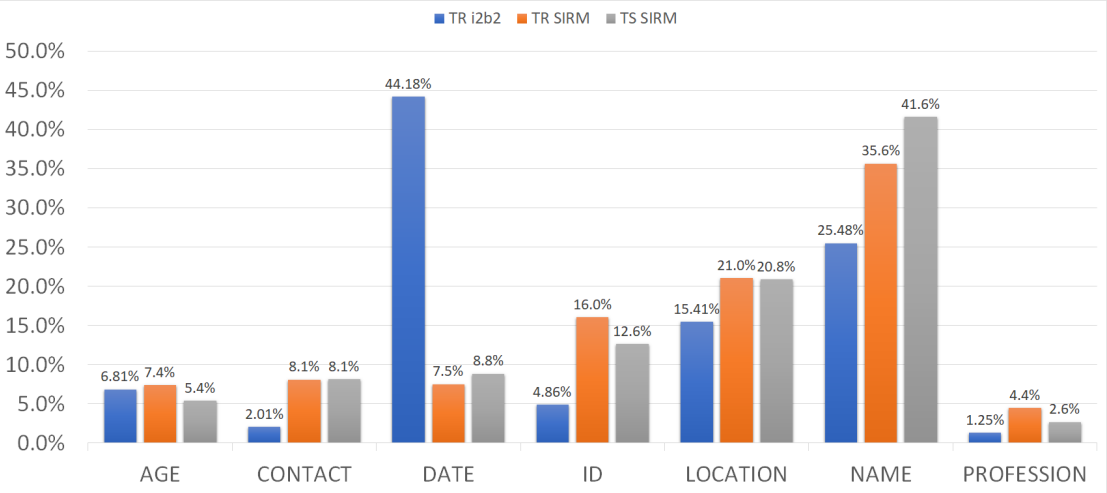


Figure 3.1: Clustered column chart. Distribution of the entities in the data sets.

3.2 System architectures

This section explains which system architectures have been used. In detail, the system architecture based on the Bidirectional Long Short-Term Memory (Bi-LSTM) neural network plus Conditional Random Field (CRF) is described in Section 3.2.1, while the system architecture based on the Bidirectional Encoder Representations from Transformers (BERT) neural network is described in Section 3.2.2.

These system architectures are currently considered the state of the art for NER tasks in NLP: the results obtained in the literature do not allow to identify a significantly superior architecture in the case of clinical de-identification but, depending on the specific scenario (conditioned by language, size of data sets, training strategies and so on), one architecture tends to prevail over the other.

A different discussion deserves the time complexity. Given a sentence of length N (number of tokens composing the sentence), systems based on transformers like BERT process all tokens together, so the time complexity is O(1) while for a Bi-LSTM + CRF it is O(N) (Strubell et al. 2017; Li et al. 2020): this is mainly due to the fact that transformers were designed to run on parallel hardware architectures such as GPU, TPU and so on, resulting faster (Vaswani et al. 2017; P. J. Liu et al. 2018; Kitaev and Klein 2018; Li et al. 2020) whereas the second is intrinsically serial.

3.2.1 Bi-LSTM + CRF based architecture

One of the best performing sequence labeling architecture recognized by scientific literature is represented by the Bi-LSTM + CRF model, as demonstrated by Huang, W. Xu, and Yu 2015, who tested several architectures such as LSTM, Bi-LSTM, CRF, LSTM+CRF and Bi-LSTM + CRF for sequence labeling task. In detail, a network constituted by a Bi-LSTM layer and by a CRF layer is able to use more efficiently two types of information: the input features thanks to the Bi-LSTM network and the relations between the tags at sentence level thanks to the CRF network. Both networks work in a similar way, exploiting the bidirectionality given by the use of information, i.e. input features and relations between tags, coming both from the past (the preceding words) and from the future (the following words), hence the ability to handle long-term dependencies. As it has been widely demonstrated starting from the experiments of Huang, W. Xu, and Yu 2015 and confirmed by the consequent scientific literature, employing these networks at the same time for sequence labeling tasks shows superior and more robust performance, as well as higher accuracy, than employing only (Bi-)LSTM or CRF. Figure 3.2 introduces the general architecture of the proposed Bi-LSTM + CRF system. In particular, according to what was said in section 3.1.3 concerning the possible strategies of fine-tuning and knowledge transfer, the example input types EN, IT, MIX and EN-IT are shown below the figure.

Figure 3.3 shows an overview of the processes required for NER-based clinical de-identification. First of all, a collection of clinical records documents is needed, which have to be re-organised in the form of an annotated dataset. Secondly, there is a preliminary step to prepare the input data, i.e. tokenization: it is performed in order to split each input sentence $s = w_1 w_2 ... w_n$ (where w_t, with $1 \leq t \leq n$, represents the generic token) within raw clinical notes into a sequence of tokens. At this point the real neural network comes into play, whose first input layer is called the embedding layer: here the tokens (which can be words, sub-words or characters, depending on the type of embedding chosen) are transformed into numeric vectors that, in the more sophisticated versions, try to incorporate different aspects of the tokens (grammatical, morphological, syntactic, semantic and so on) from a general point of view. Then, the Bi-LSTM

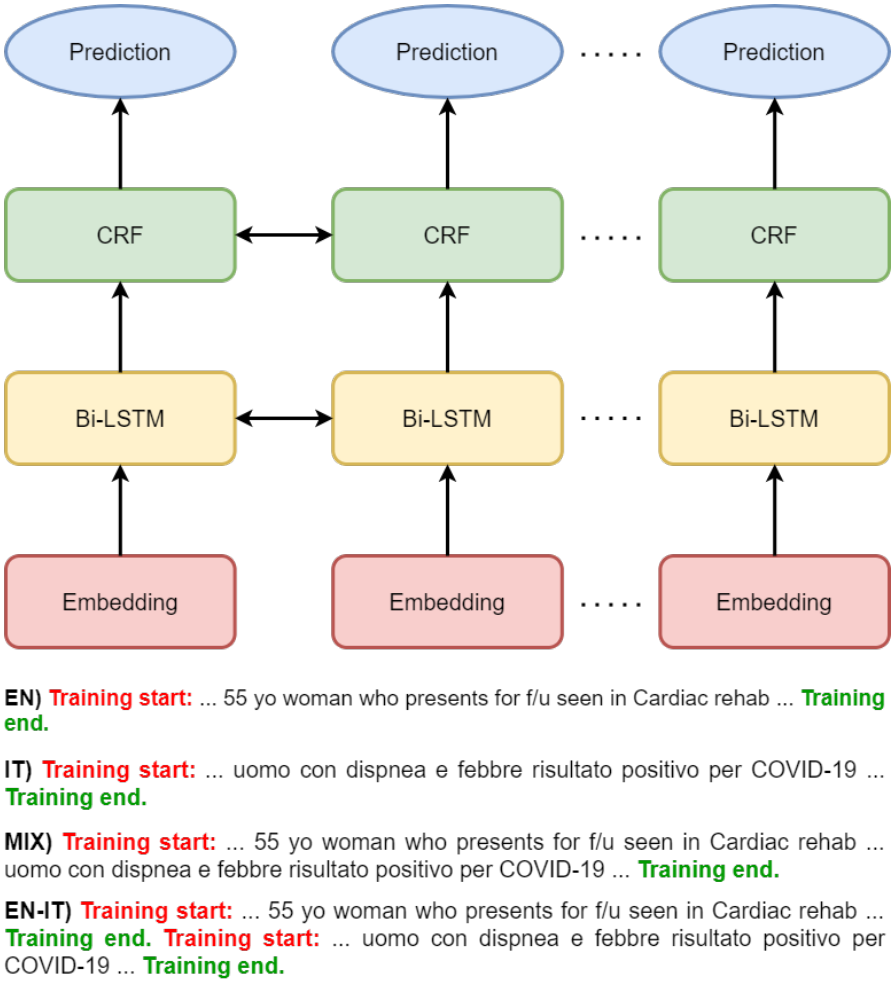


Figure 3.2: Bi-LSTM + CRF overall system architecture. EN, IT, MIX and EN-IT types represent the possible inputs according to training strategies.

specialises these representations by exploiting the ability to analyze the bidirectional context (taking into account its memory limits), i.e. the correlations between the tokens according to the texts in which they are placed. Finally, the CRF layer, working in an analogous way to the Bi-LSTM layer, takes care of providing the predictions of the output labels, trying to preserve their coherence: for instance, it is unlikely that the sequel of a token of type ”beginning of person name” is a token of type ”continuous of thing name”. Finally, a high-level algorithmic representation of the proposed NER-based clinical de-identification management method is illustrated with Algorithm 1.

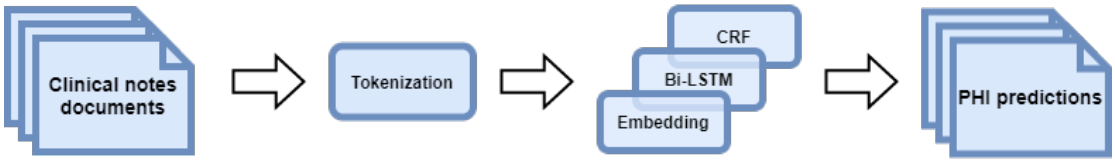


Figure 3.3: NER-based clinical de-identification processes overview.

Algorithm 1: Algorithmic representation of the proposed method.

```
Result: Most likely PHI label for each token
Given pre-processed EHRs;
while EHRs not finished do
    Tokenize documents;
    if Tokenization is successful then
        Embedding tokens;
        Specializing embeddings through context analysis
         given by Bi-LSTM;
        Providing probabilities of each PHI label for each
         token given by CRF;
    else
        Change EHRs pre-processing steps;
    end
end
```

Embedding layer

The embedding layer was chosen on a case-by-case basis according to the requirements to be met. In detail, different types of embeddings have been selected and mixed, so an exhaustive overview of the different solutions adopted for this layer is given below. It is important to remember that, as shown by Alsentzer et al. 2019, the use of specific embeddings for the clinical de-identification task, i.e. clinical or biomedical versions, does not provide improvements, so versions of embeddings trained on generic domains have been used.

FastText Built and pre-trained over very large corpora by Bojanowski et al. 2017, these embeddings are static related to context and work on the subword-level. In this way FastText embeddings attempt to capture morphological information to induce word em-

beddings and deal better with out of vocabulary words.

MultiBPEmb It[5] is the multilingual version of BPEmb[6] (Heinzerling and Strube 2018). The basis of these embeddings is one large multilingual segmentation model. Consequently, corresponding embeddings with a sub-word vocabulary, i.e. pre-trained sub-word embeddings, are shared among all 275 supported languages. The training corpus is based on Wikipedia: thanks to the underlying algorithms, which are language-agnostic but not language-independent, the article texts of all Wikipedia editions can be concatenated. This way, a sub-word segmentation model and sub-word embeddings are learnt. In detail, SentencePiece[7] (Kudo and Richardson 2018), the open source version of Google WordPiece, is used to learn the BPE sub-words segmentation model, while GloVe (Pennington, Socher, and Manning 2014) is used to train sub-word embeddings. In particular, the dimensionality of the sub-word embeddings is set at 300, while the vocabulary size can be 100000, 320000, 1000000. Generally, embedding a word though BPE means that the word is subdivided into sub-words, whose embeddings vectors are subsequently combined. In sequence tagging problems with word-based gold annotations, these sub-word embeddings vectors are usually condensed into one, and this procedure can be done in several ways (e.g. arbitrarily choosing one then losing some information, using a composition function such as addition, leveraging a RNN, and so on). In this case, in order to condense the sub-word embeddings into one, the first and last sub-words embedding vector have been concatenated, leaving GloVe as the embedding algorithm. The vocabulary size has been chosen equal to 1000000, so that words can be more easily represented through sub-words.

Flair Recently, Akbik, Blythe, and Vollgraf 2018 proposed their embeddings, called Flair and described as *contextual string embeddings*, along with their Flair NLP framework (Akbik, Bergmann, Blythe, et al. 2019). The novelty of these embeddings is the ability to capture latent syntactic-semantic information, unseen by standard word embeddings, leveraging two important principles: firstly,

[5]https://nlp.h-its.org/bpemb/multi/
[6]https://nlp.h-its.org/bpemb/
[7]https://github.com/google/sentencepiece

they model words as sequences of characters because they are trained without any explicit notion of words and, secondly, the surrounding text contextualizes them so that the same word will have different embeddings derived from its contextual use, generating appropriate embeddings for polysemic words. Such embeddings are usually employed taking advantage of both forward and backward version, pre-trained on large unlabeled corpora.

Multilingual versions are also available. It is possible to distinguish between the *multi* version, pre-trained on more than 300 languages using the JW300 corpus as proposed by (Agic and Vulic 2019), and the *multi-fast* version pre-trained on English, German, French, Italian, Dutch and Polish, mixing several corpora (Web, Wikipedia, Subtitles and News). The embeddings dimensionality is set at 1024 and 2048 for *multi-fast* and *multi* respectively, one for forward and one for backward embeddings. The interesting property of these character-level embeddings is related to their vocabulary size: it is not as computationally heavy as word-level embeddings that have millions of distinct words to consider, but it only counts a bunch of hundreds of distinct characters, so it is really easy to train. Finally, character-level language models deal well with Out-Of-Vocabulary (OOV), rare and misspelled words and sub-words like prefixes and suffixes, hence with morphologically rich languages like Italian.

Stacked Each embedding can detect different features within the text, so their combination through concatenation, ensemble or weighting (Y. Kim, Heider, and Stéphane M. Meystre 2018; Akbik, Blythe, and Vollgraf 2018), can be useful and further improve performance as demonstrated by several studies also in the field of clinical NER (Y. Wu, Xi Yang, et al. 2018; Si et al. 2019; M. Jiang, Sanger, and X. Liu 2019; Kalyan and Sangeetha 2020).

GloVe and Flair As suggested by Akbik, Blythe, and Vollgraf 2018, stacking Flair with GloVe (Pennington, Socher, and Manning 2014) through concatenation potentially adds more capacities.

As a result, each input token w_t is transformed into its numerical representation x_t, in the form of a stacked embedding with the

following shape:

$$\mathbf{x}_t = \begin{bmatrix} \mathbf{x}_t^{GloVe} \\ \mathbf{x}_t^{Flair} \end{bmatrix} \tag{3.1}$$

where $\mathbf{x}_t^{GloVe}$ is a classic word-level embedding while $\mathbf{x}_t^{Flair}$ adds its own capabilities.

The concatenated embedding is generally followed by a *reprojection layer*: a map on top of the pure concatenated embedding layer. This is a fully connected linear layer on each concatenated embedding: in essence it is a layer that remaps the concatenated embedding to another embedding space, making it possible to achieve some of the effect of fine-tuning modifying the embeddings representation before moving on to successive layers.

MultiBPEmb and Flair Another concatenation was made using MultiBPEmb and Flair for a multilingual scenario, getting the stacked embedding x_t of each word as:

$$\mathbf{x}(t) = \begin{bmatrix} \mathbf{x}_t^{MultiBPEmb} \\ \mathbf{x}_t^{Flair} \end{bmatrix} \tag{3.2}$$

where $\mathbf{x}_t^{MultiBPEmb}$ and $\mathbf{x}_t^{Flair}$ are respectively the MultiBPEmb word embedding and a type of Flair contextual string embedding.

As said, Akbik, Blythe, and Vollgraf 2018 have demonstrated how the combination of Flair embeddings with GloVe embeddings is the one capable of achieving the best performance for NER. But the use of a stacked embedding in multilingual environments is able to achieve better performance when the pre-training languages have similar characteristics, otherwise the risk is to increase the confusion introduced in the network then degrade its performance. For this reason, in a bilingual scenario like the one under consideration, the optimal choice would have been to use English-Italian bilingual embeddings but, unfortunately, both Flair embeddings and GloVe embeddings are not available in such combinations. Therefore it was decided to use *Flair embeddings multi fast* (with far fewer languages than *Flair embeddings multi*) together with *MultiBPEmb*, which continue to use the GloVe algorithm but adding the ability to work at sub-word level, as already explained.

LSTM layers

A LSTM layer works through its hidden layers h. It takes in input a sequence of embeddings [$x(1)$, $x(2)$, ..., $x(n)$] where each $x(i)$ is the numerical representation sequence of the i-th token. The representation of the generic i-th token at step t, i.e. $x_t(i)$, will be referred to as x_t for simplicity of notation. In detail $x_t = d_1 d_2 ... d_m$, where d_i is the i-th digit and m is the dimension of the embedding vector x_t. The LSTM layers output a new token representation sequence $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$ with $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ representing the concatenated outputs of forward and backward LSTM respectively, at step t. In detail, the so-called forward LSTM produces a representation of the left context of the sentence for each token, while another LSTM that reads the same sequence in reverse, the so-called backward LSTM, obtains a representation of the right context of the sentence for the same token. Hence, the overall output h_t is the concatenation of both left and right context representations. This couple of LSTMs is referred to as a bidirectional LSTM, whose superiority over unidirectional architectures for sequence tagging tasks such as NER has been widely demonstrated in literature Graves and Schmidhuber 2005 thanks to its ability to efficiently make use of both left (via forward LSTM) and right context (via backward LSTM) representations. Therefore, the representation of a token, e.g. a word, obtained using this model is an effective representation of a token in context.

In particular, the LSTM unit is composed by three gates (an i input gate, a f forget gate and an o output gate) and a c memory cell implemented to work at step t as follows:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t\text{-}1} + \mathbf{W}_{ci}\mathbf{c}_{t\text{-}1} + \mathbf{b}_i) \tag{3.3}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t\text{-}1} + \mathbf{W}_{cf}\mathbf{c}_{t\text{-}1} + \mathbf{b}_f) \tag{3.4}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t\text{-}1} + \mathbf{i}_t \odot tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t\text{-}1} + \mathbf{b}_c) \tag{3.5}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t\text{-}1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o) \tag{3.6}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot tanh(\mathbf{c}_t) \tag{3.7}$$

where σ is the element-wise appropriate sigmoid activation function (i.e. logistic or softmax) and $\odot$ is the element-wise product; $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{c}_t$ and $\mathbf{o}_t$ are the input gate, forget gate, cell and output gate vectors, which all have the same dimensions as the hidden vector $\mathbf{h}_t$; $\mathbf{W}\_i$,

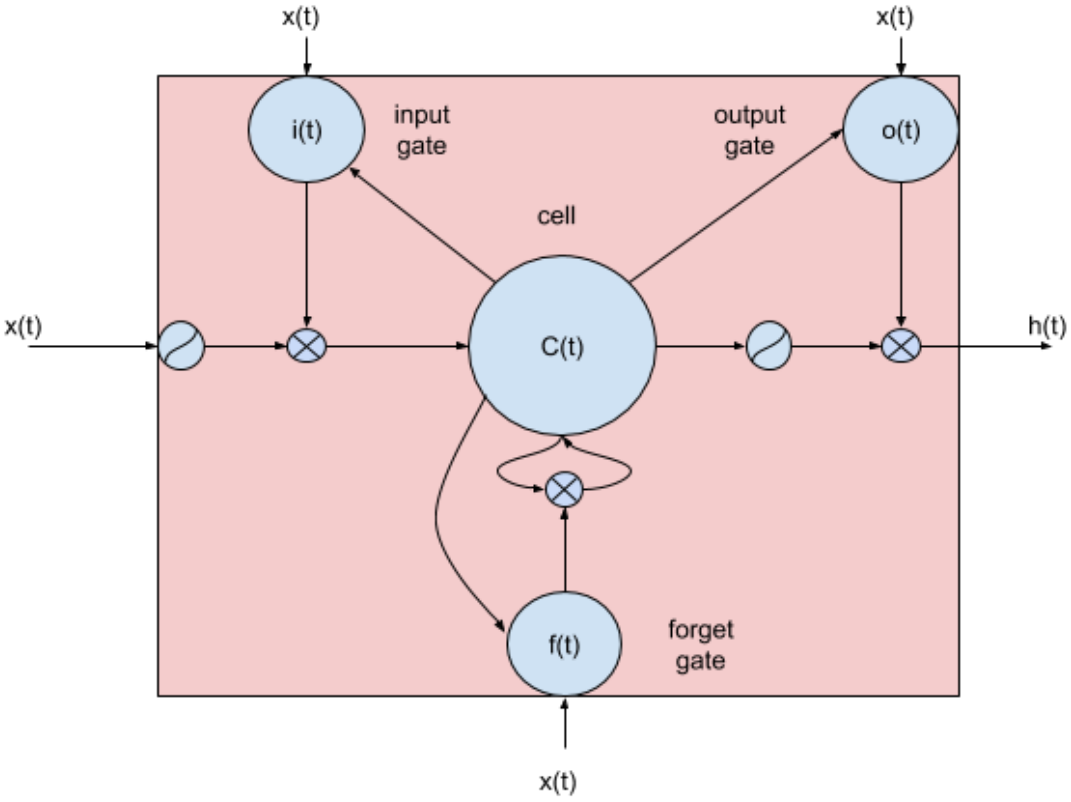


Figure 3.4: Long Short-Term Memory cell representation.

$\mathbf{W}\_f$, $\mathbf{W}\_c$, $\mathbf{W}\_o$ (with subscripts x, h or c in place of _) are the weight matrices for input $\mathbf{x}_t$, hidden state $\mathbf{h}_t$ and memory cell $\mathbf{c}_t$ respectively, to be calculated during the training process: for example, the notation $\mathbf{W_{xo}}$ represents the weight matrix of the input-output gate. Lastly $\mathbf{b}_i$, $\mathbf{b}_f$, $\mathbf{b}_c$ and $\mathbf{b}_o$ represent the bias vectors. Figure 3.4 represents a Long Short-Term Memory cell as described in Huang, W. Xu, and Yu 2015.

