Università degli Studi di Napoli “Federico II”
Facoltà di Ingegneria

Dottorato di Ricerca in Ingegneria Informatica ed Automatica
XXVIII Ciclo
Dipartimento di Ingegneria Elettrica e delle Tecnologie
dell’Informazione

A Knowledge Multidimensional Representation Model for Automatic Text Analysis and Generation: Applications for Cultural Heritage

Fiammetta Marulli
Ph.D. Thesis

Tutor
Prof. Angelo Chianese

Coordinator
Prof. Francesco Garofalo

February 2016

“I don't care that they stole my idea...I care that they don't have any of their own.”

Nikola Tesla

Abstract

Knowledge is information that has been contextualized in a certain domain, where it can be used and applied. Natural Language provides a most direct way to transfer knowledge at different levels of conceptual density. The opportunity provided by the evolution of the technologies of Natural Language Processing is thus of making more fluid and universal the process of knowledge transfer. Indeed, unfolding domain knowledge is one way to bring to larger audiences contents that would be otherwise restricted to specialists. This has been done so far in a totally manual way through the skills of divulgators and popular science writers. Technology provides now a way to make this transfer both less expensive and more widespread. Extracting knowledge and then generating from it suitably communicable text in natural language are the two related subtasks that need be fulfilled in order to attain the general goal. To this aim, two fields from information technology have achieved the needed maturity and can therefore be effectively combined. In fact, on the one hand Information Extraction and Retrieval (IER) can extract knowledge from texts and map it into a neutral, abstract form, hence liberating it from the stylistic constraints into which it was originated. From there, Natural Language Generation can take charge, by regenerating automatically, or semi-automatically, the extracted knowledge into texts targeting new communities.

This doctoral thesis provides a contribution to making substantial this combination through the definition and implementation of a novel multidimensional model for the representation of conceptual knowledge and of a workflow that can produce strongly customized textual descriptions.

By exploiting techniques for the generation of paraphrases and by profiling target users, applications and domains, a target-driven approach is proposed to automatically generate multiple texts from the same information core. An extended case study is described to demonstrate the effectiveness of the proposed model and approach in the Cultural Heritage application domain, so as to compare and position this contribution within the current state of the art and to outline future directions.

Acknoledgements

I would like to thank my supervisor, Prof. Angelo Chianese, for the great opportunity and the trust he gave me, guiding and suggesting me into following research topics and directions compliant with my personal interestes and skills. He also introduced me into the Cultural Heritage Domain and the Databenc District, giving me several effective opportunities to develop novel ideas and to perform effective tests and experiments.

A special thanks goes also to my co-supervisor and friend, Eng. Paolo Benedusi, from In.Tel.Tec. s.p.a., for his precious input, advices and the opportunity to work by experimenting professional software tools, just introducing me to novel and effective technological skills; for the long discussions we had concerning the problems encountered during my research activities.

I can’t forget to thank my Ph.D. colleagues at DIETI, Doc. Mario Barbareschi, Doc. Roberto Nardone and Doc. Giancarlo Sperlì, for their precious support in solving my student doubts.

A lovely thanks goes to all my friends and to my family for their support and patience all these years. Especially to Luca V., Giancarlo D.G., Augusto S., Antonio C., Aaron V., Stefania L., Stefania G., Luisa C., Mary C., Ciro M., Silvio D., Ivana T., Alessio P., Lilly M., Kira M., Marco M., Patrizia B., Biagina L., who were always there when I needed them, to support me and help me go through many difficulties and disheartening moments, lifting up my spirit when needed.

Last but not the least, I’d like to thank Proff. Mariella Tortorella and Lerina Aversano, from University of Sannio, for the great regard they always had to me and to my research dream.

Finally, thanks to my not yet visible loving friends, always standing with me, in my heart and in my mind.

Preface

Some of the research and results described in this Ph.D. thesis has undergone peer review and has been published in, or at the date of this printing is being considered for publication in, academic journals, books, and conferences. In the following I list all the papers developed during my research work as Ph.D. student.