CRF layer

It was demonstrated that, using the CRF (Lafferty, McCallum, and Pereira 2001) layer at the top of the Bi-LSTM, the overall prediction performance of the sequence tagger classifier can be improved (Huang, W. Xu, and Yu 2015).

A CRF layer associates to the $h = h_1 h_2 ... h_n$ input sequence, that is a generic token representation sequence for sentence s outputted by the Bi-LSTM layer, the output $y = y_1 y_2 ... y_n$ that represents the most plausible label sequence for the sentence s, e.g. I-tag should

not follow O. Said TR the generic training set and said θ the generic CRF layer, the parameters of the latter are estimated maximizing the following log-likelihood function:

$$L(\theta) = \sum_{(s,y)\in TR} logp(y|h,\theta) \tag{3.8}$$

Said $Z_\theta(h,y)$ the score of label sequence y for the sentence h, the conditional probability p can be calculated as:

$$p(y|h,\theta) = \frac{e^{Z_\theta(h,y)}}{\sum_{y'} e^{Z_\theta(h,y')}} \tag{3.9}$$

where $y^{'}$ is a possible sequence of labels of h. Therefore the log-likelihood of p is equal to:

$$logp(y|h,\theta) = Z_\theta(h,y) - log\sum_{y'} e^{Z_\theta(h,y^{'})} \tag{3.10}$$

To take into account the dependencies among neighboring labels, a transition matrix T is added to an emission matrix E to provide $Z_\theta(h,y)$ in this way:

$$Z_\theta(h,y) = \sum_{t=1}^{n}(E_{y_t,t} + T_{y_{t-1},y_t}) \tag{3.11}$$

where $E_{y_t,t}$ is the probability of the token h_t versus the label y_t, and T_{y_{t-1},y_t} is the probability of the token h_{t-1} with the label y_{t-1} followed by h_t with the label y_t.

It is possible to maximize the log-likelihood of the equation 3.8 on the whole TR training set by taking advantage of dynamic programming (Rabiner 1989), and find the best sequence of labels for each sentence s by maximizing the score given by the equation 3.11 using the Viterbi algorithm (Forney 1973), therefore exploiting an *argmax* function, during the test.

3.2.2 BERT based architecture

The Bidirectional Encoder Representations from Transformers (Devlin et al. 2019), is a general purpose language model trained on

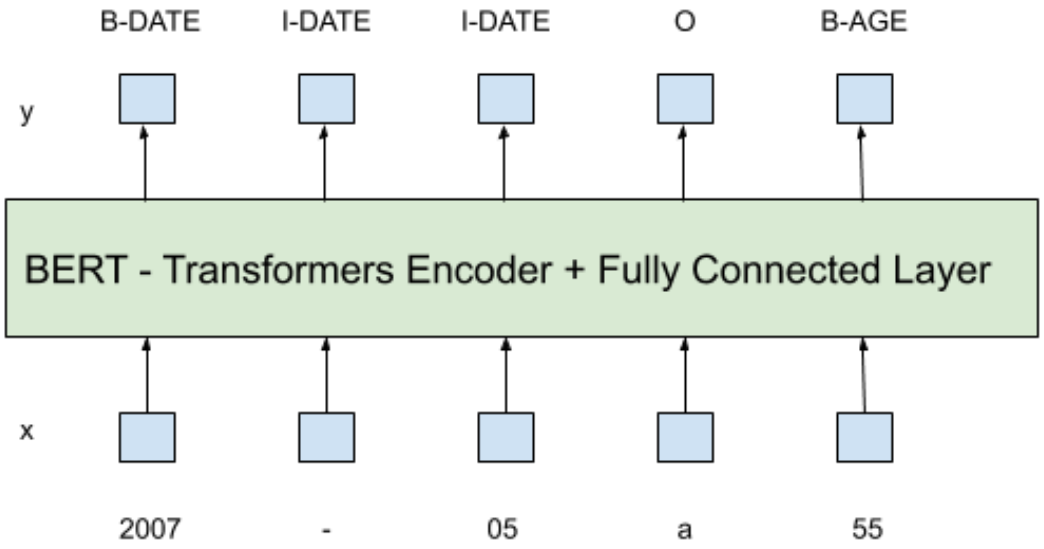


Figure 3.5: Simple BERT network topology for Entity Recognition Task

a large text corpus (like Wikipedia), which can be used for various downstream NLP tasks, such as NER, Relation Extraction, and Question Answering, without heavy task-specific feature engineering. In detail, BERT architecture is based on 12 encoder layers, called Transformers Blocks, 12 attention heads or Self-Attention (Vaswani et al. 2017), and feed forward networks with a hidden size of 768. A simple network topology is shown in Figure 3.5.

BERT accepts embedding and encoder input/output vectors that have a dimension of 512, called Maximum Sequence Length. Some special tokens are employed: the first is *[SEP]*, used for segments separation. The second one corresponds to the first input token supplied, the *[CLS]* token (*CLS* stands for *Classification*), which produces an output vector, of *hidden size* dimension, that can be used as the input for an arbitrarily chosen classifier. In particular, for NER tasks, BERT is fine-tuned following a general tagging task approach without a CRF layer as output layer. As input to the token-level classifier, working over the NER label set, the representation of the first sub-token is used.

Formally, the final hidden representation h_i of each token i is passed into softmax function. The probability P is calculated as follows:

$$P(t|h_i) = softmax(W_o H_i + b_o) \tag{3.12}$$

where $t \in T$, W_o and b_o are weight parameters. Furthermore, during the training, categorical cross-entropy as loss function is used.

Transformer

At the base of BERT is the Transformer (Vaswani et al. 2017). Say $\mathbf{x}$ and $\mathbf{y}$ a sequence of subwords from a couple of sentences. The token [CLS] is placed before $\mathbf{x}$ and after both $\mathbf{x}$ and $\mathbf{y}$ the token [SEP]. Called E the embedding function and called LN the normalization layer (Vaswani et al. 2017), it is possible to get the embedding in this way:

$$\hat{h}_i^0 = E(x_i) + E(i) + E(1_{\mathbf{x}}) \tag{3.13}$$

$$\hat{h}^0{}_{\mathrm{j}+|x|} = E(y_j) + E(j + |x|) + E(1_{\mathbf{y}}) \tag{3.14}$$

$$\hat{h}_\cdot^0 = Dropout(LN(\hat{h}_\cdot^0)) \tag{3.15}$$

Hence the embeddings follow M transformer blocks. Defined the element-wise Gaussian Error Linear Units (GELU) activation function (Hendrycks and Gimpel 2016) and called MHSA the Multi-Heads Self-Attention function and FF the Feed Forward layer, in each of these blocks it applies:

$$\hat{h}_\cdot^{\mathrm{i}+1} = Skip(FF, Skip(MHSA, h_\cdot^i)) \tag{3.16}$$

$$Skip(f, h) = LN(h + Dropout(f(h))) \tag{3.17}$$

$$FF(h) = GELU(h\mathbf{W}_1{}^\top + \mathbf{b}_1)\mathbf{W_2}{}^\top + \mathbf{b}_2 \tag{3.18}$$

where $h^i \in \mathbb{R}^{(|\mathbf{x}|+|\mathbf{y}|)\times \mathrm{d_h}}$, $\mathbf{W}_1 \in \mathbb{R}^{4\mathrm{d_h}\times \mathrm{d_h}}$, $\mathbf{b}_1 \in \mathbb{R}^{4\mathrm{d_h}}$, $\mathbf{W}_2 \in \mathbb{R}^{4\mathrm{d_h}\times \mathrm{d_h}}$, $\mathbf{b}_2 \in \mathbb{R}^{4\mathrm{d_h}}$ and one new position $\hat{h}_i$ is calculated as follows:

$$\begin{aligned}[...,\hat{h}_i,...] = MHSA([h_1,...,h_{|\mathrm{x}|+|\mathrm{y}|}]) = \\ \mathbf{W}_o Concat(h_i^1,...,h_i^N) + \mathbf{b}_o\end{aligned} \tag{3.19}$$

While in each attention, also called attention head, it applies:

$$h_i^j = \sum_{k=1}^{|\mathbf{x}|+|\mathbf{y}|} Dropout(\alpha_k^{(i,j)})\mathbf{W}_V^j h_k \tag{3.20}$$

$$a_k^{(i,j)} = \frac{exp\frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_k}{\sqrt{d_h/N}}}{\sum_{k'=1}^{|\mathbf{x}|+|\mathbf{y}|} exp\frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_{k'}}{\sqrt{d_h/N}}} \tag{3.21}$$

where N is the number of attention heads, $h_i^j \in \mathbb{R}^{(d_h/N)}$, $\mathbf{W}_o \in \mathbb{R}^{\mathrm{d_h} \times \mathrm{d_h}}$, $\mathbf{b}_o \in \mathbb{R}^{\mathrm{d_h}}$ and $\mathbf{W}_Q^j, \mathbf{W}_K^j, \mathbf{W}_V^j \in \mathbb{R}^{\mathrm{d_h/N} \times \mathrm{d_h}}$.

To date BERT is released in two sizes $\mathrm{BERT_{base}}$ and $\mathrm{BERT_{large}}$. $\mathrm{BERT_{base}}$ is made of 12 layers (transformer blocks), 768 hidden size, 12 attention heads, and 110 million parameters, whereas $\mathrm{BERT_{large}}$ is composed of 24 layers, 1024 hidden size, 16 attention heads and, 340 million parameters. The multilingual version of BERT is released only in the base size and is case sensitive. However, additionally, experiments by the scientific community have widely demonstrated that Cased versions of BERT and variants are superior to Uncased versions in the NER task, where the relevant entities often have capitalized initials (Mayhew, Tsygankova, and Roth 2019). For this reason, the cased version of mBERT has been used in this work.

3.3 Use cases and experimental setups

In this section, the use cases and their experimental setups are presented. In particular, Section 3.3.1, 3.3.2 and 3.3.3 describe the three use cases examined.

3.3.1 First use case

The first use case is related to the employment of a stacked embedding consisting of Flair and GloVe in combination with an appropriate exploitation of the full memory capacity of the Bi-LSTM + CRF architecture, using the English i2b2/UTHealth 2014 de-identification corpus. Figure 3.6 shows the architecture overview of the proposed NER-based clinical de-identification system.

This implementation used Flair[8] (Akbik, Bergmann, Blythe, et al. 2019), a straightforward framework for NLP tasks, such as NER, part-of-speech (PoS) tagging, sense disambiguation and classification. For this study, stochastic gradient descent (SGD) algorithm was used to estimate neural networks parameters, using the hyperparameters shown in Table 3.6. Each training was started using a learning rate equal to 0.1: after 3 (*patience* parameter) successive epochs without any loss reduction, the learning rate is multiplied by the annealing factor. When learning rate becomes smaller

[8]https://alanakbik.github.io/flair.html

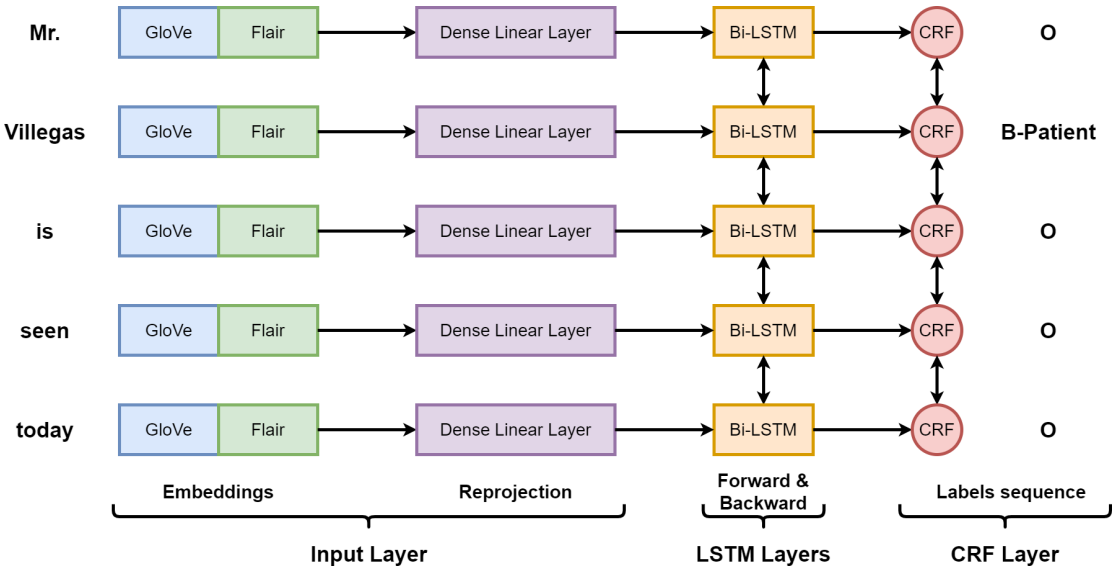


Figure 3.6: First use case: architecture overview.

than 0.0001 the system reaches an early stopping condition. So the number of training epochs equal to 500 constitutes an upper-bound. Therefore, all neural network models have the same hyper-parameters shown in Table 3.6 but the training time changes depending on the early stopping condition. Gradient clipping value was 5.0, Bi-LSTM hidden size 256, word dropout 0.05 and variational dropout 0.5. Lastly, batch size was set to 4. In the system used the dimensionality of a reprojected embedding is equal to 4196 that is the sum of the dimensionalities of the stacked embeddings, i.e. 4096 for Flair Forward and Backward embeddings plus 100 for GloVe embeddings. Furthermore, the versions of embeddings used are all trained on generic domain: as demonstrated by Alsentzer et al. 2019, using domain-specific embeddings for the de-identification task does not lead to an improvement in performance. All experiments were performed on an IBM POWER9 cluster with NVIDIA V100 GPUs.

An additional test criterion was adopted in the wake of considerations made by Khandelwal et al. 2018, who found that LSTM-based systems have an effective context size of about 200 tokens on average. First, a *sentences grouping factor* (SGF) was defined: it indicates the number of sentences grouped in the clinical training, validation and testing notes before producing the training, validation and testing files for the proposed system and conducting what is hereafter called a *sub-document level analysis*. Then SGFs were chosen equal to 1, 2, 4, 8, 16 and 32 taking into account the average extent of

Table 3.6: Hyper-parameters

Hyperparameter	**Value**
Annealing factor	0.5
Batch size	4
Dropout	0.05 (word) - 0.5 (variational)
Epochs	up to 500
Gradient clipping value	5
Hidden size	256
Learning rate	from 0.1 up to 0.0001
Patience	3
RNN Layers	1

the sentences in terms of number of tokens (see Table 3.7) found for each SGF in the training, validation and test files thus generated.

Finally all models were trained, evaluated and tested using the official division of the data set, repeating the procedure five times for each configuration and reporting the arithmetic mean of the results, rounded to the fourth decimal place.

Table 3.7: Average of tokens per sentence

SGF	**Training data set**	**Validation data set**	**Test data set**
1	7.3240	7.3858	7.5039
2	14.6150	14.7432	14.9676
4	29.1031	29.4298	29.8001
8	57.8398	58.6173	59.1131
16	113.5571	115.9395	116.3578
32	219.2904	228.6881	225.3407

3.3.2 Second use case

The second use case is related to the experimentation of a stacked embedding consisting of Flair and FastText, never tested before for the clinical de-identification scenario in Italian, on a novel ad-hoc created data set, the Italian SIRM COVID-19 de-identification corpus.

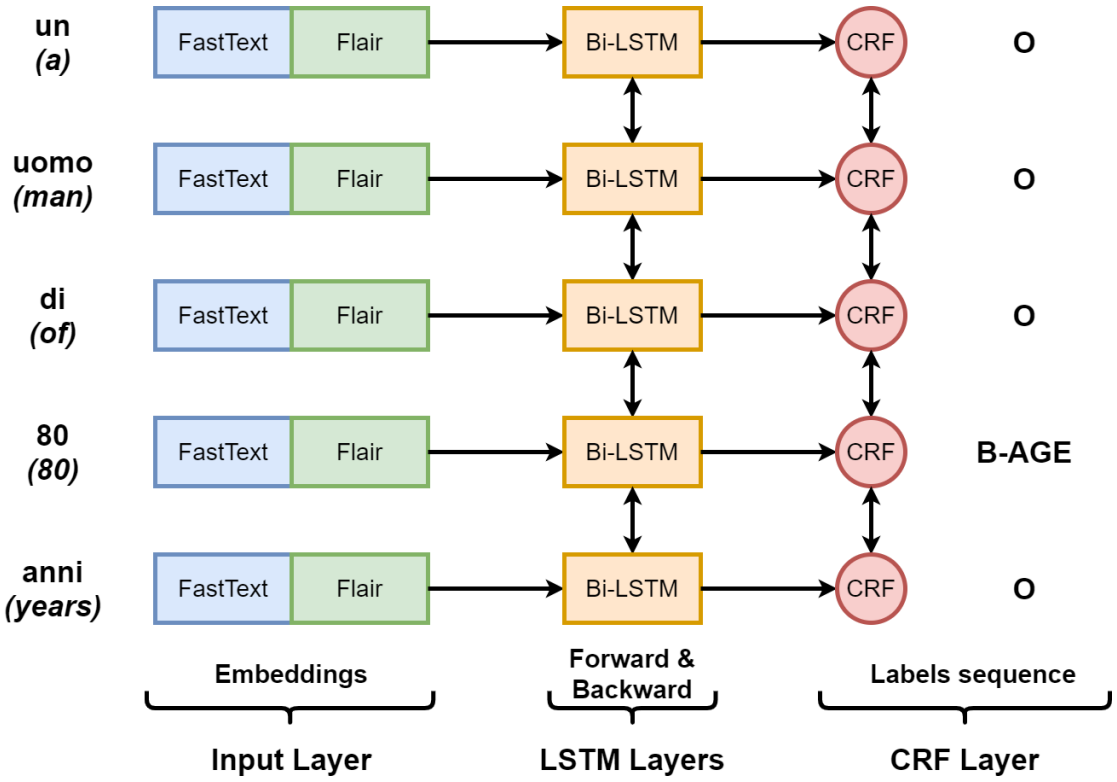


Figure 3.7: Second use case: architecture overview.

The architecture overview of the proposed clinical de-identification system is shown in Figure 3.7.

These experiments used Flair framework[9] (Akbik, Bergmann, Blythe, et al. 2019) for Bi-LSTM + CRF model implementation. It provides state-of-the-art general-purpose architectures with thousands of pre-trained models in over a hundred languages for NLP tasks, such as NER, part-of-speech (PoS) tagging, sense disambiguation and classification.

Flair framework was used with the hyper-parameters reported in Table 3.8 and stochastic gradient descent (SGD) algorithm was used to estimate neural networks parameters. On the one hand we used only Italian FastText embeddings or only Flair (forward and backward) embeddings, on the other hand we stacked Italian FastText and Italian Flair (forward and backward) embeddings concatenating them.

As far as we know, there are no other works for the particular NER task of clinical de-identification in Italian, since there are no publicly available Italian data sets. Hence, beside the Bi-LSTM

[9]https://alanakbik.github.io/flair.html

Table 3.8: LSTM-based model hyper-parameters.

Hyperparameter	**Value**
Annealing factor	0.5
Batch size	16
Dropout (Variational)	0.5
Dropout (Word)	0.05
Epochs	up to 500
Gradient clipping	5
Hidden size	256
Learning rate	from 0.1 up to 0.0001
Patience (early stopping parameter)	3
RNN Layers	1

+ CRF model, the BERT model was tested too, which is another common state-of-the-art language model for different NLP tasks. In detail, the Hugging Face Transformers[10] framework for BERT-based models was used, the main architecture is shown in Figure 3.8.

In particular, the BERT architecture Devlin et al. 2019, which stands for Bidirectional Encoder Representations from Transformers, is a general purpose language model trained on a large text corpus (like Wikipedia), which can be used for various downstream NLP tasks, such as NER, Relation Extraction, and Question Answering, without heavy task-specific engineering. $\text{BERT}_{\text{BASE}}$ architecture is based on 12 encoder layers, known as Transformers Blocks, 12 attention heads (or Self-Attention as introduced in Vaswani et al. 2017), and feed forward networks with a hidden size of 768. Instead, $\text{BERT}_{\text{LARGE}}$ is based on 24 encoder layers, 16 attention heads and feed forward networks with a hidden size of 1024. For simplicity, if not specified, we will refer to $\text{BERT}_{\text{BASE}}$ in the following.

$\text{BERT}_{\text{BASE}}$ Maximum Sequence Length fixes the accepted embedding and encoder input/output vectors dimension to 512. Two special tokens are used: *[CLS]* and *[SEP]*. The *[CLS]*, which stands for *Classification*, is the first input token and produces an output vector of dimension equal to *hidden size* that can be used as the input for an arbitrarily chosen classifier. Instead, *[SEP]* stand for *segments separation*.