1. Fiammetta Marulli and Luca Vallifuoco, *The Imitation Game in Cultural Heritage: A Human-like Interaction Driven Approach for Supporting Art Recreation*. InProceedings of 5th EAI International Conference: ArtsIT, Interactivity & Game Creation (ArtsIT2016). Springer, 2016.
2. Paolo Benedusi, Angelo Chianese, Fiammetta Marulli and Francesco Piccialli, *An Associative Engines Based Approach supporting Collaborative Analytics in the Internet of Cultural Things*. In Journal of Future Generation Computer System (FGCS), Elsevier, 2016.
3. Angelo Chianese, Fiammetta Marulli and Francesco Piccialli, *Sensitivity Mining in Social Pulses to address Cultural Heritage Competitive Intelligence*. In International Journal of Knowledge Society Research (IJKSR), IGI Global, 2016. (Submitted in February, 2016, Under Review Process).
4. Angelo Chianese, Fiammetta Marulli and Francesco Piccialli, *A perspective on applications of in-memory and associative approaches supporting Cultural Big Data Analytics*. In International Journal of Computational Science and Engineering (IJCSE), Elsevier, Interscience, 2016. (Submitted in February, 2016, Under Review Process).
5. Fiammetta Marulli, Remo Pareschi and Daniele Baldacci, *The Internet of Speaking Things and its Applications to Cultural Heritage*. In Proceedings of the International Conference on

Internet of Things and Big Data (IOTBD2016), SCITEPRESS, 2016.

6. Angelo Chianese, Fiammetta Marulli and Francesco Piccialli, *Cultural Heritage and Social Pulse: A semantic approach for CH Sensitivity Discovery in social media data*. In Proceedings of 10th International Conference on Semantic Computing (ICSC2016), IEEE, 2016.
7. Paolo Benedusi, Fiammetta Marulli, Adriano Racioppi and Luca F. Ungaro, *What's the matter with Cultural Heritage tweets? An Ontology–based approach for CH Sensitivity Estimation in Social Network Activities*. In Proceedings of 11th International Conference on Signal Image Technology and Internet Based System (SITIS2015), IEEE, 2015.
8. Paolo Benedusi, Angelo Chianese, Fiammetta Marulli and Francesco Piccialli, *An Associative Engines Based Approach supporting Collaborative Analytics in the Internet of Cultural Things*. In Proceedings 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC2015), IEEE, 2015.
9. Fiammetta Marulli, *IoT to enhance understanding of Cultural Heritage: Fedro authoring platform, artworks telling their fables*. In Proceedings of 1st EAI International Conference on Future access enablers of ubiquitous and intelligent infrastructures (FABULOUS2015), Springer, 2015.
10. Francesco Bifulco, Angelo Chianese, Fiammetta Marulli, Francesco Piccialli and Isabella Valente, *La gestione della conoscenza per DATABENC*. Published in the CNR Journal Archeologia e Calcolatori, Supplemento 7 – 2015, pp. 117-129, 2015.
11. Angelo Chianese, Fiammetta Marulli, Vincenzo Moscato and Francesco Piccialli, SmARTweet: *A Location-based smart application for Exhibits and Museums*. In Proceedings of 9th International Conference on Signal Image Technology and Internet Based System (SITIS2013), IEEE, 2013.
12. Angelo Chianese, Fiammetta Marulli, Francesco Piccialli and Isabella Valente, *A novel challenge into Multimedia Cultural*

Heritage: an integrated approach to support cultural information enrichment. In Proceedings of 9th International Conference on Signal Image Technology and Internet Based System (SITIS2013), IEEE, 2013.

13. Angelo Chianese, Fiammetta Marulli, Vincenzo Moscato and Francesco Piccialli, *A "smart" multimedia guide for indoor contextual navigation in Cultural Heritage applications*. In Proceedings of 4th International Conference on Indoor Positioning and Indoor Navigation (IPIN2013), IEEE, 2013.