[10]https://github.com/huggingface/transformers

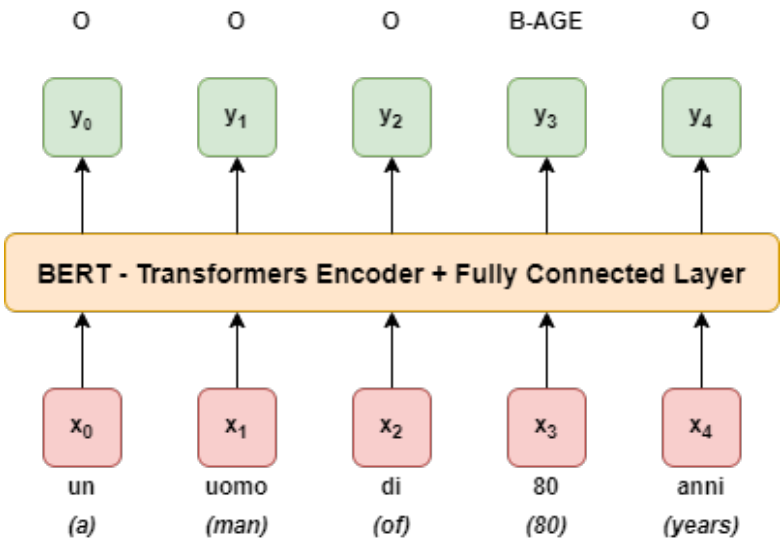


Figure 3.8: BERT architecture overview.

BERT, when used for NER, is fine-tuned without a CRF layer as output layer, following a diffused tagging task approach. Operating on the NER label set, the input provided to the token-level classifier uses the representation of the first sub-token. In detail, the final hidden representation h_i of each token i passes through the softmax function and the probability P is then calculated as follows:

$$P(t|h_i) = softmax(W_o H_i + b_o) \tag{3.22}$$

where $t \in T$ while W_o and b_o are weight parameters. During training the loss function used is categorical cross-entropy.

In particular, the Italian BERT models used with the Hugging Face framework are those made available by the MDZ Digital Library team at the Bavarian State Library[11]. For this study, the hyper-parameters shown in Table 3.9 were used. In detail, BERT-based models have $110M$ of parameters. Batch size and Maximum Sequence Length were set to 32 and 512 respectively, while the model was fine-tuned for 5 epochs. Attention heads, hidden size and hidden layers were 12, 768 and 12 respectively. The Italian BERT was trained on a source of data consist made by a recent Wikipedia dump and various texts from the OPUS corpora[12] collection with a final corpus size equal to about 13 GB and more than 2 billions tokens. Both the cased and the uncased versions were used.

[11]https://huggingface.co/dbmdz/
[12]http://opus.nlpl.eu/

Table 3.9: BERT-based model hyper-parameters.

Hyperparameter	**Value**
Attention heads	12
Batch size	32
Epochs	5
Hidden size	768
Hidden layers	12
Maximum Sequence Length	512
Parameters	110 M

All experiments were performed on an IBM POWER9 cluster with NVIDIA V100 GPUs. All models were trained and tested using the chosen division of the data set between training and testing, reporting the results rounded to the fourth decimal place.

3.3.3 Third use case

The third use case is related to the investigation of strategies for cross-linguistic transfer learning between high and low resources languages, specifically English and Italian, when applied to clinical de-identification, leveraging the mentioned corpora.

For the sake of clearness, an overview of the research aspects covered by this use case is given in Figure 3.9. In detail, Italian Medical Records constitute the primary input information, while English Medical Records constitute the broader additional information indicated in Figure with a red arrow and an extended graphical representation. The output is given by PHI predictions that represent the information to anonymise in order to make the Italian input documents compliant with privacy regulations. The central block represents all the different combinations of (i) network topologies, (ii) pre-trained embeddings and (iii) different training strategies as detailed in the yellow balloons.

To implement the systems described in this use case, two different frameworks have been used that offer different possibilities for NLP tasks like classification, NER, Part-of-Speech tagging, sense disambiguation and so on. The first one is Flair[13] (Akbik, Bergmann,

[13]https://alanakbik.github.io/flair.html

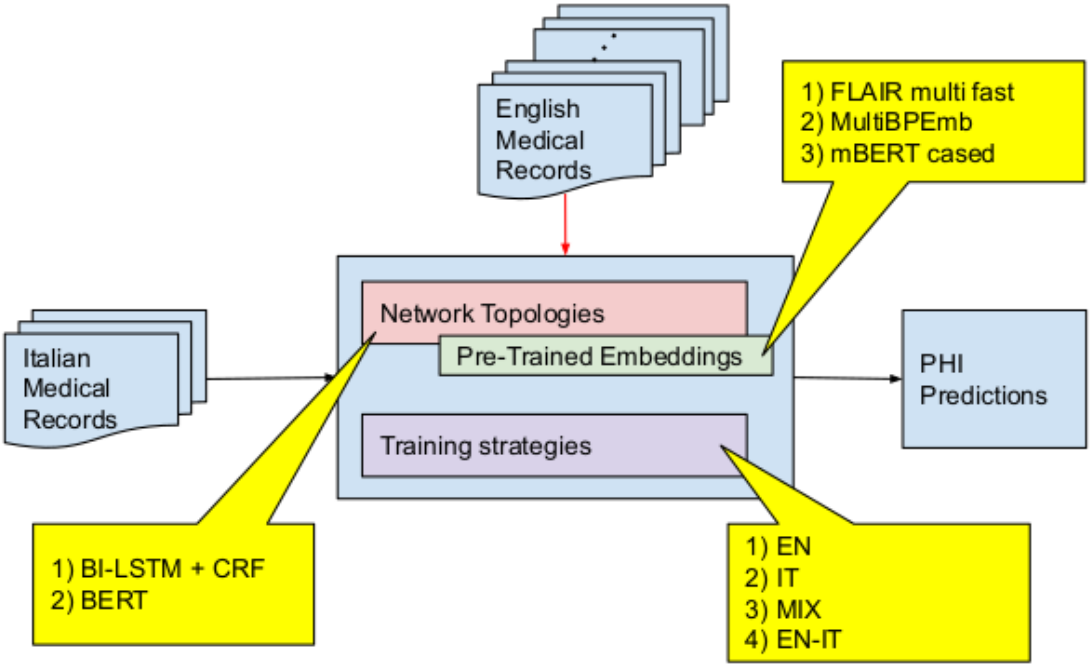


Figure 3.9: Third use case: research aspects overview.

Blythe, et al. 2019), written in Python: this framework has been used to implement the neural network based system Bi-LSTM + CRF, leaving default values not of interest and setting as shown in Table 3.10 the values relevant to the experimentation. The second is Hugging Face Transformers[14], also written in Python: this framework has been used to implement the system based on Transformers, so BERT. Similarly to what was done previously, only the values relevant to the experimentation have been modified and reported in Table 3.11. The hyper-parameters modified and reported in Table 3.10 and Table 3.11 are described below.

Regarding Flair, the stochastic gradient descent (SGD) was used to update neural network parameters. Every 3 epochs without improvement the learning rate is reduced according to *Patience* hyper-parameter, by multiplying the annealing factor, so it goes from 0.1 to 0.0001, the latter being a system condition of early stopping. For this reason, the 500 limit of training epochs is never reached but the number of epochs used is different for each trained model. Other hyper-parameters are: gradient clipping 5.0, Bi-LSTM hidden size 256, variational dropout 0.5, word dropout 0.05 and batch size 16.

[14]https://github.com/huggingface/transformers

Table 3.10: Bi-LSTM + CRF hyper-parameters

Hyperparameter	**Value**
Annealing factor	0.5
Batch size	16
Dropout	0.5 (variational) 0.05 (word)
Epochs	up to 500
Gradient clipping	5
Hidden size	256
Learning rate	0.1 - 0.0001
Patience	3
RNN Layers	1

Table 3.11: $\text{BERT}_{\text{base}}$ and mBERT hyper-parameters

Hyperparameter	**Value**
Attention heads	12
Batch size	32
Epochs	5
Hidden size	768
Languages	104
Hidden layers	12
Maximum Sequence Length	512
Parameters	110 M

Regarding HuggingFace Transformers, $\text{BERT}_{\text{base}}$ and mBERT implementations have both $110M$ of parameters. Batch size and Maximum Sequence Length were set to 32 and 512 respectively, while the model was fine-tuned for 5 epochs. Attention heads, hidden size and hidden layers were 12, 768 and 12 respectively.

An IBM POWER9 cluster with NVIDIA V100 GPUs was used to run the experiments. In detail, the tested models were based on:

1. Bi-LSTM + CRF with stacked embedding consisting of *MultiBPEmb* and *Flair embedding multi-fast* (both forward and backward);

2. mBERT Cased.

The models were trained using the strategies introduced in Section 3.1.3. In particular, the EN and IT strategies perform a training on i2b2 2014 and SIRM COVID-19 data sets respectively, while the MIX strategy provides a single concatenated i2b2 2014/SIRM COVID-19 data set for training, finally the EN-IT strategy performs a first training on i2b2 2014 data set and a second training on SIRM COVID-19 data set. All models were tested on SIRM COVID-19 testing data set (50 of 115 clinical records).

Finally all models were trained and tested repeating the procedure five times for each configuration and reporting the arithmetic mean of the results, rounded to the fourth decimal place.

3.4 Evaluation metrics

To assess the performance of the models and compare them the F_1 measure was used, defined as the harmonic mean of precision P and recall R. Defined TP as the number of true positives, FP the number of false positives and FN the number of false negatives, these metrics can be defined as:

$$F_1 = \frac{2 * P * R}{P + R} \tag{3.23}$$

$$P = \frac{TP}{TP + FP} = \frac{\#\ of\ correctly\ predicted\ items}{\#\ of\ predicted\ items} \tag{3.24}$$

$$R = \frac{TP}{TP + FN} = \frac{\#\ of\ correctly\ predicted\ items}{\#\ expected\ items} \tag{3.25}$$

where items are entities or tokens, depending on the evaluation criteria used, hereinafter described.

In the case of multi-class problems, the calculation of precision and recall can be done in different ways, considerably changing the resulting F_1 value. The most common calculation methods are the following:

- Micro-Averaging. The number of correct, predicted and expected entities of each class is added up. With their total values, precision and recall are calculated. In binary classification problems, Micro-Averaged F_1 is the same as accuracy.

- Macro-Averaging. The precision and recall values are calculated for each class. Then precision and recall are calculated as the arithmetic average of the precision and recall values. Hence F_1 is calculated by 3.23[15]. It should be reported its standard deviation also.

- Weighted Macro-Averaging. The precision and recall values are calculated for each class. Then precision and recall are calculated as the weighted average (related to the number of expected entities for each class) of the precision and recall values. Hence F_1 is calculated by 3.23[15].

In order to have the best degree of comparability with the baseline systems Micro-Averaging was used. Five evaluation criteria were adopted to produce the results: on the one hand, *entity* and *token*, on the other hand *binary*, *i2b2 category*, *i2b2 subcategory*. The *entity* criterion checks if a predicted entity exactly matches the correspondent in the so-called *gold standard* (also known as *ground truth*), i.e. when all tokens belonging to the entity are correctly recognized, while the *token* criterion checks only if there is a token match, which is considered correct even if it only partially covers the entity. This reasoning applies with increasing difficulty using the second set of criteria: in the case of *binary* criterion it is sufficient to distinguish entities and non-entities (or tokens and non-token), then for *i2b2 category* and *i2b2 subcategory* it is necessary to recognize categories and subcategories to which the entities or tokens belong respectively. Therefore, in *token-binary* cases the highest scores are obtained, while in *entity-subcategory* cases the lowest scores are obtained: the latter should be the main criterion to adopt.

[15]The final F_1 value should be calculated by 3.23 and not by arithmetic or weighted average of the F_1 values of the classes.

Chapter 4

Results and Discussions

This section collects results and discussions concerning the three use cases presented in the previous Section 3.3, presenting them in the same order in the following Sections 4.1, 4.2, 4.3 respectively.

4.1 First use case

In Table 4.1 the Micro-Averaged F_1 scores of the analysis at sub-document level are reported according to the criteria given in Section 3.4. In detail, the results were listed by different SGFs and grouped by *entity* and *token* levels, which in turn were divided by the *i2b2 sub-category*, *i2b2 category* and *binary* criteria.

Among the 6 different SGFs, the sub-document level with a $SGF = 32$ achieves the highest Micro-Averaged F_1 scores by adopting any possible combination of the criteria. Comparing the best configuration with $SGF = 32$ to the reference system at sentence level, i.e. $SGF = 1$, there are increases of 2.41%, 1.34% and 1.28% at entity level in sub-category, category and binary cases respectively. On the other hand, at token level, for sub-category, category and binary there are increases of 2.35%, 0.79% and 0.76%.

The detailed results of the best configuration, i.e. $SGF = 32$, with regard to i2b2 subcategories, categories and binary are reported in Table 4.2. Precision P, recall R and Micro-Averaged score F_1 are reported with E or T subscripts depending on whether they are at entity or token level.

Table 4.1: Sub-document level analysis - Micro-Averaged F_1 scores

	Entity Level			Token Level		
SGF	Sub-category	Category	Binary	Sub-category	Category	Binary
1	0.9239	0.9445	0.9486	0.9443	0.9713	0.9756
2	0.9343	0.9481	0.9508	0.9581	0.9741	0.9784
4	0.9409	0.9523	0.9559	0.9631	0.9760	0.9805
8	0.9435	0.9545	0.9581	0.9646	0.9774	0.9821
16	0.9446	0.9552	0.9590	0.9655	0.9776	0.9822
32	**0.9480**	**0.9579**	**0.9614**	**0.9678**	**0.9792**	**0.9832**

For the sake of completeness $\text{BERT}_{\text{base}}$ was also tested in its default configuration, batch size of 32 for 5 fine-tuning epochs to perform NER on the data set, also experimenting different *sequences grouping factors*. In detail, it was evaluated both with $SGF = 1$ (and *maximum sequence length* fixed to 256) as baseline and with $SGF = 11$ (and *maximum sequence length* fixed to 512): this choice is dictated by the limit of BERT to be able to manage maximum up to 512 token, so it has been chosen a SGF that prevents the 512 token from being exceeded in training, validation and testing data sets. The maximum number of tokens detected by BertTokenizer, according to which the *maximum sequence length* was chosen, is shown in Table 4.3 for each data set and for each SGF. Although a further research point could be to compare the results of BERT versions that have a *maximum sequence length* of 1024 or 2048, it must be taken into account that the hardware resources already required to fine-tune $\text{BERT}_{\text{base}}$ are much higher than those required by Bi-LSTM + CRF-based systems. The results of BERT obtained with the two different SGFs are in Table 4.4, showing worse results than the best setup.

In Table 4.5 the Micro-Averaged F_1 scores obtained with the best configuration are reported, then compared with those reported by H. Yang and Garibaldi 2015, Z. Liu, Tang, et al. 2017, Y. Kim, Heider, and Stéphane M. Meystre 2018 and Tang, D. Jiang, et al. 2019 which are the highest obtained for clinical de-identification using the i2b2 2014 data set to date. Where *NA* is reported means that no results have been provided for that level of analysis. In particular, the best system always obtains better scores compared with H. Yang and Garibaldi 2015 and Y. Kim, Heider, and Stéphane M. Meystre 2018,

Table 4.2: Best SGF - Averaged scores

i2b2 Sub-category	$\mathbf{P_E}$	$\mathbf{R_E}$	$\mathbf{F_{1E}}$	$\mathbf{P_T}$	$\mathbf{R_T}$	$\mathbf{F_{1T}}$
AGE	0.9644	0.9442	0.9542	0.9668	0.9418	0.9541
CITY	0.8230	0.8800	0.8506	0.8645	0.9159	0.8894
COUNTRY	0.8029	0.6615	0.7254	0.9228	0.6985	0.7951
DATE	0.9821	0.9776	0.9798	0.9918	0.9874	0.9896
DEVICE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DOCTOR	0.9629	0.9550	0.9589	0.9735	0.9740	0.9738
EMAIL	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
FAX	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
HOSPITAL	0.9145	0.8679	0.8906	0.9529	0.9263	0.9394
IDNUM	0.8746	0.8226	0.8478	0.9506	0.8890	0.9188
LOCATION_OTHER	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MEDICALRECORD	0.9683	0.9834	0.9758	0.9751	0.9837	0.9794
ORGANIZATION	0.7834	0.5293	0.6317	0.9046	0.5852	0.7107
PATIENT	0.9409	0.9354	0.9382	0.9766	0.9561	0.9663
PHONE	0.9729	0.9684	0.9706	0.9858	0.9778	0.9818
PROFESSION	0.8511	0.7665	0.8066	0.9575	0.8238	0.8856
STATE	0.8491	0.8768	0.8628	0.9021	0.8592	0.8802
STREET	0.8833	0.9015	0.8923	0.9588	0.9856	0.9720
USERNAME	0.9955	0.9565	0.9756	0.9955	0.9565	0.9756
ZIP	0.9802	0.9914	0.9858	0.9879	0.9932	0.9906
Total	0.9552	0.9409	**0.9480**	0.9754	0.9602	**0.9677**
i2b2 Category	$\mathbf{P_E}$	$\mathbf{R_E}$	$\mathbf{F_{1E}}$	$\mathbf{P_T}$	$\mathbf{R_T}$	$\mathbf{F_{1T}}$
AGE	0.9644	0.9442	0.9542	0.9668	0.9418	0.9541
CONTACT	0.9796	0.9697	0.9746	0.9953	0.9815	0.9883
DATE	0.9821	0.9776	0.9798	0.9918	0.9874	0.9896
ID	0.9588	0.9386	0.9486	0.9838	0.9645	0.9740
LOCATION	0.9170	0.8777	0.8969	0.9748	0.9344	0.9542
NAME	0.9724	0.9641	0.9682	0.9908	0.9835	0.9871
PROFESSION	0.8511	0.7665	0.8066	0.9575	0.8238	0.8856
Total	0.9653	0.9506	**0.9579**	0.9870	0.9716	**0.9792**
i2b2 Binary	$\mathbf{P_E}$	$\mathbf{R_E}$	$\mathbf{F_{1E}}$	$\mathbf{P_T}$	$\mathbf{R_T}$	$\mathbf{F_{1T}}$
NAMED ENTITY	0.9693	0.9536	**0.9614**	0.9910	0.9755	**0.9832**

Table 4.3: Maximum number of tokens detected by BertTokenizer

SGF	**Training data set**	**Validation data set**	**Test data set**
1	212	242	203
11	474	425	464

Table 4.4: BERT results

Model	Entity Level			Token Level		
	Sub-category	Category	Binary	Sub-category	Category	Binary
BERT SGF=1	0.9215	0.9430	0.9509	0.9470	0.9698	0.9758
BERT SGF=11	0.9302	0.9452	0.9520	0.9598	0.9730	0.9786

while it has slightly different performance, at category level and at binary level, compared with Z. Liu, Tang, et al. 2017 and Tang, D. Jiang, et al. 2019.

Table 4.5: Micro-Averaged F_1 scores comparison

Model	Entity Level			Token Level		
	Sub-category	Category	Binary	Sub-category	Category	Binary
H. Yang and Garibaldi 2015	NA	0.9360	NA	NA	0.9611	NA
Z. Liu, Tang, et al. 2017	NA	0.9511	0.9650	NA	0.9698	0.9828
Y. Kim, Heider, and Stéphane M. Meystre 2018	NA	0.9573	NA	NA	NA	NA
Tang, D. Jiang, et al. 2019	NA	0.9550	**0.9685**	NA	0.9748	**0.9870**
$SGF = 32$	0.9480	**0.9579**	0.9614	0.9678	**0.9792**	0.9832

The results show the effectiveness of the sub-document analysis method introduced, thanks to which it is possible to identify the best grouping of sentences to be provided to the system in order to maximize the exploitation of contextual information, as stated by Khandelwal et al. 2018. In this way it is possible to obtain results that improve the state of the art of the NER task applied to de-identification as a multi-class problem, therefore at category level which, unlike previous works focused on the binary level, is much more important to be able to properly exploit such systems for anonymisation.

With regard to BERT, its performance confirms what was reported by Devlin et al. 2019, i.e. the way BERT works is problematic when applied as a feature-based approach to NER. Moreover, the need to insert *special* tokens inside the text by BERT, such as *[CLS]* and *[SEP]*, reduces the number of useful tokens managed by BertTokenizer and consequently the useful context.

Compared with the other state-of-the-art systems, the proposed system obtains better scores at category level: this is extremely relevant in a de-identification scenario, where the next step is anonymization and, thus, it is important to get the best possible results at a finer level than binary to better replace entities with

their surrogates. Moreover, even if the proposed system obtains a slightly worse score at the binary level than Z. Liu, Tang, et al. 2017 and Tang, D. Jiang, et al. 2019, this is counterbalanced by not using feature engineering and handcrafted rules, saving time for analysis and implementation. In fact, based on the data publicly available, Z. Liu, Tang, et al. 2017 and Tang, D. Jiang, et al. 2019 get better results at the binary level for those categories (e.g. fax, email, device) that are present both in small amounts and in different forms, and often split into multiple tokens, which is the worst case for learning by a neural network without the help of, for example, regular expressions.

4.1.1 Some examples of entity classification with the proposed system

In the proposed approach embeddings that work both at word and character level and in a contextual manner are stacked: in this way it is possible to detect more entities and classify almost all of them correctly.

In particular, working at character level the proposed system identifies entities introduced by an abbreviated form or an acronym, for instance entities of type *NAME: DOCTOR*, like *David Mccall*, that are preceded by the abbreviated form *Dr.*. In the same way, some tokens after the full name, both the shortened names of doctors, such as *RX* and *GV*, and those written with a lowercase initial, such as *rosenberg*, are correctly detected. Morpho-syntactic variations caused by writing errors are also overcome: for example entities of type *PROFESSION buidling construction* and *mathmatics* are correctly recognized. Similarly, *'s* that are part of *DATE*s are detected correctly. Moreover, polysemy is correctly managed within the embeddings: as reported in Table 4.6, the token *Jordan* is annotated in some cases as *LOCATION: CITY*, in others as *NAME: DOCTOR*, and the recognition system is able to take into account the multiplicity of meaning by predicting the exact label depending on the case.

The use of word level embedding, such as GloVe, allows us to identify entities never seen before. An example are the two *PROFESSION*s *Ironworker* and *vocational instructor*: *Ironworker* is semantically similar to the word *worker*, already present as *PROFES-*

Table 4.6: Polysemous entities

i2b2 category: subcategory	Extracted sentence
LOCATION: CITY	Thank you for referring this interesting patient. She was seen and examined with Dr. Voss. Sincerely, Pamela Imperial, M.D. PI / waldron cc: Charles Van, M.D. **Jordan**, FL 83712
LOCATION: CITY	Record date: 2074-06-20 HHH Cardiovascular Division CVD Rm 5 89 Buck St JORGENSON, VIVIANLEE Sioux City, FL 76546 47190847 (179)732-8159 06/20/74 Charles F. Van, M.D. 66 Kessler Farm Drive, Puite # 9488 **Jordan**, FL 83712
NAME: DOCTOR	Return to see in a week's time, understands. He can discuss any management issues regarding his insulin doses or others anytime. Yale **Jordan**, M.D. YJ: isenberg: 74:585:40; 00-10142
NAME: DOCTOR	Thank you for allowing me to participate in the care of your patient. I will continue to follow along with you. Sincerely, **Jordan** N. Akers, M.D., Ph.D. Department of Neurology Stroke Division Orange City Hospital Prestonsburg, MS 54151

SION in the training data set, while for *vocational instructor* the semantic similarity is exploited with other labeled terms of which, for the sake of completeness, both cosine similarity and number of occurrences in the training data set are reported in Table 4.7. In the same way, the never seen entities of type *LOCATION: CITY*, such as *Pecos* and *Turlock*, are identified thanks to the presence of the introductory formulas *life in* and *came from* respectively, semantically close to similar forms present in the training data set to introduce entities of type *LOCATION: CITY* such as *lives in*, *lived in*, *living in*, *from*, *comes back from*.

Table 4.7: The most semantically similar words to *instructor* in GloVe embeddings

Word	Cosine similarity	Occurrences
instructor	1	1
instructors	0.7111440300941467	0
teacher	0.7077120542526245	1
training	0.6565847396850586	0
technician	0.6412791013717651	5
taught	0.6199019551277161	0
mechanic	0.6190917491912842	0
gunnery	0.6142827272415161	0
sergeant	0.6117113828659058	0
graduate	0.6066259145736694	0
teaching	0.5975379943847656	0

Finally, by exploiting a wider context thanks to a higher SGF,

the neural network is able to find the necessary patterns to identify other unknown entities. Some examples are provided in Table 4.8 where, for clarity, it is reported only a subset of relevant sentences. So, using only $SGF = 1$, an entity like *goods purchaser* belonging to the i2b2 category *PROFESSION* is not detected within a short sentence like *Used to be goods purchaser for 34 years.*. This is because with $SGF = 1$ the generally recurrent patterns that allow to recognize entities despite the little contextual information content are different in these particular cases. For example, in the case of *PROFESSION*s, the common pattern recurring within short sentences is similar to *Was a* or *Retired*, thus followed by the profession. Instead, in the cases examined there are either (1) different phrasal expressions, such as *Used to be* and *Used to do*, or (2) expressions with both grammatical errors that contribute to changing the meanings at stake, for example *He is retried*, and longer entities composed of more than one or two tokens, for example *Motor Vehicle Body Repairer*, complicating recognition. Or, in the case of *LOCATION: ORGANIZATION*s, there are (3) mainly unusual expressions, e.g. *Then*, compared to more canonical introductory formulas, e.g. *worked for*, but, extending the context, there is a list of organizations still introduced by a known pattern. By looking at the sentences in Table 4.8 it is possible to observe how the extended context can provide further information. Professions and organizations are often reported within a specific part of the structure of the medical record titled *Social History*, written in full or abbreviated, upper or lower case, so that it is reasonable to wait for such an entity in the next sentences: in essence the pattern *Social History* influences the weights of the network in relation to the most appropriate label prediction.

Moreover, it is interesting to note that with the use of the proposed system that work with a character-level language model able to take into account the context as Flair, it was possible to correctly recognize a badly annotated entity with a pattern not perfectly identical to the training or validation entities. An important example in this regard is the entity *778 210-2105*, which was annotated as *CONTACT: PHONE*, even if within the text it is preceded by the text *Fax:*. In this case the proposed system recognizes it correctly for us humans as *CONTACT: FAX*, but being wrong the annotation this unfortunately contributes to a lowering of the score.

Table 4.8: Entities identified through an extended context with $SGF > 1$

Extracted sentences with identified entities in bold
His father died at 65 from Alzheimer's disease, and his mother at 50 from an unspecified cancer. No other known family members with heart disease. SOCIAL HISTORY Presently lives in Lake Pocotopaug with wife and son Brandon. Used to be **goods purchaser** for 34 years. He has a 60 pack - year smoking history, but has since quit. Last drink of EtOH was 1.5 years ago.
Allergies Penicillins - Angiodema, Hives, Erythromycins - Hives, Iv Contrast - convulsions, Hypotension Social History: Single, divorced. 3 children, 4 grandchildren. Retired, used to do **immigration policing**. Lives in Havre De Grace with his brother. Dating current girlfriend x 3 years now, sexually active.
Family history: Noncontributory. No coronary artery disease or MI. Social history: He is retried **Motor Vehicle Body Repairer**. He used to work at GM. He lives with his wife, who comes in today with him. He has three children. Nondrinker.
Social History Married. Lives with wife. Sexually active, usually needs Viagra. Communications senior manager, marketing, worked for Brinker International, now Facebook. Then **Twitter**.

4.1.2 Error analysis and distribution

In the following sections, in order to understand the limits of the proposed system, its errors are examined in detail trying to understand the reasons for missing or wrong classifications. In particular, false positives and false negatives distribution are analyzed, hence some examples of unidentified entities in Section 4.1.2 are reported. Instead, in Section 4.1.2, those entities that are difficult to identify are inspected.

FP and FN distribution

The performance of the best system are analyzed in detail, examining the distribution of False Positives (FPs) and False Negatives (FNs), whether at entity or token level, whose results are summarized in the Figure 4.1.

The different distribution of false positives and false negatives of the categories between entity level and token level is also justified by the different token-entity ratio for each of them, as shown in Figure 4.2.

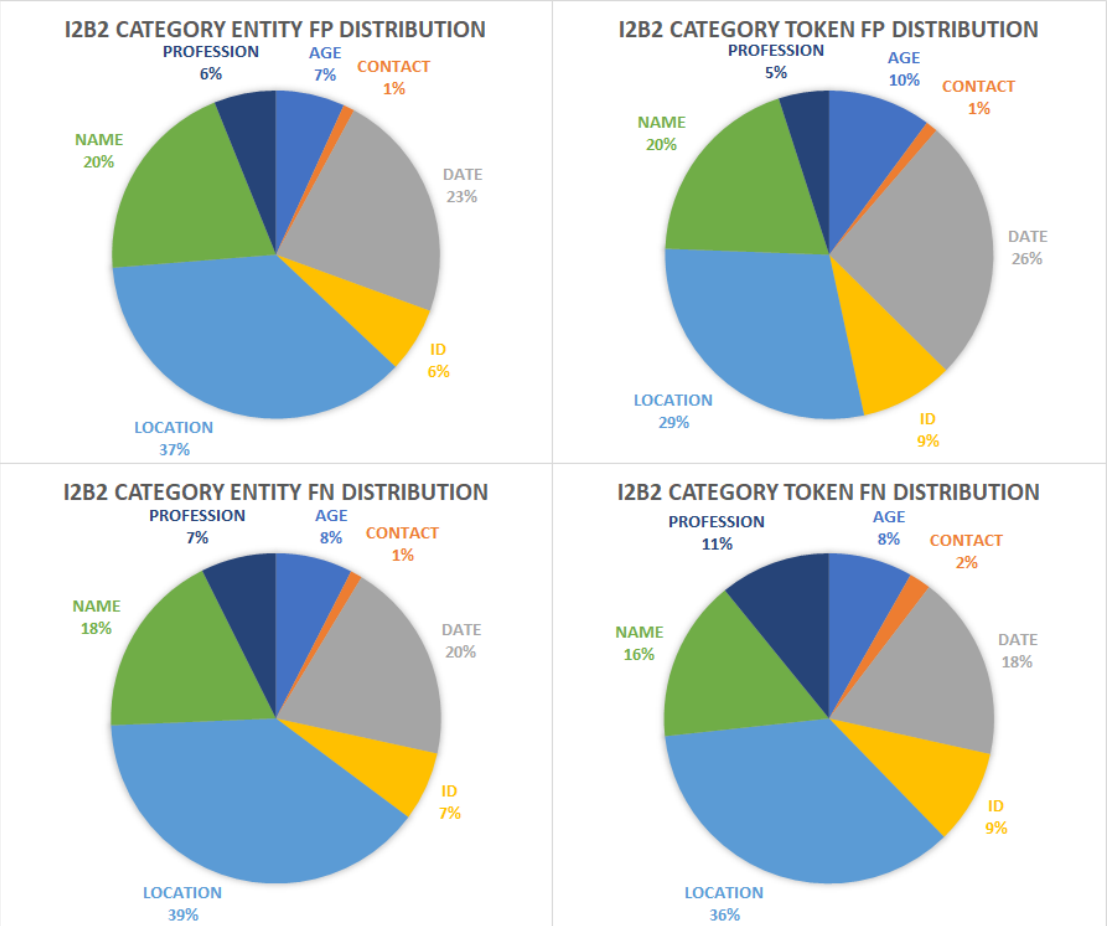


Figure 4.1: Error distribution helps to understand the weaknesses of the specific architecture used for the specific data set. This figure is related to the best proposed system with $SGF = 32$.

Moreover, FPs and FNs have been associated to four main sources of errors, in line with Dernoncourt, J. Y. Lee, Uzuner, et al. 2017: (1) the abbreviations, due to brevity and vagueness; (2) the ambiguities, due to the tokens themselves and/or the context; (3) the debatable annotations, i.e. the tokens and the entities marked as PHI in a questionable way; (4) the scarcity and sparsity of data, with respect to the training, validation and testing data sets.

In Table 4.9 some examples of unidentified entities due to previous listed error types are shown: in detail AB, AM, D, S stand for error types (1), (2), (3) and (4) respectively as reported above. The examined entity is underlined and the unrecognized part is in bold, while the black/red variation indicates tokenization. There were no entities of type *LOCATION: ZIP CODE* to which no label was

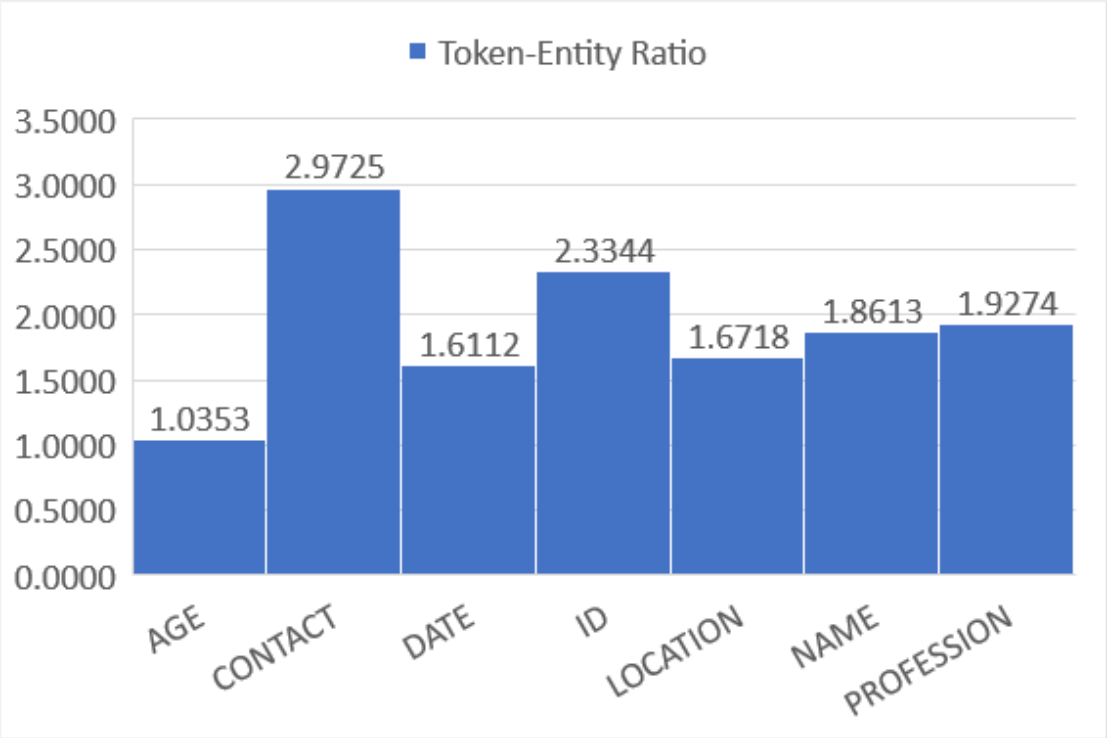


Figure 4.2: The token-entity ratio of i2b2 categories.

assigned, so *NA* is reported. Instead, for the entities of type *CONTACT: EMAIL*, *CONTACT: FAX*, *ID: ID NUMBER*, and *LOCATION: OTHER* the main source of error leading to non-recognition or misrecognition is due both to the scarcity and poor distribution of data within data sets and to non-repetitive patterns: there is a dedicated analysis to these entities in Section 4.1.2.

Challenging entities analysis

The most challenging entities in the i2b2 data set are listed in Table 4.10 to better investigate. They belong to i2b2 subcategories *CONTACT: EMAIL*, *CONTACT: FAX*, *ID: DEVICE* and *LOCATION: OTHER* and are fully reported, i.e. there are no other entities in the 2014 i2b2 data set other than those enumerated. The black-red variation indicates the tokenization: for example, the first entity of the *CONTACT: FAX* subcategory in the test data set, i.e. *(385)031-7905*, consists of 6 tokens. For each of the four subcategories there are important considerations to be made.

Consider *CONTACT: EMAIL* and *CONTACT: FAX*: in both cases there are few examples but there is an important difference

Table 4.9: Unidentified entities examples

PHI category: subcategory	Context	Error Type
AGE	Synthroid (LEVOTHYROXINE SODIUM) **150MCG** TABLET PO variable Torsemide	D
	FH: M & # 8211; CAD **60s**, F & # 8211; CAD **80s**	AB, AM
CONTACT: PHONE	Wheatland Manor: 154-735-1487, x **557** (4th floor)	AB, AM
	type 2 Hypertriglyceridemia H/O paroxysmal afib VNA **171-311-7974** =======	AM
DATE	HISTORY: -ESRD (FSGS) on HD **Mon**, **We d**, **Fri** s/p failed in renal transplants	AB, AM
	Month(s) Supply, Q Sun, **Mon**, **We d**, **Fri** Aspirin (ACETYLSALICYLIC ACID)	AB, AM
ID: ID NUMBER	By: CHANEY, QUENTIN eScription document: 7-9617124 **SJZvdbs** ***** ***	D, AM
	Qiana Solomon, MD eScription document: 9-2784353 **KUQlhv** **Egq** DD: 03/06/87	D, AM
ID: MEDICAL RECORD	Patient: Vincent Ware (71417347 **2Y**) Student: Casey Best,	AM
LOCATION: CITY	She never has felt the **Pecos** as home and is home sick for	D
	M.D. cc: Dirk O. Reece, M.D., PCP, **Pune**, ME SRH / valdovinos	AM
LOCATION: COUNTRY	AS HPI : 54 y/o **Columbian** speaking male with HTN	D
	General Appearance PLEASANT GENTLEMAN OF **GERMAN** EXTRACTION	D
LOCATION: HOSPITAL	Mike Ivan, MD, **EHMS** pager 84710	AB, AM
	Dr. Earle to contact from the **WBM** office, however, my calls	AB, AM
LOCATION: ORGANIZATION	some high in Na (**atkins** he tried) and others not good for	AM
	Formerly in the **marines**. Had lived in Poland.	AM, D
LOCATION: STATE	M.D. cc: Dirk O. Reece, M.D., PCP, Pune, **ME** SRH / valdovinos	AM
LOCATION: STREET	POC 112 11/05/2095 HGBA1C 7.8 (*) Spanish **30** Dan Chan	AM, D
LOCATION: ZIP CODE	*NA*	*NA*
NAME: DOCTOR	Attending: **YBARRA** CODE: FULL HPI: 70	AM
	He has met with Drs. Eagle and **Yzaguirre**.	AM
	anti-biotics as outlined by Dr. Infant-**Nickel**. His ARF has	AM
	Best wishes, **O**	AB
NAME: PATIENT	Your patient **Earnest Branch** came in the office today for	S
	UHER, OLGA 12/31/63 instituted to achieve an LDL cholesterol	S
NAME: USERNAME	Peter Quale, **IJ6** pgr 20951	AM, S
PROFESSION	Social History **NP** in Laplace	AB, AM, S
	Volunteers - **animal rescue**. No current or previous tobacco.	AM, S

between the two categories. While in the first case there is at least one entity in the validation data set and the same pattern is always respected, i.e. with @ and *.org*, in the second case a greater presence of examples corresponds to an absence of recurrent patterns between the training, validation and testing data sets: this situation can only worsen the performance of a recognition system whose learning is based exclusively on deep neural networks.

Look at the cases of *ID: DEVICE* and *LOCATION: OTHER*. This time, in addition to the scarcity of examples and the absence of repetitive patterns, there is also an improper distribution of entities in the data set, resulting in the complete absence of examples in the validation data set and contributing to confusing the learning system that becomes more susceptible to overfitting. Between the two categories, however, there is an important difference: if in the first case no features suitable for recognition seem to have been acquired, in the second case the presence of two successive words both with capital letters seems to be a strong clue for the presence of an entity: this assumption is supported by the massive presence within the data set of this pattern also in other categories, such as *LOCA-*

Table 4.10: Challenging entities

Entity sub-category	Train data set	Validation data set	Test data set	Proposed system prediction
CONTACT: EMAIL	yfcooley@wsh.org yfcooley@wsh.org gmichael@KCM.ORG	vmeadows@sbhnc.org	iparedes@oachosp.org	CONTACT: EMAIL
CONTACT: FAX	966-221-9723 595-442-5450 664-577-0339 534-184-9285 648-875-5821	192 7991 320.821.2954 320.821.2954	(385)031-7905 (251)628-xxxx	CONTACT: PHONE O
ID: DEVICE	193062 358892 8068103 17722GNP 30058 56520YOG 9YS7EZ		QQ 626 CTE 226 5435 4712198 5167DH/20 SQ462162 0049EO/46 LX39426	O O O ID: MEDICAL RECORD O ID: ID NUMBER O ID: ID NUMBER
LOCATION: OTHER	the Midwest the midwest Fountain Of The Four Rivers GOLDEN GATE BRIDGE		Cape Cod Central Park Rockefeller Centre global Storting Capitol Rockefeller Centre global Storting Acropolis long island long island long island	LOCATION: CITY LOCATION: CITY LOCATION: ORGANIZATION O O O LOCATION: HOSPITAL O O O O O O

TION and *NAME*, which constitute almost half of the annotations. In fact, other entities are either too generic to be identified (perhaps even annotated), like *global*, or suffer even more ambiguity caused by the absence of capital initials, like *long island*. Finally, entities like *Storting*, *Capitol* and *Acropolis* fall within the problem of out-of-vocabulary words: other de-identification systems, unlike the one proposed, have solved this problem through the use of gazetteers and improved F_1 accordingly.

4.1.3 Ablation analysis

To establish the importance of the various main components making up the proposed Bi-LSTM + CRF-based architecture, 5 variants of the model are tested by eliminating several elements one at a time or in pairs, as Dernoncourt, J. Y. Lee, Uzuner, et al. 2017 and Akbik, Blythe, and Vollgraf 2018. Figure 4.3 presents the results of the ablation tests.

GloVe embeddings have the lowest weight, whose removal leads to a reduction of 0.33% and 0.31% at entity and token level respectively, compared to significant -7.13% and -3.98% by removing the

Flair embeddings. Although there is a decrease in recognition ability probably caused by a poor capture of polysemy and therefore of semantics due to the removal of GloVe embeddings working at word level, removing Flair embeddings is an evidence of the fact that their ability to work at character level and capture morpho-syntactic variations is particularly useful in the case of handwritten texts due to the second lowest score obtained.

The Bi-LSTM + CRF was replaced with a linear feedforward architecture, i.e. a multinomial logistic regression (Menard 2002), with and without CRF decoding layer. In fact, starting from the linear hidden layer $\mathbf{h}_t = \mathbf{W}_h\mathbf{x}_t + \mathbf{b}_h$ given by the feedforward network, the label prediction is given by $P(\mathbf{y}_t = j|\mathbf{h}_t) = softmax(\mathbf{h}_t)[j]$. It is interesting to note that the removal of the LSTM layers and the CRF layer has different impacts depending on whether you analyze the entity level (-3.69% and -3.13%) or the token level (-2.16% and -1.04%): the Bi-LSTM has a greater impact on performance at both token and entity level than the CRF. This is due to the fact that sentence level tag information passing through a CRF layer makes it more efficient to use closer past and future tags to predict the current tag, while observing a long series of close tags through the Bi-LSTM layer provides the best representation of the context and, therefore, of structurally correct tag sequences, helping to correctly recognize more single and multi-token entities. The non-linear and exceptionally large reduction obtained by removing both layers (-14.95% at the entity level and -5.7% at the token level) presents a discrepancy of about 10 percentage points between the entity and token levels, underlining the importance of the former as the main measure for the evaluation of NER systems. Overall, in these approaches predictions are made directly on the basis of the embeddings provided without further learning from both the recurrence of Bi-LSTM and the presence of the context, justifying lower scores.