Contents

List of Figures

List of Tables

Abbreviations

AI	**A**rtificial **I**ntelligence
CH	**C**ultural **H**eritage
IER	**I**nformation **E**xtraction and **R**etrieval
IoT	**I**nternet **o**f **T**hings
LOCH	**L**anguage of **T**hought **H**ypothesis
LOD	**L**inked **O**pen **D**ata
MVC	**M**odel **V**iew **C**ontroller
NLG	**N**atural **L**anguage **G**eneration
NLP	**N**atural **L**anguage **P**rocessing
RTE	**T**extual **E**ntailment **R**ecognition
RTM	**R**epresentational **T**heory of the **M**ind
UGC	**U**ser **C**ontent **G**eneration

Chapter 1

Introduction

Knowledge is not a simple concept to define, and although many definitions have been given of it, only a few describe the concept with enough detail to grasp it in practical terms. Knowledge is information that has been contextualised in a certain domain, to be used or applied. Any piece of knowledge is related with more knowledge in a particular and different way in each individual. Knowledge can have many facets (Ramires, 2012), but it is basically constituted by static components, called concepts or facts, and dynamic components, called skills, abilities, procedures, actions, etc., which together allow general cognition, including all different processes typically associated to it, such as perceiving, distinguishing, abstracting, modelling, storing, recalling, remembering, etc., which are part of three primary cognitive processes: learning, understanding and reasoning. No one of those processes can live isolated or can be carried out alone, actually it can be said that those processes are part of the dynamic knowledge, and dynamic knowledge typically requires of conceptual or factual knowledge to be used.

Knowledge represents the basic core of our Cultural Heritage and Natural Language provides us with prime versatile means of construing experience at multiple levels of organization, storing and exchanging knowledge and information encoded as linguistic meaning. By means of its internal structure and organization, natural language allows us to pass on what we learn about the world from one individual to the other and from one generation to the next.

We can thus observe in scientific texts the construal of domain knowledge by means of enfolding taxonomic relations obtaining between

lexical items; we can likewise observe the relational organisation of a text by which parts of a text make reference to one another, which can be described in terms of conjunctive relations and, on a more abstract scale, rhetorical structure.

Nowadays, the task of generating easily understandable information for people using natural language is being addressed by two fields which, independently until now, have researched the processes this task involves from different perspectives: the natural language generation (NLG) field and the knowledge and information extraction and retrieval (IER) field. The natural language generation field consists in the creation of texts which provide information contained in other kind of sources (numerical data, graphics, taxonomies and ontologies or even other texts), with the aim of making such texts indistinguishable, as far as possible, from those created by humans. On the other hand, the knowledge extraction, basing on text mining and text analysis tasks, as examples of the many applications born from computational linguistic, provides summarization, categorization, topics extractions from textual resources using linguistic concepts, which deal with the imprecision and ambiguity of human language.

Although nowadays in the scientific community there is generally agreement that knowledge about how the world works, or common-sense knowledge is vital for natural language understanding, there is, however, much less agreement or understanding about how to define common-sense knowledge (LoBue, 2012), and what its components are (Feldman, 2002). Likewise, most knowledge extraction systems focus on extracting one specific kind of knowledge from text, often factual relationships, although other specialized extraction techniques exist as well.

Text mining or knowledge discovery is that sub process of data mining, which is widely being used to discover hidden patterns and significant information from the huge amount of unstructured written material. The proliferation of clouds, research and technologies are responsible for the creation of vast volumes of data. This kind of data cannot be used until or unless specific information or pattern is discovered. For this text mining uses techniques of different fields like machine learning, visualization, case-based reasoning, text analysis, database technology statistics, knowledge management, natural language processing and information retrieval. Text mining is largely growing field of computer science simultaneously to big data and artificial intelligence. There are several technology premises for

mining the text. Some of them are represented by summarization, information extraction, Categorization, visualization, clustering, topic tracking, question answering, sentiment and opinion minig. Text mining is a field towards which scientific community interest showed, in the last 10 years, incredibly increased: it became one of the most deeply explored fields, as evidenced by the increasing number of scientific contributions and conferences born in the few years. (Kaushik, 2016) provides a review of text mining techniques, tools and various applications, at current date.