Finally, each removal has resulted in a more or less marked reduction in performance. This suggests that the insights at the basis of the choices made to build the proposed architecture have each its own dignity, all contributing to improve the results.

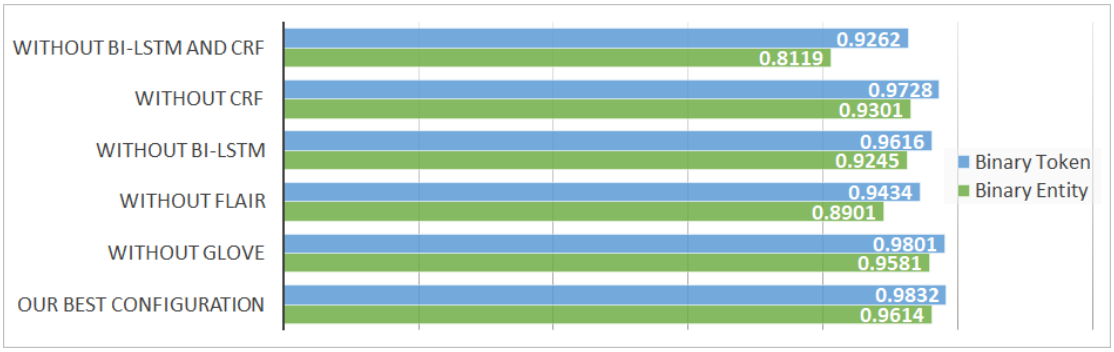


Figure 4.3: Ablation test performance.

Table 4.11: Micro-Averaged F_1 results.

Model	Embedding	Entity Level			Token Level		
		Subcategory	Category	Binary	Subcategory	Category	Binary
Bi-LSTM + CRF	FastText	0.7034	0.7130	0.7297	0.7821	0.8155	0.8395
Bi-LSTM + CRF	Flair	**0.8100**	0.8224	0.8289	0.8797	0.9045	0.9211
Bi-LSTM + CRF	FastText + Flair	0.8063	**0.8294**	**0.8308**	**0.8850**	**0.9116**	0.9211
$\text{BERT}_{\text{BASE}}$ Uncased	-	0.6442	0.6667	0.6848	0.7667	0.8083	0.8796
$\text{BERT}_{\text{BASE}}$ Cased	-	0.7553	0.7880	0.7969	0.8561	0.8979	**0.9260**

4.2 Second use case

The Micro-Averaged F_1[1] scores of all tested models and related embeddings are shown in Table 4.11, ordered in accordance with the criteria given in Section 3.4.

In regard to the Bi-LSTM + CRF model, FastText embedding, working at the sub-word level and managing semantic similarity accordingly, can better detect entities. Flair embedding, instead, relies more on its ability to exploit the context and manage polysemy. While individually FastText and Flair embeddings have comparable performance, a stacked embedding of their combination improves overall performance and is also the best method.

In addition, $\text{BERT}_{\text{BASE}}$ Uncased achieves significantly lower results than the Cased version: this underlines the importance of training systems capable of distinguishing upper and lower case for clinical

[1]Depending on how precision and recall are calculated, different types of F_1 can be obtained. In Micro-Averaging, the number of correct, predicted and expected entities or tokens of each class is added up and, with their total values, precision and recall are calculated. In Macro-Averaging, precision and recall values are calculated for each class, then overall precision and recall are calculated as the arithmetic average of class values. Instead, in Weighted Macro-Averaging, overall precision and recall are calculated as the weighted average (related to the number of expected entities or tokens for each class) of the precision and recall values.

de-identification. In fact, in this sub-task of the NER, the Named Entities are often proper names, of people, places, or things, and therefore written with capital letters.

According to the results obtained, the Bi-LSTM + CRF model with the proposed stacked embedding (FastText plus Flair) performs better than all the others. It outperforms models made with FastText or Flair embeddings only and the $\text{BERT}_{\text{BASE}}$ Uncased model. Instead, the $\text{BERT}_{\text{BASE}}$ Cased model is outperformed in all metrics except one: it is important to underline that the $\text{BERT}_{\text{BASE}}$ Cased model outperforms the Bi-LSTM + CRF model with the FastText plus Flair stacked embedding only at binary token level, the least significant to evaluate the performance of a NER system. It is of particular importance to consider this aspect in a de-identification scenario: in fact, the next step in this process is generally anonymisation, so it is necessary to obtain correct results at the most refined level of classification in order to replace the identified entities with valid surrogates Vincze and Farkas 2014, e.g. replacing a date with the surrogate of an ID number would allow the reader to easily identify the point of substitution by opening the door for an unwanted re-identification.

Therefore, although the data set is modest in size, using pretrained embeddings and language models it is possible to obtain good performance. The Bi-LSTM + CRF model with the proposed stacked embedding made by FastText plus Flair showed superior performance compared to all other models analyzed: its detailed results are reported in Table 4.12. The subscripts E and T indicate Entity or Token level respectively.

Analysing the results obtained, it is possible to identify some aspects undoubtedly related to the type of data set. To better support this analysis, it is introduced in Table 4.13 the Token/Entity ratio (indicated as T/E in the Table for short) for each subcategory, calculated on the basis of the entities present in the data set and on how many tokens make up each entity.

First of all, the *AGE* category is the only one to obtain high and identical results both at entity and token level: this is due to the general coincidence between the two levels, being the Token/Entity ratio equal to 1 in this case. Moreover all the entities are of numerical type, with few exceptions as for example the entity *sei* (six) and *47aa* (47yo).

Table 4.12: Detailed results obtained by the best model Bi-LSTM + CRF with stacked FastText + Flair embedding

i2b2 Sub-Category	$\mathbf{P_E}$	$\mathbf{R_E}$	$\mathbf{F_{1E}}$	$\mathbf{P_T}$	$\mathbf{R_T}$	$\mathbf{F_{1T}}$
AGE	1.0000	0.8909	0.9423	1.0000	0.8909	0.9423
CITY	0.4872	0.4191	0.4494	0.8452	0.4804	0.6108
COUNTRY	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DATE	0.9022	0.7445	0.8151	0.9414	0.6914	0.7966
DOCTOR	0.8893	0.8312	0.8590	0.9825	0.8659	0.9203
HOSPITAL	0.6073	0.7955	0.6884	0.8891	0.9490	0.9180
IDNUM	0.9981	0.8140	0.8966	0.9981	0.6562	0.7919
LOCATION OTHER	0.5000	0.1000	0.1643	0.5000	0.1000	0.1643
ORGANIZATION	0.3186	0.3555	0.3305	0.8620	0.6000	0.6830
PHONE	0.9600	0.5428	0.6897	0.9600	0.5428	0.6897
PROFESSION	0.8773	0.5778	0.6960	1.0000	0.5692	0.7246
URL	0.9422	0.9421	0.9421	0.9974	0.8064	0.8918
Total	0.8323	0.7819	**0.8063**	0.9408	0.8355	**0.8850**
i2b2 Category	$\mathbf{P_E}$	$\mathbf{R_E}$	$\mathbf{F_{1E}}$	$\mathbf{P_T}$	$\mathbf{R_T}$	$\mathbf{F_{1T}}$
AGE	1.0000	0.8909	0.9423	1.0000	0.8909	0.9423
CONTACT	0.9426	0.9084	0.9251	0.9951	0.7881	0.8796
DATE	0.9022	0.7445	0.8151	0.9414	0.6914	0.7966
ID	0.9981	0.8140	0.8966	0.9981	0.6562	0.7919
LOCATION	0.6785	0.7080	0.6928	0.9499	0.9341	0.9419
NAME	0.8893	0.8312	0.8590	0.9825	0.8659	0.9203
PROFESSION	0.8773	0.5778	0.6960	1.0000	0.5692	0.7246
Total	0.8626	0.7986	**0.8294**	0.9691	0.8606	**0.9116**
i2b2 Binary	$\mathbf{P_E}$	$\mathbf{R_E}$	$\mathbf{F_{1E}}$	$\mathbf{P_T}$	$\mathbf{R_T}$	$\mathbf{F_{1T}}$
NAMED ENTITY	0.8664	0.7982	**0.8308**	0.9792	0.8697	**0.9211**

Table 4.13: Token/Entity ratio per subcategories.

PHI C:Subcategory	**# of Tokens**	**# of Entities**	**T/E**
AGE	118	118	1.0000
C:PHONE	10	10	1.0000
C:URL	160	142	1.1268
DATE	200	154	1.2987
I:ID NUMBER	297	266	1.1165
L:CITY	139	101	1.3762
L:COUNTRY	6	6	1.0000
L:HOSPITAL	1297	266	4.8759
L:ORGANIZATION	56	13	4.3077
L:OTHER	9	9	1.0000
N:DOCTOR	1564	733	2.1337
N:PATIENT	3	3	1.0000
PROFESSION	96	65	1.4769

The *CONTACT* category, although not as high and symmetrical, still obtains important results. In detail, this category is composed mainly of entities of type *URL* and minimally by entities of type *PHONE*. In particular, the entities of type *URL* can rely on rather repetitive patterns and, if broken on several tokens, on always the same introductory formulas (e.g. *http* and *www*). In the case of the entities of type *PHONE*, the only entity present is *118*: the subcategory is reduced in this case to a single numerical almost always recognized.

The *DATE* category, both at entity and token level, averages around a F_1 of 80%. Several considerations about the existing entities come into play here. The most often recurrent pattern is that of the type *gg/mm/yyyy* but not always in the same variant and for this reason it is not always identified: in some cases it is found *g/m/yyyy* or *gg/m/yyyy* or *gg.mm.yyyy* or *g/m* or *gg/m* or *yyyy - mm - dd*. Equally often there are the single entities *2020* or *marzo* and *febbraio* but it is often possible to find the English variants of the months of the year *January*, *February* or *Feb*, *March* and *April* or *April* because they refer to international studies of medical colleagues. Therefore the abundance of patterns not always numerous makes the recognition task less easy.

The category *ID* presents instead many mono-token entities introduced by the same formula (e.g. the numbers from 1 to 115 that indicate the medical records preceded by the pattern *COVID-19: caso* (COVID-19: case)) that contribute to keep the result especially high at entity level. However, the presence of a few scarcely recurrent if not unique and multi-token patterns lowers the performance at token level: in fact we have entities of the type *e200067*, *S2352302620301095*, *ehaa254* or multi-token as *10.1148/radiol.2020 200823* in which the black-red alternation indicates the different component tokens.

The category *LOCATION* gets good results at token level but not entity. This behavior is generally due to the presence of several subcategories, such as *CITY*, *COUNTRY*, *HOSPITAL*, *ORGANIZATION* and *OTHER*. If for the entities of type *CITY* and *HOSPITAL* there is a sufficient number of samples more or less distributed between training and test sets, the same cannot be said for the other three categories, mainly present in the test set and with a small number of samples. In addition, for the *CITY* type entities there is an additional disadvantage due to the presence of a certain number of abbreviations, such as *VV*, *CE* and *VR* for *Vibo Valentia*, *Caserta* and *Verona* respectively, which are not very numerous and therefore difficult to recognize. On the other hand, for the HOSPITAL type entities there are tokens that are often repetitive components within the entities, as for example *UOC*, *ASST*, *AO* or *PO* even in the dotted versions, e.g. U.O.C., but the disturbing element is often the presence within the entity, as part of the hospital name, of entities that could also be indicated as *NAME* or LOCATION.

The category *NAME* achieves good results and in practice consists only of the subcategory *DOCTOR*. In this case, despite the token/entity ratio greater than 2, the results at entity level are not very far from those at token level. In fact, there are two recurring patterns: *Name Surname* or *N. Surname*, although in the latter case it may happen to find the entity constituted by a single token *N.Surname* which becomes more difficult to interpret, explaining the lower F_{1E}.

Finally, the *PROFESSION* category has the worst performance: this result is not unexpected as in NER tasks, and in de-identification tasks in particular H. Yang and Garibaldi 2015; Z. Liu, Tang, et al. 2017, it is quite common. This behavior is due to the peculiar-

Table 4.14: Examples of polysemous entities

i2b2 category: subcategory	Extracted sentence
LOCATION: HOSPITAL	Viene ricoverato inizialmente nel **reparto** di Osservazione Breve - Covid. (He was initially admitted to the Short Observation **department** - Covid.)
PROFESSION	presidio Ospedaliero di Vigevano, direttore ff **reparto** di radiologia Elena Belloni (Vigevano Hospital, director ff radiology **department** Elena Belloni)
LOCATION: HOSPITAL	UOC Radiologia Pediatria PO G. Di **Cristina** ARNAS Civico Palermo (UOC Radiology Pediatrics PO G. Di **Cristina** ARNAS Civico Palermo)
NAME: DOCTOR	**Cristina** Veirana, Alessandro Gastaldo UOC Radiologia, Ospedale San Paolo (**Cristina** Veirana, Alessandro Gastaldo UOC Radiology, San Paolo Hospital)

Table 4.15: Cosine similarity between words in Italian FastText embeddings

Word 1	Word 2	Cosine similarity
vibonese	Vibo	0.50009230
lodigiano	Lodi	0.58718747
Veneto	Lombardia	0.62249140

ity of this category: the professions are various and hardly recurrent in the medical records if used as a descriptive part of the personal information of patients, as for example *dipendente di un albergo* (hotel employee) and *medico di continuità assistenziale* (continuity of care doctor). On the other hand, to describe the roles in hospital facilities, if the medical records are rather sectorial as in this case, it is possible to find recurrent entities, such as *Direttore* (Director), which are always recognized.

4.2.1 Qualitative analysis

The Bi-LSTM + CRF model with the proposed stacked embedding made by FastText plus Flair works both at the sub-word level and at the character level exploiting the context: the results show that this proposed stacked embedding is particularly effective in improving the ability to detect and classify entities.

The presence of Flair embedding and its ability to work at character level allow the identification of a series of entities that FastText embedding alone is not able to detect, such as DOCTOR type entities where the surname is attached to the pointed name, such as *U.Burgio*, *M.Castiglia*, *L.Ferraro*, *M.Finazzo*, *G.Marsala*, *L.Putignano* and *A.Re*. Instead, the ability to exploit polysemy

and context is effective both when entities are multi-token hence difficult to identify like HOSPITAL type entities such as *reparto di Osservazione Breve* (Short Observation Department), *U.O.C. di Malattie Infettive* (Infectious Disease Complex Operating Unit), *PO G. Di Cristina* (PO G. Di Cristina) and when entities are in foreign language, hence unusual, but mentioned in a specific context like DOCTOR type entities such as *Wang*, *Ruchong* and *Chunli*. In a similar way these capabilities make it easy to identify URL type entities such as *https://doi.org/10.3760/cma.j.cn112147-20200217-00106*, *https://doi.org/10.2214/ajr.20.22954* and *https://doi.org/10.1148/radiol.2020200823*. Some examples of polysemous entities are reported in Table 4.14.

The use of sub-word level embedding, such as FastText, allows to identify semantically similar entities. In fact FastText embedding, unlike Flair one, is able to identify entities like *vibonese* and *lodigiano*: these are other ways to indicate the provinces of *Vibo Valentia* (often recurring as *Vibo*) and *Lodi* respectively and, although these entities are never seen before, their semantic similarities at sub-word level allow the system to recognize them. Similarly the entity *Veneto* when introduced by the term *regione* (region) is correctly recognized: in the training data set there is a similar introductory formula for another region, i.e. *Lombardia*. For the sake of completeness, the cosine similarity between entities are reported in Table 4.15.

The combination of Flair and FastText embeddings, despite the contextual capabilities, is not always able to recognize single token entities. Some examples in this sense are given by the entities *domenica*, *April* and *March* of type DATE, or by the entity *Reggio* of type CITY, as well as numerous DOCTOR type entities belonging to foreign but particularly short names such as *Han*, *Shi*, *Cao*, *Pan* and *Sun*.

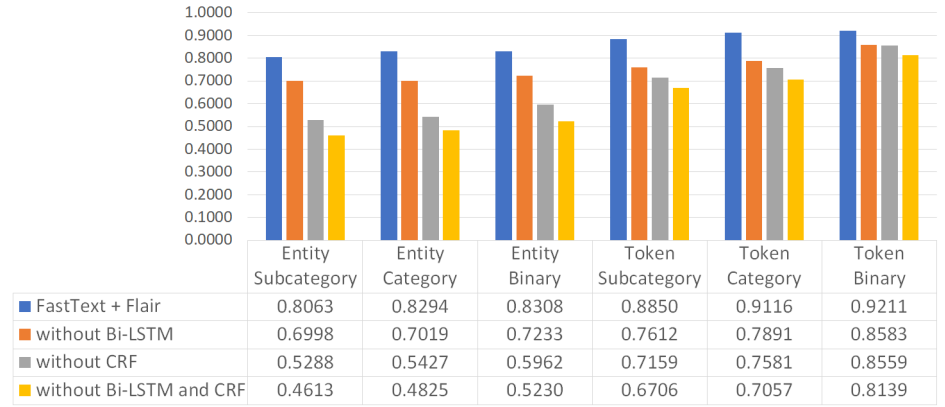
It is interesting to note that there are some entities that are detected by the Bi-LSTM + CRF model with FastText plus Flair embeddings but not by $\text{BERT}_{\text{BASE}}$ Cased model, and this is probably due to a different work at character and sub-word level: for example we have the *118* entity of type PHONE, or the *SOC Radiodiagnostica* (complex radiodiagnostic operating structure) entity of type HOSPITAL only partially detected by $\text{BERT}_{\text{BASE}}$ Cased model.

Some significant examples of challenging entities for all the models have been reported in Table 4.16. In some cases, as for the entities

Table 4.16: Examples of Unidentified Entities; in blue are identified the entities belonging to LOCATION category whereas in red the ones belonging to PROFESSION category.

[...] La Pz viene assistita e trattata in MU (reparto dedicato ai pazienti COVID-19) e dopo circa 1 settimana dimostra notevoli miglioramenti [...] [...] The patient is assisted and treated in emergency medicine (department dedicated to COVID-19) patients and after about 1 week shows significant improvements) [...]
[...] in accordo con i colleghi clinici del Pronto Soccorso con attribuzione di uno score radiologico per quantificare l'estensione di malattia. [...] [...] in agreement with clinical colleagues in the ER with the attribution of a radiological score to quantify the extent of the disease. [...]
[...] esami ematochimici con PCR lievemente aumentata. Mamma dipendente di industria chimica con casi positivi al tampone naso - faringeo. [...] [...] blood chemistry tests with slightly increased C-reactive protein. Mother employee of chemical industry with positive cases of nose - pharyngeal swab. [...]
[...] Giunge al PS di Serra San Bruno (VV) per riferita febbre (da almeno 5 giorni) [...] [...] He arrives at the emergency room of Serra San Bruno (VV) for reported fever (for at least 5 days) [...]

of type PROFESSION *clinici* (clinics) and *dipendente di industria chimica* (chemical industry employee) the recognition is difficult due to the lack of examples in the training data set combined with ambiguities and complex patterns respectively. On the other hand, the HOSPITAL entity *reparto dedicato ai pazienti COVID-19* (ward dedicated to COVID-19 patients) is rather ambiguous and annotated in a questionable way, therefore difficult to identify. Finally, among CITY entities, it remains very difficult to recognize *VV* which is an



	Entity Subcategory	Entity Category	Entity Binary	Token Subcategory	Token Category	Token Binary
FastText + Flair	0.8063	0.8294	0.8308	0.8850	0.9116	0.9211
without Bi-LSTM	0.6998	0.7019	0.7233	0.7612	0.7891	0.8583
without CRF	0.5288	0.5427	0.5962	0.7159	0.7581	0.8559
without Bi-LSTM and CRF	0.4613	0.4825	0.5230	0.6706	0.7057	0.8139

Figure 4.4: Ablation analysis.

abbreviation, albeit present in an extended form and with capitalized initials within the context.

Entities of type *COUNTRY*, such as *Italy*, *Inghilterra* and *China* are not recognized by any system because of the lack of representativeness and disparities within data sets: in the training system we find only *Italia* of type *COUNTRY*.

4.2.2 Ablation analysis

The ablation analysis allows to understand the weight of the main components of a system within a given scenario (Dernoncourt, J. Y. Lee, Uzuner, et al. 2017; Akbik, Blythe, and Vollgraf 2018). Here it can be seen which layer makes the greatest contribution to clinical de-identification in a low-resource language scenario with a small data set. In the specific it goes to compare a baseline, constituted by the best model that is the BiLSTM + CRF with FastText plus Flair embedding, with three ablated models: one will not have the CRF layer, the second one will have a simple Feed Forward layer instead of the BiLSTM layer and the third one without both CRF layer and Bi-LSTM layer (substituted by the Feed Forward one).

When the Bi-LSTM layer is replaced by a linear Feed Forward layer, i.e. when a multinomial logistic regression (Menard 2002) is applied, then the label prediction is obtained as $P(\mathbf{y}_t = j|\mathbf{h}_t) = softmax(\mathbf{h}_t)[j]$ where the hidden layer $\mathbf{h}_t$ is equal to $\mathbf{W}_h\mathbf{x}_t + \mathbf{b}_h$.

This analysis allows to highlight two key aspects for this particular scenario:

- the combination of a BiLSTM layer and a CRF layer always achieves better performance than the individual layers;
- as the level of classification difficulty increases, it is possible to better distinguish the contributions of the different layers: in fact, if the CRF layer and the BiLSTM layer seem to have almost the same weight in a binary token scenario, the difference in favor of the model with the CRF layer becomes more evident proceeding towards the entity subcategory scenario.

To sum up, if on the one hand each removal has resulted in a marked reduction in performance suggesting that the choices made to assemble the analyzed architecture are correct, on the other hand

it is possible to underline that, unlike what previously proposed by the scientific literature, it is not sufficient to conduct such a study limiting itself to the binary token layer as it could obtain misleading indications on the performance of the different layers composing the model.

4.3 Third use case

The Micro-Averaged F_1 results are shown in Table 4.17. In particular, column *Model* indicates the trained model, while column *Strategy* indicates the training strategy adopted. On the other columns, as explained in Section 3.4, there are six evaluation critera: S, C and B represent respectively *i2b2 subcategory*, *i2b2 category*, *binary*, while the subscripts E and T stand for *entity* and *token*.

Table 4.17: Micro-Averaged F_1 results

Model	Strategy	S_E	C_E	B_E	S_T	C_T	B_T
Bi-LSTM+CRF: BPEmb (IT) + Flair (IT)	IT	0.8110	0.8278	0.8317	0.8856	0.9115	0.9190
Bi-LSTM+CRF: MultiBPEmb + Flair multi fast	EN	0.2662	0.2948	0.3134	0.4103	0.4914	0.5797
	IT	0.7910	0.8118	0.8159	0.8826	0.9060	0.9183
	MIX	0.8371	**0.8602**	0.8618	0.8970	0.9304	0.9417
	EN-IT	**0.8391**	0.8595	0.8619	**0.9033**	0.9321	**0.9449**
$BERT_{base}$ (IT) Cased	IT	0.7553	0.7880	0.8561	0.7969	0.8979	0.9260
mBERT Cased	EN	0.4585	0.5029	0.6878	0.5498	0.6097	0.6878
	IT	0.7768	0.8207	**0.9449**	0.8923	**0.9353**	**0.9449**
	MIX	0.7696	0.8105	0.9379	0.8833	0.9245	0.9379
	EN-IT	0.7228	0.7576	0.8969	0.8241	0.8678	0.8969

In particular, two pre-trained Italian language models were used as baselines: the *Bi-LSTM + CRF: BPEmb (IT) + Flair (IT)* and the *$BERT_{base}$ (IT) Cased* models. In these cases the only possible training strategy involves the exclusive use of the Italian training set. Observing the results it is possible to understand how it is feasible to obtain better performance by using strategies based on transfer learning approaches: in this way it is easy to increase the training set by using data available in languages with high resources such as English.

Furthermore, the results further confirm what (Devlin et al. 2019) have already expressed in the literature: although the results at token level suggest the use of BERT-based architectures for the NER task, this assumption is actually misleading. It is important to remember that the NER, hence de-identification as the basis of the

anonymisation process, should be evaluated at the level of multi-class entities, i.e. category and subcategory. In all other scenarios, in fact, entities could be replaced by the wrong surrogates, which would leave ample room for re-identification (Vincze and Farkas 2014). As a result, it is possible to consider the model Bi-LSTM + CRF with *MultiBPEmb + Flair multi fast* stacked embedding trained with strategy *EN-IT* as the most suitable for the clinical de-identification in a low-resources scenario such as that of the Italian language. Hereinafter, the *Bi-LSTM + CRF: BPEmb (IT) + Flair (IT)* will be referred to as monolingual system, while the *Bi-LSTM + CRF: MultiBPEmb + Flair multi fast* model trained with MIX or EN-IT strategies as crosslingual systems.

4.3.1 Embeddings ablation analysis

For the sake of completeness, it was analysed the importance of each embedding type within the EN-IT crosslingual system. Results are reported in Table 4.18.

As can be easily seen from the results, neither Flair alone nor MultiBPEmb alone can achieve results comparable to their combination: exploiting a contextual model that works at character level proves to be a less performing choice compared to the use of a subword model in the case of a low-resources language. But the considerable detachment that is obtained by combining the two different embedding suggests that, in a clinical de-identification task such the one under analysis, the use of a subword model that can also exploit contextuality is particularly effective.

Table 4.18: Micro-Averaged F_1 results for embeddings ablation analysis of the EN-IT crosslingual system.

Embedding	S_E	C_E	B_E	S_T	C_T	B_T
MultiBPEmb	0.7614	0.7743	0.7914	0.8201	0.8569	0.8835
Flair multi fast	0.7621	0.7851	0.7972	0.8529	0.8801	0.8963
MultiBPEmb + Flair multi fast	**0.8391**	**0.8595**	**0.8619**	**0.9033**	**0.9321**	**0.9449**

4.3.2 Embeddings space analysis

The Figure 4.5 reports an embedding scatter plot obtained applying a 3D Principal Component Analysis (PCA) on the original embedding space representing the Original-Embedding (MultiBPEmb + Flair multi fast) on the first row and the Reprojected-Embedding after the double training strategy (EN-IT) on the second row. On the other hand, the first column presents the tokens related to English sentences whereas the second column the ones related to Italian sentences.

The two data sets used for the plots are:

- the Italian SIRM COVID-19 test set composed by 1185 sentences;
- the first 1185 sentences of the English i2b2 2014 training set.

This choice was made in order to have about the same data points for both English and Italian scenarios.

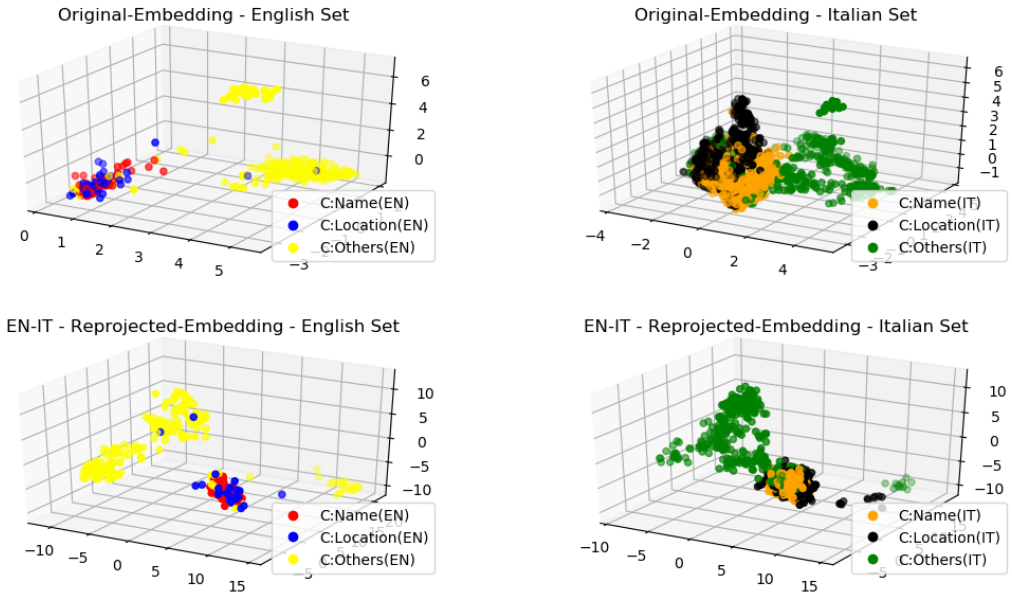


Figure 4.5: Scatter plots of three dimensional principal component analysis of the embedding points.

For the sake of clarity, only the two most represented categories were considered, *C:Name* and *C:Location*, respectively the red and blue points for the English set and the orange and black points for

the Italian set. All other categories are labeled with *C:Other*, using the colors yellow (EN) and green (IT).

The scatter plots highlight two major insights:

1. *Column view*: the training process has generated a redistribution of clusters on the reprojected embedding space depending on the NER task adapting their own position on the analyzed categories;

2. *Row view*: the better alignment of embeddings clusters considering the relative position of the clusters related to each category: in fact, in the reprojected embedding space, the clusters of the same category remain in the same area for both languages.

4.3.3 Considerations

The results obtained allow to identify what may be the best approaches to manage clinical de-identification using NER systems for Italian language. First of all it is possible to notice that the use of the crosslingual system trained in English and tested in Italian does not obtain exciting results, on the contrary it obtains worse results than a monolingual system with training and testing in Italian.

In detail, this study allows to highlight that, even if used in a scenario of limited resources such as that of the Italian language, crosslingual systems properly used can obtain better results than monolingual systems provided some caution during training, so as to take full advantage of the beneficial effects due to the transfer learning. In fact, crosslingual systems that are trained with a mixed English - Italian data set or with a double training first in English then in Italian, can obtain better results than monolingual systems. Moreover, this study shows that it is slightly preferable to adopt a strategy with double training, rather than single training with a mixed data set: this finding leads to think that in the first case the ”noise” introduced within the network is more limited, favoring a better settlement of the weights of the neural network.

Along with these aspects, it is important to add another crucial consideration: the world of research today is strongly interconnected, which is why it is increasingly common to come across bibliometric references, often in English or transliterated Chinese, within medical

records written in any other language. Hence there is a phenomenon sometimes similar to code-switching, although references do not constitute expressions in different languages within the same expression, but rather sentences disconnected from the rest of the discourse and reported as notes at the bottom of the page. For these reasons, crosslingual systems can, in addition, succeed in obtaining superior performance, at least in terms of recognition of entities in languages other than the target language. An example can be the entity *State Administration of Traditional Chinese Medicine*, written in English within a predominantly Italian text and correctly recognized only by the crosslingual systems as LOCATION: ORGANIZATION.

4.3.4 Strengths and weaknesses: monolingual vs crosslingual systems

In order to clarify the advantages and disadvantages of the best crosslingual system, i.e. trained with EN-IT strategy, compared to the monolingual system, a comparative analysis of the entities surveyed is proposed below. For simplicity, the monolingual system will be indicated by the abbreviation IT while the crosslingual system with EN-IT training strategy will be indicated by the abbreviation EN-IT.

Entities analysis

Table 4.19 shows the entities that are correctly recognized only by one system, IT or EN-IT. The output of the tokenization process is emphasized by the alternation of black and red colors.

First of all, the only type of case study in which the IT system has an advantage over the EN-IT system: it refers to all those situations in which there are multi-token entities in the target language that are complex and specific. An example is given by the entity of type LOCATION: HOSPITAL *Unità di terapia intensiva* "Intensive Care Unit": the IT system correctly recognizes the entity, instead the EN-IT system succeeds in a random way because of the complexity and peculiarity of the entity, which neither presents the same number of tokens as the English correspondent (4 vs 3) nor has the same roots for all the words in the other training language (*terapia* vs "care"). On the other hand, the EN-IT crosslingual system has a number

of advantages in different scenarios, which can be grouped in three cases.

The first case concerns all those entities in languages other than the target language, but present because they may be quotes. A frequent example is the one given by foreign names, English or Chinese, such as the entity *Liu, Bin* of type NAME: DOCTOR. Even if the Beginning or Inside of the entity is not correct, maybe because of an unusual pattern compared to the Italian, i.e. *Surname, Name*, it is possible to get better results at token level. Here the crosslingual is clearly superior to the monolingual approach.

The second case, on the other hand, concerns those entities which generally belong to the LOCATION category. What can be identified is a higher accuracy, especially finer-grained and therefore at subcategory level, of the EN-IT system than the IT system. In detail, some entities of type LOCATION: CITY or LOCATION: OTHER as *Milano*, *Marcianise*, *Vibo Valentia*, *Wuhan* and *Veneto* are generally correctly identified by the EN-IT system, instead with wrong subcategories or unseen by the IT system. The motivation is probably to be found in the ability of the crosslingual system to rely to a greater extent on contextual patterns derived also from the English language that suggest the presence of an entity of type LOCATION: CITY. Instead, the IT system tends to identify them as LOCATION: HOSPITAL: this error is induced by the fact that in the Italian language the names of cities or places are often used also to give the name to the hospital that oversees the city or place.

Finally, the third case considers those entities of type AGE, CONTACT: PHONE, DATE which, although not present in large numbers, are expressed through recurrent patterns also in other languages such as English: some examples can be *47aa* "47yo", *118*, *12.02.2020*.

4.3.5 Challenging entities

In this section the focus is on the entities that are difficult to identify for both systems, with the aim of providing some explanation. As already mentioned by (Dernoncourt, J. Y. Lee, Uzuner, et al. 2017), the main sources of error are generally due to (1) abbreviations, whose brevity and variety contribute to confuse the learning system, (2) ambiguities, due to polysemic tokens used in unclear contexts,

Table 4.19: Examples of recognized entities. The alternation of black and red words is used to emphasize the output of the tokenization process.

i2b2 Category: Subcategory	**Entity**	**Recognized by**
AGE	47aa	EN-IT
CONTACT: PHONE	118	EN-IT
DATE	12.02.2020	EN-IT
LOCATION: CITY	Milano	EN-IT
	Marcianise	EN-IT
	Vibo Valentia	EN-IT
	Wuhan	EN-IT
LOCATION: HOSPITAL	Unità di terapia intensiva (*intensive care unit*)	IT
LOCATION: OTHER	Veneto	EN-IT
NAME: DOCTOR	Liu, Bin	EN-IT

(3) debatable annotations, i.e. annotation errors, shortcomings or variations with respect to the guidelines and (4) both scarcity and sparsity of certain types of entities within the data sets. These error sources have been indicated by the abbreviations AB, AMB, D and S, respectively, and used in the *Motivation* column of Table 4.20. In detail, this table shows the entities that are most difficult to identify and the alternation of the colors black and red indicates how tokenization works.

Some examples are LOCATION: CITY entities such as *Fabrizia* or *Melito*, scarcely present, or LOCATION: CITY abbreviations widely present as *VV* and *CE* to indicate the cities of Vibo Valentia and Caserta respectively, or LOCATION: OTHER entities such as *vibonese* and *lodigiano*, which represent unusual ways of identifying the provinces Vibo Valentia and Lodi.

Moreover, the systems under analysis are not able to successfully identify those complexly structured entities such as *reparto dedicato ai pazienti COVID-19* "COVID-19 patient department" or *HUB di riferimento Covid* "Covid Reference HUB", labeled as LOCATION: HOSPITAL but not predicted in any way, probably because of the too ambiguous way of identifying specific places without even using capital letters.

While ID: ID NUMBER or CONTACT: URL entities such as *10.1186/s40779-020-00240-0* and *http://yzs.satcm.gov.cn/zhengcewenjian/2020-02-19/13221.html*

Table 4.20: Challenging entities. The alternation of black and red words is used to emphasize the output of the tokenization process.

i2b2 Category: Subcategory	**Entity**	**Motivation**
CONTACT: URL	http://yzs.satcm.gov.cn/zhengcewen-jian/2020-02-19/13221.html	AMB
DATE	domenica (*sunday*)	S
ID: ID NUMBER	10.1186/s40779-020-00240-0	AMB
LOCATION: CITY	Fabrizia	S
	Melito	S
	VV	AB
	CE	AB
LOCATION: HOSPITAL	reparto dedicato ai pazienti COVID-19 (*COVID-19 patients department)*	AMB, D
	HUB di riferimento Covid (*Covid reference HUB*)	AMB, D
LOCATION: OTHER	vibonese	AMB, S
	lodigiano	AMB, S
PROFESSION	clinici (*clinician*)	S
	dipendente di industria chimica (*chemical industry employee*)	S
	medico di Pronto Soccorso (*Emergency Room medical doctor*)	S

that the tokenizer tends to break into several sub-tokens are never correctly recognized by either system and, in addition, also ambiguities contribute to lower the score: for example, within the second entity it might be easy to confuse the *2020-02-19* part with an entity of type DATE.

Furthermore, those entities of type PROFESSION, such as *clinici* "clinicians", *dipendente di industria chimica* "chemical industry employee" or *medico di Pronto Soccorso* "Emergency Room doctor", are not detected by the systems because of the scarcity, as the number of entities in training is too small.

Likewise, entities that do not recur in the training set but that also present a completely different morphology such as *domenica* "sunday" (type DATE) are not detected at all.

Chapter 5

Conclusions

The purpose of this final chapter is to summarise what was seen in the previous chapters and to summarise the most important results obtained in the various proposed use cases.

Firstly, a novel approach to clinical de-identification through NER was proposed. The evaluation on the i2b2 2014 de-identification data set showed that leveraging the proposed method is possible to obtain results that are on par or outperform the state of the art without any feature engineering or the use of handcrafted rules, which indicates the validity of the model. As main results, the proposed de-identification system has achieved the highest Micro-Averaged scores of F_1 of 94.80%, 95.79%, 96.14% and 96.78%, 97.92%, 98.32% at entity level and token level respectively for i2b2 subcategory, i2b2 category and binary recognition with $SGF = 32$ and in detail it establishes a new state of the art at the category level which is the main evaluation method for de-identification aiming to replace entities with surrogates for anonymisation purposes. The main limitations of this research work concern (1) the limited validation scope due to the restricted diffusion of other de-identification data sets and (2) both the scarce memorisation capacity and the lack of parallelism of the Bi-LSTM + CRF architecture. All the results leave a further room for improvement: in future works it will be possible to both create other de-identification data sets of the same dimensions on which the method proposed will be validated and experiment other architectures, such as those based on Transformers, to manage long documents. In addition, the impact of the re-projection of pre-trained embeddings as input to the Bi-LSTM + CRF architecture will be deeply investigated, by considering more data

sets and addressing other NER problems. Furthermore, clustering techniques, such as the ones proposed in (Abualigah, Khader, and Hanandeh 2018; Abualigah 2019), will be investigated to facilitate and improve the annotation process of entities for creating novel de-identification data sets as well as to extend the set of features useful for the sequence labelling and the classification of PHI. Finally, similarly also to other Artificial Intelligence-based applications (Xiaohui Yang et al. 2020), it will be possible to investigate how to combine big unstructured medical records with incremental learning in order to continuously refine the performance of the proposed system in real-world scenarios.

Secondly, a novel Italian data set was proposed for clinical de-identification. This data set was created from the COVID-19 medical records made available by the Italian Society of Radiology. It was labelled by three Italian native speakers and assessed by using two different indexes with a substantial agreement between them. Moreover, a Bi-LSTM + CRF architecture in combination with a stacked embedding composed by FastText embedding plus Flair (forward and backward) embeddings was tested for clinical de-identification, on the proposed Italian data set. Furthermore, another state-of-the-art architecture, i.e. $\text{BERT}_{\text{BASE}}$, was tested leveraging the Italian models made available by the MDZ Digital Library team at the Bavarian State Library. The Bi-LSTM + CRF architecture with the stacked embedding obtained the best results among the others. These results showed that it is desirable to adopt both contextualised and character-level language models in combination with sub-word embeddings: this way the system is capable to capture, on the one hand, the polysemy of words, their morpho-syntactic variations, rare words and/or misspelled ones and, on the other hand, the latent semantic and syntactic similarities. In the future it might be interesting to compare other Italian versions of BERT or existing language models to see which ones are best suited for a clinical de-identification scenario and to assess if they can outperform the combination of the Bi-LSTM + CRF architecture with Italian FastText plus Flair stacked embedding herein tested.

Thirdly, two cutting-edge NER architectures, Bi-LSTM + CRF and BERT, suitable for de-identification, were analysed in order to understand their behaviour on COVID-19 medical records with respect to a low-resource language scenario like the Italian one. For

this purpose, both English i2b2 2014 and SIRM COVID-19 de-identification data sets were used. Additionally, four strategies were tested to pinpoint the best to apply in this particular context. Performed tests showed that the best strategy to adopt was a double training, before in English then in Italian, exploiting a Bi-LSTM + CRF architecture in combination with MultiBPEmb and Flair Multilingual Fast embeddings. The results obtained leave further room for improvement, although they have allowed to highlight how, in this situation, it is desirable to proceed with clinical de-identification given the low-resources language problem. An interesting future development could be the comparison of different architectures even among those not available for multilingual purposes, to understand if at the moment the results obtained are the best possible. The real limitation of this research area remains the size of the data sets available for clinical de-identification: it would be appropriate to increase the availability of de-identification data sets of the same size as the English i2b2 2014, so as to allow a fair comparison with monolingual systems and provide strong baselines of reference before attempting necessary approaches to low resources case studies.

Further studies of interest could include the development of models capable of automatically choosing the best knowledge transfer strategy based on the input data and the task to be performed, or capable of standardizing the different possible input PHI into a predefined high-level set whatever the set of input labels, as well as testing and then extending the techniques employed to datasets from other areas.

Bibliography

Abualigah, Laith Mohammad (2019). *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering*. Vol. 816. Studies in Computational Intelligence. Springer. ISBN: 978-3-030-10673-7. DOI: 10.1007/978-3-030-10674-4. URL: https://doi.org/10.1007/978-3-030-10674-4.

Abualigah, Laith Mohammad, Ahamad Tajudin Khader, and Essam Said Hanandeh (2018). "Hybrid clustering analysis using improved krill herd algorithm". In: *Appl. Intell.* 48.11, pp. 4047–4071. DOI: 10.1007/s10489-018-1190-6. URL: https://doi.org/10.1007/s10489-018-1190-6.

Agic, Zeljko and Ivan Vulic (2019). "JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, pp. 3204–3210. DOI: 10.18653/v1/p19-1310. URL: https://doi.org/10.18653/v1/p19-1310.