On the other hand, if compared to scientific contributions in text mining and knowledge extraction approaches and techniques, from stochastic-static methods (machine learning based) to rule based approaches, typical for Artificial Intelligence, the linguistic verbalization of segmented data, also known as text generation, is a young field still in its early stages.

It has a solid formal base and but its real potential is still waiting to be uncovered. As reported in (Ramos, 2016), although nowadays there are relevant research results in this domain, most of them (theoretical ones aside) present simple use cases whose application in real problems seems somehow limited, since the complexity of descriptions for real problems in terms of natural language is in general higher than what quantified sentences and the most complex linguistic descriptions currently provide.

Another challing issue, object of recent interest and increasing investigation is Textual Entailment Recognition (RTE). RTE is defined as the capability of a system to recognize that the meaning of a portion of text (usually one or few sentences) entails the meaning of another portion of text. Subsequently, the task has also been extended to recognizing specific cases of non-entailment, as when the meaning of the first text contradicts the meaning of the second text. Although the study of entailment phenomena in natural language was addressedmuch earlier, the novelty of the RTE evaluation was to propose a simple text-to-text task to compare human and system judgments, making it possible to build data sets and to experiment with a variety of approaches. Two main reasons likely contributed to the success of the initiative: First, the possibility to address, for the first time, the complexity of entailment phenomena under a data-driven perspective; second, the text-to-text approach allows one to easily incorporate a textual entailment engine into applications (e.g., question answering, summarization, information extraction) as a core inferential component.

Recognizing textual entailment (RTE) has been proposed as a task in computational linguistics under a successful series of annual evaluation campaigns started in 2005, as evidenced in (Ferro, 2016; Dagan, 2015; Androutsopoulos, 2010). Another task of increasing interest for LNG community is Paraphrasing task. Paraphrasing can be seen as bidirectional textual entailment and methods from the two areas are often very similar. Both kinds of methods are useful, at least in principle, in a wide range of natural language processing applications, including question answering, summarization, text generation, and machine translation (Malakasiotis, 2011).

The problem of automatic production of natural language texts becomes more and more salient with the constantly increasing demand for production of technical documents in multiple languages; intelligent help and tutoring systems which are sensitive to the user's knowledge; and hypertext which adapts according to the user's goals, interests and prior knowledge, as well as to the presentation context. This section will outline the problems, stages and knowledge resources in natural language generation.

Natural Language Generation (NLG) systems produce language output (ranging from a single sentence to an entire document) from computer-accessible data usually encoded in a knowledge or data base. Often the input to a generator is a high-level communicative goal to be achieved by the system (which acts as a speaker or writer). During the generation process, this high-level goal is refined into more concrete goals which give rise to the generated utterance. Consequently, language generation can be regarded as a goal-driven process which aims at adequate communication with the reader/hearer, rather than as a process aimed entirely at the production of linguistically well-formed output. In order to structure the generation task, most existing systems divide it into three main stages, which are often organised in a pipeline architecture: Content Determination, Text Planning, Surface Realization (EAGLES96, 1996). The first and second stages involves, respectively, decisions regarding the information which should be conveyed to the user (content determination) and the way this information should be rhetorically structured (text planning). Many systems perform these tasks simultaneously because often rhetorical goals determine what is relevant. Most text planners have hierarchically-organised plans and apply decomposition in a top-down fashion following AI planning techniques. However, some planning approaches rely on previously selected content - an

assumption which has proved to be inadequate for some tasks (e.g., a flexible explanation facility). Surface realization involves generation of the individual sentences in a grammatically correct manner, e.g., agreement, reflexives, morphology.