Ahmad, Wasi Uddin, Zhisong Zhang, Xuezhe Ma, Eduard H. Hovy, Kai-Wei Chang, and Nanyun Peng (2019). "On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing". In: *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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, pp. 2440–2452. DOI: 10.18653/v1/n19-1253. URL: https://doi.org/10.18653/v1/n19-1253.

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf (2019). "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP". In: *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, Demonstrations*. Ed. by Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh. Association for Computational Linguistics, pp. 54–59. DOI: 10.18653/v1/n19-4010. URL: https://doi.org/10.18653/v1/n19-4010.

Akbik, Alan, Tanja Bergmann, and Roland Vollgraf (2019). "Pooled Contextualized Embeddings for Named Entity Recognition". In: *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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, pp. 724–728. DOI: 10.18653/v1/n19-1078. URL: https://doi.org/10.18653/v1/n19-1078.

Akbik, Alan, Duncan Blythe, and Roland Vollgraf (2018). "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Association for Computational Linguistics, pp. 1638–1649. URL: https://www.aclweb.org/anthology/C18-1139/.

Alfalahi, Alyaa, Sara Brissman, and Hercules Dalianis (2012). "Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus". In: *LREC 2012, May 23-24-25, 2012, Istanbul, Turkey*.

Alsentzer, Emily, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott (June 2019). "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://www.aclweb.org/anthology/W19-1909.

Arkhipov, Mikhail, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin (2019). "Tuning Multilingual Transformers for Lan-

guage-Specific Named Entity Recognition". In: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics. DOI: 10.18653/v1/w19-3712. URL: https://doi.org/10.18653/v1/w19-3712.

Arora, Abhinav, Akshat Shrivastava, Mrinal Mohit, Lorena Sainz-Maza Lecanda, and Ahmed Aly (2020). "Cross-lingual Transfer Learning for Intent Detection of Covid-19 Utterances". In:

Aroyo, Lora and Chris Welty (2015). "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation". In: *AI Mag.* 36.1, pp. 15–24. DOI: 10.1609/aimag.v36i1.2564. URL: https://doi.org/10.1609/aimag.v36i1.2564.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2018). "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, pp. 789–798. DOI: 10.18653/v1/P18-1073. URL: https://www.aclweb.org/anthology/P18-1073/.

Beckwith, Bruce, Rajeshwarri Mahaadevan, Ulysses J. Balis, and Frank Kuo (2006). "Development and evaluation of an open source software tool for deidentification of pathology reports". In: *BMC Medical Informatics Decis. Mak.* 6, p. 12. DOI: 10.1186/1472-6947-6-12. URL: https://doi.org/10.1186/1472-6947-6-12.

Bingel, Joachim and Johannes Bjerva (2018). "Cross-lingual complex word identification with multitask learning". In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications@NAACL-HLT 2018, New Orleans, LA, USA, June 5, 2018*. Ed. by Joel R. Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis. Association for Computational Linguistics, pp. 166–174. DOI: 10.18653/v1/w18-0518. URL: https://doi.org/10.18653/v1/w18-0518.

Bobicev, Victoria and Marina Sokolova (2017). "Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*. Ed. by Ruslan Mitkov and Galia Angelova. INCOMA Ltd., pp. 97–102. DOI: 10.26615/978-954-

452-049-6_015. URL: https://doi.org/10.26615/978-954-452-049-6_015.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information". In: *Trans. Assoc. Comput. Linguistics* 5, pp. 135–146. URL: https://transacl.org/ojs/index.php/tacl/article/view/999.

Borowik, Piotr, Piotr Brylicki, Mariusz Dzieciatko, Waldemar Jeda, Lukasz Leszewski, and Piotr Zajac (2019). "De-Identification of Electronic Health Records Data". In: *Information Technology in Biomedicine, ITIB 2019, Kamień Śląski, Poland, 18-20 June, 2019*. Ed. by Ewa Pietka, Pawel Badura, Jacek Kawa, and Wojciech Wieclawek. Vol. 1011. Advances in Intelligent Systems and Computing. Springer, pp. 325–337. DOI: 10.1007/978-3-030-23762-2_29. URL: https://doi.org/10.1007/978-3-030-23762-2_29.

Chen, Tao, Richard M. Cullen, and Marshall Godwin (2015). "Hidden Markov model using Dirichlet process for de-identification". In: *J. Biomed. Informatics* 58, S60–S66. DOI: 10.1016/j.jbi.2015.09.004. URL: https://doi.org/10.1016/j.jbi.2015.09.004.

Chiu, Jason P. C. and Eric Nichols (2016). "Named Entity Recognition with Bidirectional LSTM-CNNs". In: *Trans. Assoc. Comput. Linguistics* 4, pp. 357–370. URL: https://transacl.org/ojs/index.php/tacl/article/view/792.

Chklovski, Timothy A. and Rada Mihalcea (2003). "Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation". In: *Recent Advances in Natural Language Processing (RANLP) Conference, 2003, Borovetz, Bulgaria*.

Conneau, Alexis and Guillaume Lample (2019). "Cross-lingual Language Model Pretraining". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, pp. 7057–7067. URL: http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.

Coombs, Crispin R. (2020). "Will COVID-19 be the tipping point for the Intelligent Automation of work? A review of the debate and implications for research". In: *Int. J. Inf. Manag.* 55, p. 102182.

DOI: 10.1016/j.ijinfomgt.2020.102182. URL: https://doi.org/10.1016/j.ijinfomgt.2020.102182.

Dehghan, Azad, Aleksandar Kovacevic, George Karystianis, John A. Keane, and Goran Nenadic (2015). "Combining knowledge- and data-driven methods for de-identification of clinical narratives". In: *J. Biomed. Informatics* 58, S53–S59. DOI: 10.1016/j.jbi.2015.06.029. URL: https://doi.org/10.1016/j.jbi.2015.06.029.

Dernoncourt, Franck, Ji Young Lee, and Peter Szolovits (2017). "NeuroNER: an easy-to-use program for named-entity recognition based on neural networks". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 - System Demonstrations*. Ed. by Lucia Specia, Matt Post, and Michael Paul. Association for Computational Linguistics, pp. 97–102. DOI: 10.18653/v1/d17-2017. URL: https://doi.org/10.18653/v1/d17-2017.

Dernoncourt, Franck, Ji Young Lee, Özlem Uzuner, and Peter Szolovits (2017). "De-identification of patient notes with recurrent neural networks". In: *J. Am. Medical Informatics Assoc.* 24.3, pp. 596–606. DOI: 10.1093/jamia/ocw156. URL: https://doi.org/10.1093/jamia/ocw156.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/n19-1423. URL: https://doi.org/10.18653/v1/n19-1423.

Eddy, Sean R (1996). "Hidden Markov models". In: *Current Opinion in Structural Biology* 6.3, pp. 361–365.

Elman, Jeffrey L. (1990). "Finding Structure in Time". In: *Cogn. Sci.* 14.2, pp. 179–211. DOI: 10.1207/s15516709cog1402_1. URL: https://doi.org/10.1207/s15516709cog1402_1.

Forney, G.D. (1973). "The viterbi algorithm". In: *Proceedings of the IEEE* 61.3, pp. 268–278. DOI: 10.1109/proc.1973.9030. URL: https://doi.org/10.1109/proc.1973.9030.

Freund, Yoav and Robert E. Schapire (1995). "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Computational Learning Theory, Second European Conference, EuroCOLT '95, Barcelona, Spain, March 13-15, 1995, Proceedings*. Ed. by Paul M. B. Vitányi. Vol. 904. Lecture Notes in Computer Science. Springer, pp. 23–37. DOI: 10.1007/3-540-59119-2_166. URL: https://doi.org/10.1007/3-540-59119-2_166.

Friedlin, F. Jeff and Clement J. McDonald (2008). "Application of Information Technology: A Software Tool for Removing Patient Identifying Information from Clinical Documents". In: *J. Am. Medical Informatics Assoc.* 15.5, pp. 601–610. DOI: 10.1197/jamia.M2702. URL: https://doi.org/10.1197/jamia.M2702.

Gaudet-Blavignac, Christophe, Vasiliki Foufi, Eric Wehrli, and Christian Lovis (Sept. 2018). "De-identification of French medical narratives". In: *Swiss Medical Informatics*. DOI: 10.4414/smi.34.00417. URL: https://doi.org/10.4414/smi.34.00417.

Giorgi, John M. and Gary D. Bader (2020). "Towards reliable named entity recognition in the biomedical domain". In: *Bioinform.* 36.1, pp. 280–286. DOI: 10.1093/bioinformatics/btz504. URL: https://doi.org/10.1093/bioinformatics/btz504.

Goller, Christoph and Andreas Küchler (1996). "Learning task-dependent distributed representations by backpropagation through structure". In: *Proceedings of International Conference on Neural Networks (ICNN'96), Washington, DC, USA, June 3-6, 1996*. IEEE, pp. 347–352. DOI: 10.1109/ICNN.1996.548916. URL: https://doi.org/10.1109/ICNN.1996.548916.

Goodman, Leo A. and William H. Kruskal (1979). "Measures of Association for Cross Classifications". In: *Measures of Association for Cross Classifications*. Springer New York, pp. 2–34. DOI: 10.1007/978-1-4612-9995-0_1. URL: https://doi.org/10.1007/978-1-4612-9995-0_1.

Graves, Alex and Jürgen Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural Networks* 18.5-6, pp. 602–610. DOI: 10.1016/j.neunet.2005.06.042. URL: https://doi.org/10.1016/j.neunet.2005.06.042.

Grouin, Cyril and Aurélie Névéol (2014). "De-identification of clinical notes in French: towards a protocol for reference corpus development". In: *J. Biomed. Informatics* 50, pp. 151–161. DOI:

10.1016/j.jbi.2013.12.014. URL: https://doi.org/10.1016/j.jbi.2013.12.014.

Grouin, Cyril, Arnaud Rosier, Olivier Dameron, and Pierre Zweigenbaum (2009). "Testing Tactics to Localize De-Identification". In: *Medical Informatics in a United and Healthy Europe - Proceedings of MIE 2009, The XXIInd International Congress of the European Federation for Medical Informatics, Sarajevo, Bosnia and Herzegovina, August 30 - September 2, 2009*. Ed. by Klaus-Peter Adlassnig, Bernd Blobel, John Mantas, and Izet Masic. Vol. 150. Studies in Health Technology and Informatics. IOS Press, pp. 735–739. DOI: 10.3233/978-1-60750-044-5-735. URL: https://doi.org/10.3233/978-1-60750-044-5-735.

Gui, Tao, Jiacheng Ye, Qi Zhang, Yaqian Zhou, Yeyun Gong, and Xuanjing Huang (2020). "Leveraging Document-Level Label Consistency for Named Entity Recognition". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. Ed. by Christian Bessiere. ijcai.org, pp. 3976–3982. DOI: 10.24963/ijcai.2020/550. URL: https://doi.org/10.24963/ijcai.2020/550.

Guillen, R et al. (2006). "Automated de-identification and categorization of medical records". In: *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*. Vol. 116.

Gupta, Dilip, Melissa Saul, and John Gilbertson (Feb. 2004). "Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research". In: *American Journal of Clinical Pathology* 121.2, pp. 176–186. DOI: 10.1309/e6k33gbpe5c27fyu. URL: https://doi.org/10.1309/e6k33gbpe5c27fyu.

Hazarika, Barenya Bikash and Deepak Gupta (2020). "Modelling and forecasting of COVID-19 spread using wavelet-coupled random vector functional link networks". In: *Appl. Soft Comput.* 96, p. 106626. DOI: 10.1016/j.asoc.2020.106626. URL: https://doi.org/10.1016/j.asoc.2020.106626.

He, Bin, Yi Guan, Jianyi Cheng, Keting Cen, and Wenlan Hua (2015). "CRFs based de-identification of medical records". In: *J. Biomed. Informatics* 58, S39–S46. DOI: 10.1016/j.jbi.2015.08.012. URL: https://doi.org/10.1016/j.jbi.2015.08.012.

Hearst, Marti A. (1998). "Trends & Controversies: Support Vector Machines". In: *IEEE Intell. Syst.* 13.4, pp. 18–28. DOI: 10.1109/5254.708428. URL: https://doi.org/10.1109/5254.708428.

Heinzerling, Benjamin and Michael Strube (2018). "BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2018/summaries/1049.html.

Heinzerling, Benjamin and Michael Strube (2019). "Sequence Tagging with Contextual and Non-Contextual Subword Representations: A Multilingual Evaluation". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, pp. 273–291. DOI: 10.18653/v1/p19-1027. URL: https://doi.org/10.18653/v1/p19-1027.

Hendrycks, Dan and Kevin Gimpel (2016). "Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units". In: *CoRR* abs/1606.08415. arXiv: 1606.08415. URL: http://arxiv.org/abs/1606.08415.

Hernandez-Matamoros, Andres, Hamido Fujita, Toshitaka Hayashi, and Héctor M. Pérez Meana (2020). "Forecasting of COVID19 per regions using ARIMA models and polynomial functions". In: *Appl. Soft Comput.* 96, p. 106610. DOI: 10.1016/j.asoc.2020.106610. URL: https://doi.org/10.1016/j.asoc.2020.106610.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Howard, Jeremy and Sebastian Ruder (2018). "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Lin-*

guistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, pp. 328–339. DOI: 10.18653/v1/P18-1031. URL: https://www.aclweb.org/anthology/P18-1031/.

Hu, Anwen, Zhicheng Dou, Jian-Yun Nie, and Ji-Rong Wen (2020). "Leveraging Multi-Token Entities in Document-Level Named Entity Recognition". 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*. AAAI Press, pp. 7961–7968. URL: https://aaai.org/ojs/index.php/AAAI/article/view/6304.

Huang, Zhiheng, Wei Xu, and Kai Yu (2015). "Bidirectional LSTM-CRF Models for Sequence Tagging". In: *CoRR* abs/1508.01991. arXiv: 1508.01991. URL: http://arxiv.org/abs/1508.01991.

Hvingelby, Rasmus, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard (2020). "DaNE: A Named Entity Resource for Danish". In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association, pp. 4597–4604. URL: https://www.aclweb.org/anthology/2020.lrec-1.565/.

Jiang, Min, Todd Sanger, and Xiong Liu (Nov. 2019). "Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study". In: *JMIR Medical Informatics* 7.4, e14850. DOI: 10.2196/14850. URL: https://doi.org/10.2196/14850.

Johnson, Andrew, Penny Karanasou, Judith Gaspers, and Dietrich Klakow (2019). "Cross-lingual Transfer Learning for Japanese Named Entity Recognition". In: *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 2*

(Industry Papers). Ed. by Anastassia Loukina, Michelle Morales, and Rohit Kumar. Association for Computational Linguistics, pp. 182–189. DOI: 10.18653/v1/n19-2023. URL: https://doi.org/10.18653/v1/n19-2023.

K, Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth (2020). "Cross-Lingual Ability of Multilingual BERT: An Empirical Study". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: https://openreview.net/forum?id=HJeT3yrtDr.

Kalyan, Katikapalli Subramanyam and Sivanesan Sangeetha (2020). "SECNLP: A survey of embeddings in clinical natural language processing". In: *J. Biomed. Informatics* 101, p. 103323. DOI: 10.1016/j.jbi.2019.103323. URL: https://doi.org/10.1016/j.jbi.2019.103323.

Khandelwal, Urvashi, He He, Peng Qi, and Dan Jurafsky (2018). "Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, pp. 284–294. DOI: 10.18653/v1/P18-1027. URL: https://www.aclweb.org/anthology/P18-1027/.

Kim, Joo-Kyung, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier (2017). "Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics, pp. 2832–2838. DOI: 10.18653/v1/d17-1302. URL: https://doi.org/10.18653/v1/d17-1302.

Kim, Youngjun, Paul M. Heider, and Stéphane M. Meystre (2018). "Ensemble-based Methods to Improve De-identification of Electronic Health Record Narratives". In: *AMIA 2018, American Medical Informatics Association Annual Symposium, San Francisco, CA, November 3-7, 2018*. AMIA. URL: http://knowledge.amia.org/67852-amia-1.4259402/t004-1.4263758/t004-1.4263759/2976309-1.4263922/2975300-1.4263919.

Kitaev, Nikita and Dan Klein (2018). "Constituency Parsing with a Self-Attentive Encoder". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, pp. 2676–2686. DOI: 10.18653/v1/P18-1249. URL: https://www.aclweb.org/anthology/P18-1249/.

Kırbaş, İsmail, Adnan Sözen, Azim Doğuş Tuncer, and Fikret Şinasi Kazancıoğlu (Sept. 2020). "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches". In: *Chaos, Solitons & Fractals* 138, p. 110015. DOI: 10.1016/j.chaos.2020.110015. URL: https://doi.org/10.1016/j.chaos.2020.110015.

Krippendorff, Klaus (1980). *Content analysis: an introduction to its methodology*.

Kudo, Taku and John Richardson (2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Eduardo Blanco and Wei Lu. Association for Computational Linguistics, pp. 66–71. DOI: 10.18653/v1/d18-2012. URL: https://doi.org/10.18653/v1/d18-2012.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*. Ed. by Carla E. Brodley and Andrea Pohoreckyj Danyluk. Morgan Kaufmann, pp. 282–289.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (2016). "Neural Architectures for Named Entity Recognition". In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. The Association for Computational Linguistics, pp. 260–270. DOI: 10.18653/v1/n16-1030. URL: https://doi.org/10.18653/v1/n16-1030.

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2018). "Word translation without parallel data". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=H196sainb.

Landis, J. Richard and Gary G. Koch (Mar. 1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, p. 159. DOI: 10.2307/2529310. URL: https://doi.org/10.2307/2529310.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang (2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinform.* 36.4, pp. 1234–1240. DOI: 10.1093/bioinformatics/btz682. URL: https://doi.org/10.1093/bioinformatics/btz682.

Li, J., A. Sun, J. Han, and C. Li (2020). "A Survey on Deep Learning for Named Entity Recognition". In: *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1. DOI: 10.1109/TKDE.2020.2981314.

Liu, Changjian, Jiaming Li, Yuhan Liu, Jiachen Du, Buzhou Tang, and Ruifeng Xu (2019). "Named Entity Recognition in Clinical Text Based on Capsule-LSTM for Privacy Protection". In: *Artificial Intelligence and Mobile Services - AIMS 2019 - 8th International Conference, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25-30, 2019, Proceedings*. Ed. by De Wang and Liang-Jie Zhang. Vol. 11516. Lecture Notes in Computer Science. Springer, pp. 166–178. DOI: 10.1007/978-3-030-23367-9_12. URL: https://doi.org/10.1007/978-3-030-23367-9_12.

Liu, Peter J., Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer (2018). "Generating Wikipedia by Summarizing Long Sequences". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=Hyg0vbWC-.

Liu, Zengjian, Yangxin Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Haodi Li, Jingfeng Wang, Qiwen Deng, and Suisong Zhu

(2015). "Automatic de-identification of electronic medical records using token-level and character-level conditional random fields". In: *J. Biomed. Informatics* 58, S47–S52. DOI: 10.1016/j.jbi.2015.06.009. URL: https://doi.org/10.1016/j.jbi.2015.06.009.

Liu, Zengjian, Buzhou Tang, Xiaolong Wang, and Qingcai Chen (Nov. 2017). "De-identification of clinical notes via recurrent neural network and conditional random field". In: *Journal of Biomedical Informatics* 75, S34–S42. DOI: 10.1016/j.jbi.2017.05.023. URL: https://doi.org/10.1016/j.jbi.2017.05.023.

Luo, Ling, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang (2018). "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition". In: *Bioinform.* 34.8, pp. 1381–1388. DOI: 10.1093/bioinformatics/btx761. URL: https://doi.org/10.1093/bioinformatics/btx761.

Luo, Ying, Fengshun Xiao, and Hai Zhao (2020). "Hierarchical Contextualized Representation for Named Entity Recognition". 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*. AAAI Press, pp. 8441–8448. URL: https://aaai.org/ojs/index.php/AAAI/article/view/6363.

Ma, Xuezhe and Eduard H. Hovy (2016). "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. DOI: 10.18653/v1/p16-1101. URL: https://doi.org/10.18653/v1/p16-1101.

Mamede, Nuno J., Jorge Baptista, and Francisco Dias (2016). "Automated anonymization of text documents". In: *IEEE Congress on Evolutionary Computation, CEC 2016, Vancouver, BC, Canada, July 24-29, 2016*. IEEE, pp. 1287–1294. DOI: 10.1109/CEC.2016.7743936. URL: https://doi.org/10.1109/CEC.2016.7743936.

Marciniak, M., A. Mykowiecka, and P. Rychlik (2010). "Medical text data anonymization". In: *Journal of Medical Informatics and Technologies* 16.

Marimon, Montserrat, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger (2019). "Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results". In: *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*. Ed. by Miguel Ángel García Cumbreras, Julio Gonzalo, Eugenio Martínez Cámara, Raquel Martínez-Unanue, Paolo Rosso, Jorge Carrillo-de-Albornoz, Soto Montalvo, Luis Chiruzzo, Sandra Collovini, Yoan Gutiérrez, Salud M. Jiménez Zafra, Martin Krallinger, Manuel Montes-y-Gómez, Reynier Ortega-Bueno, and Aiala Rosá. Vol. 2421. CEUR Workshop Proceedings. CEUR-WS.org, pp. 618–638. URL: http://ceur-ws.org/Vol-2421/MEDDOCAN_overview.pdf.

Marques, Gonçalo, Deevyankar Agarwal, and Isabel de la Torre Díez (2020). "Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network". In: *Appl. Soft Comput.* 96, p. 106691. DOI: 10.1016/j.asoc.2020.106691. URL: https://doi.org/10.1016/j.asoc.2020.106691.

Mayhew, Stephen, Tatiana Tsygankova, and Dan Roth (2019). "ner and pos when nothing is capitalized". 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 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computational Linguistics, pp. 6255–6260. DOI: 10.18653/v1/D19-1650. URL: https://doi.org/10.18653/v1/D19-1650.

Mehrabi, Ninareh, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan (2020). "Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition". In: *HT '20: 31st ACM Conference on Hypertext and Social Media, Virtual Event, USA, July 13-15, 2020*. Ed. by Ujwal Gadiraju. ACM, pp. 231–232. DOI: 10.1145/3372923.3404804. URL: https://doi.org/10.1145/3372923.3404804.

Menard, Scott (2002). *Applied Logistic Regression Analysis*. SAGE Publications, Inc. DOI: 10.4135/9781412983433. URL: https://doi.org/10.4135/9781412983433.

Meystre, Stephane M, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore (Aug. 2010). "Automatic de-identification of textual documents in the electronic health record: a review of recent research". In: *BMC Medical Research Methodology* 10.1. DOI: 10.1186/1471-2288-10-70. URL: https://doi.org/10.1186/1471-2288-10-70.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *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*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

Mohamadou, Youssoufa, Aminou Halidou, and Pascalin Tiam Kapen (2020). "A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19". In: *Appl. Intell.* 50.11, pp. 3913–3925. DOI: 10.1007/s10489-020-01770-9. URL: https://doi.org/10.1007/s10489-020-01770-9.

Mohammad, Shahnawaz Khan, Mustafa, and Yannis (Apr. 2020). "An Artificial Intelligence and NLP based Islamic FinTech Model Combining Zakat and Qardh-Al-Hasan for Countering the Adverse Impact of COVID 19 on SMEs and Individuals". In: *International Journal of Economics and Business Administration* VIII.Issue 2, pp. 351–364. DOI: 10.35808/ijeba/466. URL: https://doi.org/10.35808/ijeba/466.

Mulcaire, Phoebe, Jungo Kasai, and Noah A. Smith (2019). "Polyglot Contextual Representations Improve Crosslingual Transfer". In: *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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Asso-

ciation for Computational Linguistics, pp. 3912–3918. DOI: 10.18653/v1/n19-1392. URL: https://doi.org/10.18653/v1/n19-1392.

Nadeau, David and Satoshi Sekine (2009). "A survey of named entity recognition and classification". In: *Benjamins Current Topics*. John Benjamins Publishing Company, pp. 3–28. DOI: 10.1075/bct.19.03nad. URL: https://doi.org/10.1075/bct.19.03nad.

Neamatullah, Ishna, Margaret M. Douglass, Li-Wei H. Lehman, Andrew T. Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford (2008). "Automated de-identification of free-text medical records". In: *BMC Medical Informatics Decis. Mak.* 8, p. 32. DOI: 10.1186/1472-6947-8-32. URL: https://doi.org/10.1186/1472-6947-8-32.

Neuraz, Antoine, Ivan Lerner, William Digan, Nicolas Paris, Rosy Tsopra, Alice Rogier, David Baudoin, Kevin Bretonnel Cohen, Anita Burgun, Nicolas Garcelon, and Bastien Rance and (Aug. 2020). "Natural Language Processing for Rapid Response to Emergent Diseases: Case Study of Calcium Channel Blockers and Hypertension in the COVID-19 Pandemic". In: *Journal of Medical Internet Research* 22.8, e20773. DOI: 10.2196/20773. URL: https://doi.org/10.2196/20773.

Pan, Sinno Jialin and Qiang Yang (2010). "A Survey on Transfer Learning". In: *IEEE Trans. Knowl. Data Eng.* 22.10, pp. 1345–1359. DOI: 10.1109/TKDE.2009.191. URL: https://doi.org/10.1109/TKDE.2009.191.

Pantazos, Kostas, Søren Lauesen, and Søren Lippert (2017). "Preserving medical correctness, readability and consistency in de-identified health records". In: *Health Informatics J.* 23.4, pp. 291–303. DOI: 10.1177/1460458216647760. URL: https://doi.org/10.1177/1460458216647760.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "Glove: Global Vectors for Word Representation". In: *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*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. ACL, pp. 1532–1543. DOI: 10.3115/v1/d14-1162. URL: https://doi.org/10.3115/v1/d14-1162.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. Ed. by Marilyn A. Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/n18-1202. URL: https://doi.org/10.18653/v1/n18-1202.

Peters, Matthew E., Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih (2018). "Dissecting Contextual Word Embeddings: Architecture and Representation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Association for Computational Linguistics, pp. 1499–1509. DOI: 10.18653/v1/d18-1179. URL: https://doi.org/10.18653/v1/d18-1179.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, pp. 4996–5001. DOI: 10.18653/v1/p19-1493. URL: https://doi.org/10.18653/v1/p19-1493.

Plank, Barbara, Dirk Hovy, and Anders Søgaard (2014). "Learning part-of-speech taggers with inter-annotator agreement loss". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*. Ed. by Gosse Bouma and Yannick Parmentier. The Association for Computer Linguistics, pp. 742–751. DOI: 10.3115/v1/e14-1078. URL: https://doi.org/10.3115/v1/e14-1078.

Rabiner, L.R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2, pp. 257–286. DOI: 10.1109/5.18626. URL: https://doi.org/10.1109/5.18626.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). "Improving language understanding by generative pre-training". In:

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language models are unsupervised multitask learners". In: *OpenAI Blog* 1.8, p. 9.

Ramshaw, Lance A. and Mitch Marcus (1995). "Text Chunking using Transformation-Based Learning". In: *Third Workshop on Very Large Corpora, VLC@ACL 1995, Cambridge, Massachusetts, USA, June 30, 1995*. Ed. by David Yarowsky and Kenneth Church. URL: https://www.aclweb.org/anthology/W95-0107/.

Richter-Pechanski, Phillip, Stefan Riezler, and Christoph Dieterich (2018). "De-Identification of German Medical Admission Notes". In: *German Medical Data Sciences: A Learning Healthcare System - Proceedings of the 63rd Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology (GMDS e.V.) 2018 in Osnabrück, Germany, 2-6 September 2018, GMDS 2018*. Ed. by Ursula Hübner, Ulrich Sax, Hans-Ulrich Prokosch, Bernhard Breil, Harald Binder, Antonia Zapf, Brigitte Strahwald, Tim Beißbarth, Niels Grabe, and Anke Schöler. Vol. 253. Studies in Health Technology and Informatics. IOS Press, pp. 165–169. DOI: 10.3233/978-1-61499-896-9-165. URL: https://doi.org/10.3233/978-1-61499-896-9-165.

Røst, Thomas Brox, Laura Slaughter, Øystein Nytrø, Ashley Elizabeth Muller, and Gunn Vist (2020). "Using Deep Learning to Support High-Quality Covid-19 Evidence Mapping". In:

Ruder, Sebastian, Ivan Vulic, and Anders Søgaard (2019). "A Survey of Cross-lingual Word Embedding Models". In: *J. Artif. Intell. Res.* 65, pp. 569–631. DOI: 10.1613/jair.1.11640. URL: https://doi.org/10.1613/jair.1.11640.

Sahin, Gözde Gül, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych (2020). "LINSPECTOR: Multilingual Probing Tasks for Word Representations". In: *Comput. Linguistics* 46.2, pp. 335–385. DOI: 10.1162/coli_a_00376. URL: https://doi.org/10.1162/coli_a_00376.

Santos, Breno Santana, Ivanovitch Silva, Marcel da Câmara Ribeiro-Dantas, Gisliany Alves, Patricia Takako Endo, and Luciana Lima (Oct. 2020). "COVID-19: A scholarly production dataset report for research analysis". In: *Data in Brief* 32, p. 106178. DOI: 10.

1016/j.dib.2020.106178. URL: https://doi.org/10.1016/j.dib.2020.106178.

Scheurwegs, Elyne, Kim Luyckx, Filip Van der Schueren, and Tim Van den Bulcke (2013). "De-Identification of Clinical Free Text in Dutch with Limited Training Data: A Case Study". In: *Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013, Hissar, Bulgaria, September 13, 2013*. Ed. by Guergana Savova, Kevin Bretonnel Cohen, and Galia Angelova. INCOMA, pp. 18–23. URL: https://www.aclweb.org/anthology/W13-5103/.

Schuster, Mike and Kaisuke Nakajima (2012). "Japanese and Korean voice search". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*. IEEE, pp. 5149–5152. DOI: 10.1109/ICASSP.2012.6289079. URL: https://doi.org/10.1109/ICASSP.2012.6289079.

Schuster, Tal, Ori Ram, Regina Barzilay, and Amir Globerson (2019). "Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing". In: *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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, pp. 1599–1613. DOI: 10.18653/v1/n19-1162. URL: https://doi.org/10.18653/v1/n19-1162.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. DOI: 10.18653/v1/p16-1162. URL: https://doi.org/10.18653/v1/p16-1162.

Shakil, Mohammad Hassan, Ziaul Haque Munim, Mashiyat Tasnia, and Shahin Sarowar (Nov. 2020). "COVID-19 and the environment: A critical review and research agenda". In: *Science of The Total Environment* 745, p. 141022. DOI: 10.1016/j.scitotenv.2020.141022. URL: https://doi.org/10.1016/j.scitotenv.2020.141022.

Si, Yuqi, Jingqi Wang, Hua Xu, and Kirk E. Roberts (2019). "Enhancing clinical concept extraction with contextual embeddings". In: *J. Am. Medical Informatics Assoc.* 26.11, pp. 1297–1304. DOI: 10.1093/jamia/ocz096. URL: https://doi.org/10.1093/jamia/ocz096.

Strubell, Emma, Patrick Verga, David Belanger, and Andrew McCallum (2017). "Fast and Accurate Entity Recognition with Iterated Dilated Convolutions". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics, pp. 2670–2680. DOI: 10.18653/v1/d17-1283. URL: https://doi.org/10.18653/v1/d17-1283.

Stubbs, Amber, Christopher Kotfila, and Özlem Uzuner (2015). "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1". In: *J. Biomed. Informatics* 58, S11–S19. DOI: 10.1016/j.jbi.2015.06.007. URL: https://doi.org/10.1016/j.jbi.2015.06.007.

Stubbs, Amber and Özlem Uzuner (2015). "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus". In: *J. Biomed. Informatics* 58, S20–S29. DOI: 10.1016/j.jbi.2015.07.020. URL: https://doi.org/10.1016/j.jbi.2015.07.020.

Suri, Jasjit S., Anudeep Puvvula, Mainak Biswas, Misha Majhail, Luca Saba, Gavino Faa, Inder M. Singh, Ronald Oberleitner, Monika Turk, Paramjit S. Chadha, Amer M. Johri, J. Miguel Sanches, Narendra N. Khanna, Klaudija Viskovic, Sophie Mavrogeni, John R. Laird, Gyan Pareek, Martin Miner, and Subbaram Naidu (2020). "COVID-19 pathways for brain and heart injury in comorbidity patients: A role of medical imaging and artificial intelligence-based COVID severity classification: A review". In: *Comput. Biol. Medicine* 124, p. 103960. DOI: 10.1016/j.compbiomed.2020.103960. URL: https://doi.org/10.1016/j.compbiomed.2020.103960.

Sweeney, Latanya (1996). "Replacing personally-identifying information in medical records, the Scrub system." In: *Proceedings : a conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*, pp. 333–337.

Szarvas, György, Richárd Farkas, and Róbert Busa-Fekete (2007). "Research Paper: State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework". In: *J. Am. Medical Informatics Assoc.* 14.5, pp. 574–580. DOI: 10.1197/jamia.M2441. URL: https://doi.org/10.1197/jamia.M2441.

Tang, Buzhou, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu (2013). "Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features". In: *BMC Medical Informatics Decis. Mak.* 13.S-1, S1. DOI: 10.1186/1472-6947-13-S1-S1. URL: https://doi.org/10.1186/1472-6947-13-S1-S1.

Tang, Buzhou, Dehuan Jiang, Qingcai Chen, Xiaolong Wang, Jun Yan, and Ying Shen (2019). "De-identification of Clinical Text via Bi-LSTM-CRF with Neural Language Models". In: *AMIA 2019, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 16-20, 2019*. AMIA. URL: http://knowledge.amia.org/69862-amia-1.4570936/t004-1.4574923/t004-1.4574924/3203046-1.4574964/3201562-1.4574961.

Taylor, Wilson L. (Sept. 1953). ""Cloze Procedure": A New Tool for Measuring Readability". In: *Journalism Quarterly* 30.4, pp. 415–433. DOI: 10.1177/107769905303000401. URL: https://doi.org/10.1177/107769905303000401.

Thomas, Sean M., Burke W. Mamlin, Gunther Schadow, and Clement J. McDonald (2002). "A successful technique for removing names in pathology reports using an augmented search and replace method". In: *AMIA 2002, American Medical Informatics Association Annual Symposium, San Antonio, TX, USA, November 9-13, 2002*. AMIA. URL: http://knowledge.amia.org/amia-55142-a2002a-1.610020/t-001-1.612667/f-001-1.612668/a-156-1.612780/a-157-1.612777.

Tomanek, Katrin, Philipp Daumke, Frank Enders, Jens Huber, Katharina Theres, and Marcel Müller (2012). "An interactive de-identification-system". In: *Proceedings of SMBM*, pp. 82–86.

Trienes, Jan, Dolf Trieschnigg, Christin Seifert, and Djoerd Hiemstra (2020). "Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records". In: *Proceedings of the ACM WSDM 2020 Health Search and Data Mining Workshop, co-located with the 13th ACM International WSDM Conference, HSDM@WSDM 2020, Houston, TX, USA,*

February 3, 2020. Ed. by Carsten Eickhoff, Yubin Kim, and Ryen W. White. Vol. 2551. CEUR Workshop Proceedings. CEUR-WS.org, pp. 3–11. URL: http://ceur-ws.org/Vol-2551/paper-03.pdf.

Tu, Karen, Julie Klein-Geltink, Tezeta F. Mitiku, Chiriac Mihai, and Joel Martin (2010). "De-identification of primary care electronic medical records free-text data in Ontario, Canada". In: *BMC Medical Informatics Decis. Mak.* 10, p. 35. DOI: 10.1186/1472-6947-10-35. URL: https://doi.org/10.1186/1472-6947-10-35.

Tveit, Amund, Ole Edsberg, TB Rost, Arild Faxvaag, O Nytro, T Nordgard, Martin Thorsen Ranang, and Anders Grimsmo (2004). "Anonymization of general practioner medical records". In: *second HelsIT Conference*.

Vaishya, Raju, Mohd Javaid, Ibrahim Haleem Khan, and Abid Haleem (July 2020). "Artificial Intelligence (AI) applications for COVID-19 pandemic". In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14.4, pp. 337–339. DOI: 10.1016/j.dsx.2020.04.012. URL: https://doi.org/10.1016/j.dsx.2020.04.012.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Velupillai, Sumithra, Hercules Dalianis, Martin Hassel, and Gunnar H. Nilsson (2009). "Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial". In: *Int. J. Medical Informatics* 78.12, pp. 19–26. DOI: 10.1016/j.ijmedinf.2009.04.005. URL: https://doi.org/10.1016/j.ijmedinf.2009.04.005.

Vincze, Veronika and Richárd Farkas (2014). "De-identification in natural language processing". In: *37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014, Opatija, Croatia, May 26-30,*

2014. IEEE, pp. 1300–1303. DOI: 10.1109/MIPRO.2014.6859768. URL: https://doi.org/10.1109/MIPRO.2014.6859768.

Wellner, Ben, Matt Huyck, Scott A. Mardis, John S. Aberdeen, Alexander A. Morgan, Leonid Peshkin, Alexander S. Yeh, Janet Hitzeman, and Lynette Hirschman (2007). "Research Paper: Rapidly Retargetable Approaches to De-identification in Medical Records". In: *J. Am. Medical Informatics Assoc.* 14.5, pp. 564–573. DOI: 10.1197/jamia.M2435. URL: https://doi.org/10.1197/jamia.M2435.

Wu, Shijie and Mark Dredze (2019). "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT". 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 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computational Linguistics, pp. 833–844. DOI: 10.18653/v1/D19-1077. URL: https://doi.org/10.18653/v1/D19-1077.

Wu, Yonghui, Min Jiang, Jianbo Lei, and Hua Xu (2015). "Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network". In: *MEDINFO 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics, São Paulo, Brazil, 19-23 August 2015*. Ed. by Indra Neil Sarkar, Andrew Georgiou, and Paulo Mazzoncini de Azevedo Marques. Vol. 216. Studies in Health Technology and Informatics. IOS Press, pp. 624–628. DOI: 10.3233/978-1-61499-564-7-624. URL: https://doi.org/10.3233/978-1-61499-564-7-624.

Wu, Yonghui, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu (2017). "Clinical Named Entity Recognition Using Deep Learning Models". In: *AMIA 2017, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 4-8, 2017*. AMIA. URL: http://knowledge.amia.org/65881-amia-1.3897810/t003-1.3901461/f003-1.3901462/2730946-1.3901506/2731659-1.3901503.

Wu, Yonghui, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu (2015). "A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text". In: *AMIA 2015, American Medical Informatics Association Annual Symposium, San Francisco, CA, USA, November 14-18, 2015*. AMIA. URL: http://

knowledge.amia.org/59310-amia-1.2741865/t004-1.2745466/f004-1.2745467/2249008-1.2745489/2248738-1.2745486.

Wu, Yonghui, Xi Yang, Jiang Bian, Yi Guo, Hua Xu, and William R. Hogan (2018). "Combine Factual Medical Knowledge and Distributed Word Representation to Improve Clinical Named Entity Recognition". In: *AMIA 2018, American Medical Informatics Association Annual Symposium, San Francisco, CA, November 3-7, 2018*. AMIA. URL: http://knowledge.amia.org/67852-amia-1.4259402/t004-1.4263758/t004-1.4263759/2977069-1.4263781/2976467-1.4263778.

Xie, Jiateng, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime G. Carbonell (2018). "Neural Cross-lingual Named Entity Recognition with Minimal Resources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Association for Computational Linguistics, pp. 369–379. DOI: 10.18653/v1/d18-1034. URL: https://doi.org/10.18653/v1/d18-1034.

Xue, Hui, Songcan Chen, and Qiang Yang (2008). "Structural Support Vector Machine". In: *Advances in Neural Networks - ISNN 2008, 5th International Symposium on Neural Networks, ISNN 2008, Beijing, China, September 24-28, 2008, Proceedings, Part I*. Ed. by Fuchun Sun, Jianwei Zhang, Ying Tan, Jinde Cao, and Wen Yu. Vol. 5263. Lecture Notes in Computer Science. Springer, pp. 501–511. DOI: 10.1007/978-3-540-87732-5_56. URL: https://doi.org/10.1007/978-3-540-87732-5_56.

Yadav, Vikas and Steven Bethard (2018). "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models". In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Association for Computational Linguistics, pp. 2145–2158. URL: https://www.aclweb.org/anthology/C18-1182/.

Yang, Hui and Jonathan M. Garibaldi (2015). "Automatic detection of protected health information from clinic narratives". In: *J. Biomed. Informatics* 58, S30–S38. DOI: 10.1016/j.jbi.2015.06.015. URL: https://doi.org/10.1016/j.jbi.2015.06.015.

Yang, Xiaohui, Xiaoying Jiang, Chenxi Tian, Pei Wang, Funa Zhou, and Hamido Fujita (2020). "Inverse projection group sparse representation for tumor classification: A low rank variation dictionary approach". In: *Knowl. Based Syst.* 196, p. 105768. DOI: 10.1016/j.knosys.2020.105768. URL: https://doi.org/10.1016/j.knosys.2020.105768.

Yimam, Seid Muhie, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Stajner, Anaïs Tack, and Marcos Zampieri (2018). "A Report on the Complex Word Identification Shared Task 2018". In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications@NAACL-HLT 2018, New Orleans, LA, USA, June 5, 2018*. Ed. by Joel R. Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis. Association for Computational Linguistics, pp. 66–78. DOI: 10.18653/v1/w18-0507. URL: https://doi.org/10.18653/v1/w18-0507.

Zeman, Daniel and Philip Resnik (2008). "Cross-Language Parser Adaptation between Related Languages". In: *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*. The Association for Computer Linguistics, pp. 35–42. URL: https://www.aclweb.org/anthology/I08-3008/.

Zhao, Mengjie and Hinrich Schütze (2019). "A Multilingual BPE Embedding Space for Universal Sentiment Lexicon Induction". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, pp. 3506–3517. DOI: 10.18653/v1/p19-1341. URL: https://doi.org/10.18653/v1/p19-1341.

Zhao, Yue-Shu, Kunli Zhang, Hongchao Ma, and Kun Li (2018). "Leveraging text skeleton for de-identification of electronic medical records". In: *BMC Medical Informatics Decis. Mak.* 18.S-1, pp. 65–72. DOI: 10.1186/s12911-018-0598-6. URL: https://doi.org/10.1186/s12911-018-0598-6.

Zhu, Yi, Ivan Vulic, and Anna Korhonen (2019). "A Systematic Study of Leveraging Subword Information for Learning Word Representations". In: *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). Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, pp. 912–932. DOI: 10.18653/v1/n19-1097. URL: https://doi.org/10.18653/v1/n19-1097.