In (Androutsopoulos, 2001; Androutsopoulos, 2013), a sophisticated NLG system, for generating multilingual personalized descriptions of museum exhibits is presented. This Natural OWL system verbalizes an OWL domain ontology, exploiting a precompiled lexicon for English and Greek languages, and a flexible grammatics, whose referring expressions can be customized by system users, through a graphical user interface, provided to them. Furthermore, a user-model can be expressed in order to customize the textual output produced, by selecting the facts considered of interest for the target user and in the same way some preferred referencing expressions.
After deeply researching and studying the past and most recent literature in the aforementioned fields, we can conclude that the field concerning text analysis and mining, that is the processing of textual information supporting knowledge extraction is much more investigated, well-assessed and developed, thus providing a wide variety of approaches and solutions, even if many issues are still opened, as the RTE problem, as an example.

Going into the opposite direction, instead, composition of knowledge in structured and well-formed text, it much less investigated and is worth mentioning that there is no agreement in the NLG community on the exact problems addressed in each one of the identified steps of a NLG process, heavily varying among different approaches and systems.

One of the identified bottleneck of these kind of systems and exploited approaches is the lack of a control stategy able to orchestrate and coordinate interventions of available knowledge resources into the steps of processes.

A further aspect not yet included into these type of systems and approaches to NLG is the heavy exploitation of the large amount of information provided directly by users during their web activities. An effective customization for automatically generated texts can be achived only by a massive and effective semantic annotation of knowledge resources exploited into the generation process. Because annotating resources is a time consuming task, requiring not trivial human effort, web users annotated resources, such as folksonomies (Semeraro, 2012), could be exploited to retieve more easily, terms which have been directly used and chosen by users, for categorizing resources. Folksonomies and other types of User

Generated Contents, could be exploited to retrieve more refined characterization of users's way of expressing, which could be reused to generate more customized and users'profiled textual descriptions. In such a perspective, a model of knowledge attempting to unify the all the available knowledge resources, could be very useful in order to well address their exploitation in a text generation workflow. It could be enable the adoption of a strongly target user-profiling and application driven approach, not yet investigated in the typical approaches for automatic linguistic resources treatment.

1.1 Thesis Contribution

To face with these issues, this doctoral thesis shows the research activity conducted with the aim of exploring and scientifically describing knowledge structure and organisation in natural language text, according to different linguistic and semiotic paradigms. It focuses on the importance of linguistic knowledge representation from two perspectives: representation of knowledge by means of natural language as well as explorations and representations of knowledge and information stored in natural language text by means of other formal representations such as ontologies, taxonomies, rhetorical structure etc.

In addition to a thorough investigation of approaches concerned with aspects of knowledge representation, structure and organization, this work is concerned with computational aspects of natural language processing, focusing on computer science and language engineering approaches supporting natural language analysis and generation.

As white light passing through a prism and being split up into the colours of the spectrum, knowledge is composed of multiple diversified dimensions and facets, each exploitable in advanced and machine assisted treatments. Therefore, a novel multidimensional model for the representation of conceptual knowledge, driving a processing workflow for automatic generation of natural language textual resources, is presented. The proposed multidimensional model enhances natural language generation processes, by strongly focusing on diversificate textual generations, based on the same information sources. By exploiting paraphrases generation techniques, a

target-driven approach is proposed and adopted. The “target” term is used in this context to mean target language, target domain, target users and target application exploiting and enjoying textual representations.

In addition, an information system prototype, characterized for Cultural Heritage domain and implementing the aforementioned workflow and approach, is presented. A very extended case study is described to demonstrate the effectiveness of the proposed model and approach. A set of diversificate experiments covering and processing knowledge sources from Cultural Heritage domain, were performed to estimate the obtained results, thus providing the means for comparing and positioning this contribution with current state and future directions.

1.2 Manuscript Reminder

This doctoral thesis is structured as follows:

- Chapter 2 gives an extensive review of basic concepts behind Knowledge Representation and types of knowledge going from traditional theories such as RTM to modern ones such as LOTH and showing not only how each discipline or science, including Philosophy, Psychology, Cognitive Science, Brain Science and Computer Science, has its own approach and limitations.

- Chapter 3 provides a survey on NLG techniques, focusing on an extensive overview on current approaches and open issues. It underlines current state of art, thus introducing the main aim and the problem addressed in this doctoral thesis: the identification of a multidimensional model for knowledge representation, supporting text analysis and natural language generation processes, by adopting a users’ profiling and target applications driven approach;

- Chapter 4 describes the characteristics of the proposed solution, describing the constituting elements of the multidimensional model for knowledge representation and how it is able to support NLG processes;

- Chapter 5 presents the case study by detailing the implementation of an authoring platform, developed for supporting IoT applications in the Cultural Heritage domain, thus evidencing obtained results when compared to those ones obtained employing other NLG system, available fron scientific research community in this field. Obtained results are provided in order to verify the feasibility and the effectiveness of the proposed model and the related approach;

- Chapter 6 concludes this doctoral thesis.

Chapter 2

Knowledge Representation Basic Concepts and Models

In this chapter, a review of the basic concepts behind knowledge representation and the main types of knowledge representation models is presented. Knowledge is not a simple concept to define, and although many definitions have been given of it, only a few describe the concept with enough detail to understand it in practical terms. Knowledge has to be constructed; its construction involves the use of previous knowledge and different cognitive processes, which play an intertwined function to facilitate the development of association between the new concepts to be acquired and previously acquired concepts. Knowledge is about information that can be used or applied, that is, it is information that has been contextualised in a certain domain, and therefore, any piece of knowledge is related with more knowledge in a particular and different way in each individual. Knowledge can have many facets, but it is basically constituted by *static components*, called concepts or facts, and *dynamic components*, called skills, abilities, procedures, actions, which together allow general cognition, including all different processes typically associated to it, such as perceiving, distinguishing, abstracting, modelling, storing, recalling, remembering, etc., which are part of three primary cognitive processes: learning, understanding and reasoning. Actually it can be said that those processes are part of the dynamic knowledge, which typically requires of conceptual or factual knowledge to be used.

2.1 Knowledge: multiple definitions from different sciences

A unified definition for the concept of knowledge is difficult to grasp, diverse definitions from different backgrounds and perspectives have been proposed since the old times; some definitions complement each other and some prove more useful in practical terms. The very first and one of the most accepted definition of knowledge, occurred in philosophy, by Sir Thomas Hobbes in 1651. In his work "Leviathan" (Hobbes, 1651), he stated that knowledge is the evidence of truth, which must have four properties (Hobbes, 1969):

(1) knowledge must be integrated by concepts;
(2) each concept can be identified by a name;
(3) names can be used to create propositions;
(4) such propositions must be concluding.

Hobbes' definition of knowledge was based on the traditional Aristotelian view of ideas, known as the Representational Theory of the Mind (RTM). Till today, most works in Cognitive Science uses RTM, stating that knowledge is defined as the evidence of truth composed by conceptualisations product of the imaginative power of the mind, i.e., cognitive capabilities; ideas here are pictured as objects with mental properties, which is the way most people picture concepts and ideas as abstract objects. In the 70's, Jerry Fodors proposed a complement for RTM at a higher cognitive level by the Language of Thought Hypothesis (LOTH) (Fodors, 1975). LOTH states that thoughts are represented in a language supported by the principles of ***symbolic logic and computability***. This language is different form the one we to use to speak, it is a separate in which we can write our thoughts and we can validate them using symbolic logic.

This definition is much more useful for computer science including Artificial Intelligence and Cognitive Informatics, since it ***implies that reasoning can be formalised into symbols***; hence thought can be described and mechanised, and therefore, theoretically a machine should be able to, at least, emulate thought. More recent than Phylosophy but still directly relevant to knowledge are the branches of Psychology that study the learning

process. In Psychology through more empirical methods, a vast number of theories to understand and interpret human behaviour in relation to knowledge have been developed. Associative theories also referred to as connectionist theories, cognitive theories and constructivist theories (Chomsky, 1967) are among the most relevant theories for knowledge representation systems. Connectionist theories state that knowledge can be described as a number of interconnected concepts, each concept is connected through associations, these are the roots of semantics as means for knowledge representation (Vygotsky, 1986), i.e., what we know today as semantic knowledge representation. Semantic knowledge representation has been proven to be the main driver along with similarity behind reasoning for unstructured knowledge (Crisp-Bright, 2010). Constructivist theories on the other hand do contemplate more complex reasoning drivers such as causality, probability and context. Most constructivist theories therefore complement connectionist approaches by stating that each group of associations integrate different layers of thought where the difference between in each level is the strength of the associations. As a result, the highest layer is the concept, i.e., an organised and stable structure of knowledge and the lowest layer are loosely coupled heaps of ideas (Vygotsky, 1986). This layered structure for knowledge and the way it is built is the reason why constructivism is so relevant to semantic knowledge, because it presents mechanisms complex enough to represent how semantic knowledge is built to our current understanding.

Cognitive Science has focused on modelling and validating previous theories from almost every other science ranging from Biology and Neuroscience to Psychology and Artificial Intelligence (Eysenck, 2010); because of this, Cognitive Science is positioned as the ideal common ground where knowledge definitions from all of the above disciplines can meet computer oriented sciences, this has in fact been argued by Laird in his proposition of mental models (Laird, 1980) though this theory in reasoning rather than in knowledge.

Cognitive Science is therefore a fertile field for new theories or for the formalisation of previous ones through computer models (Marr, 1982). It is common for knowledge in this field to be described through equations, mathematical relations and computer models, for this reason approaches like connectionism in Psychology have been retaken through the modelling of neural networks and similar works (Shastri, 1988). Other famous approaches

in this field include Knowledge Space Theory (Doignon & Falmagne, 1999) which defines knowledge as a group of questions which are combined with possible answers to form knowledge states. Ackoff's (1989) distinction between data information and knowledge is helpful in providing a practical definition for knowledge in real life. Data are symbols without significance, such as numbers, information is data that also includes basic relations between such symbols in a way that provide meaning, and knowledge is context enriched information that can be used or applied, and serves a purpose or goal.

We can conclude this section stating that there are several approaches to describe and define knowledge, most of them coming from different fields. Cognitive Science has served as a common ground for comparing similar issues in the past. Figure 1 shows diffent approaches to Knowledge Representation from different disciplines, as detailed in (Ramirez, 2012).

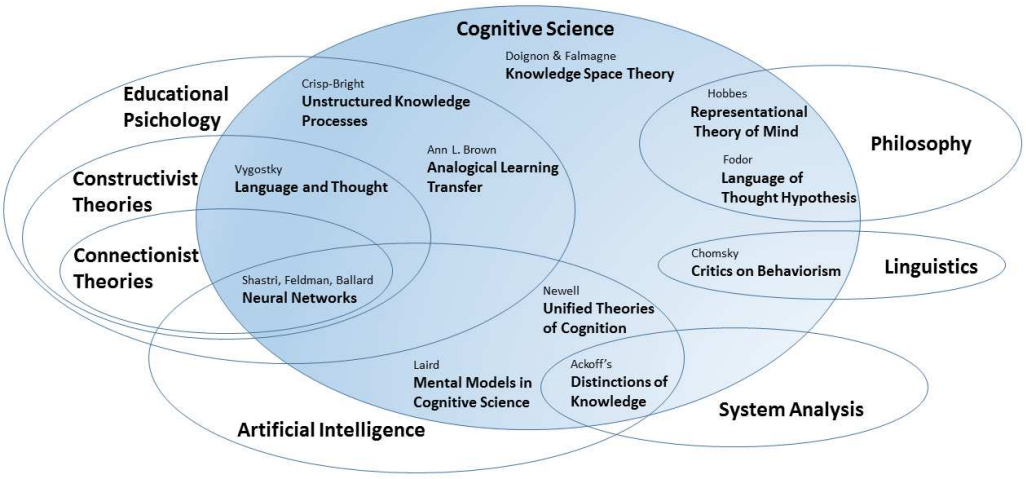

Figure 1. Multiple Approaches to Knowledge Representation from Different Disciplines

2.2 Knowledge in a computer model perspective

Among the multiple definitions provided over the times and by different disciplines, we are interested in a definition of knowledge that can be worked with and used in a computer model. For this reason, our focus is on the elements representing a common ground for knowledge representation. Any system or model for knowledge representation should consider the following:

i. Knowledge is composed of basic units, referred to as *concepts*. The approaches for representing those basic structures will be discussed in the following sections.

ii. Concepts have associations or relations to other concepts. The debate on associations is about the representational aspects regarding to the following issues:

a) What information should an association contain
b) What elements should be used to describe such information i.e., type, directionality, name, intension, extension, among others. These characteristics will be addressed in the following sections

iii. Associations and concepts build dynamic structures which tend to become stable through time. These structures are the factual or conceptual knowledge. The representation of such structures of knowledge is what varies most, in section 2.1 we will explore several different approaches used to model these structures.

From the consensus it can be assumed that these three key points are the core components of knowledge, other characteristics can be included to create more complete definitions, but these will be context dependent. With a basic notion of what knowledge is, more interesting questions can be posed in the following sections.

2.2.1 What types of knowledge do exist?

There are several ways to classify knowledge; the most common distinction is closely related to human memory: the memories related to facts and the memories related to processes, i.e., factual and procedural. Factual or declarative knowledge explains what things are e.g., *"the dogs eats meat or a dog has a tail"*. Procedural knowledge explains how things work for example what the dog needs to do in order to eat, e.g. *"if dog hungry → find food, then chew food, then swallow, then find more food if still hungry"*.

We use both types of knowledge in our everyday life; in fact it is hard to completely separate them; however, many computer models can only represent abstract ideal situations with simplified contexts in which each type of knowledge can be clearly identified, but trading off completeness for simplification. The three characteristics of knowledge, discussed in section 2.2, hold true for both types of knowledge, although they are easier to observe in declarative knowledge because on procedural knowledge concepts are integrated into processes, usually referred to as skills and competences, and the relations between them are imbued in rule sets.

Another important distinction is between structured and unstructured knowledge, since this has a strong implication on our reasoning processes. Structured knowledge relies strongly on organisation and analysis of information using higher cognitive processes, unstructured knowledge relies in lower cognitive processes such as associative knowledge and similarity.

In order for unstructured knowledge to become structured there needs to be a higher cognitive process involved in its acquisition and ordering knowledge such as taxonomy knowledge, domain knowledge, direction of causality, and description of the type of association, among others. Though some computer systems already do this in their knowledge representation such as semantic networks and Bayesian causality networks, they do so mainly on intuitive bases (Crisp-Bright A. K., 2010), where the particular reasoning process used is imbued in the heuristic or algorithm employed for information extraction and processing.

Both of these distinctions are important because they can strongly influence the way in which knowledge is represented, other common types of knowledge include domain specific knowledge which can be regarded as

a categorisation of knowledge by subject, such as taxonomic knowledge domain, ecological knowledge domain and causality knowledge domain, among others (Crisp-Bright A.K., 2010).

2.3 What are knowledge representation models?

The purpose of understanding what knowledge is, and what types of knowledge exist, is to allow us to use it in artificial systems. This long standing ambition has been fuelled by the desire to develop intelligent technologies that allow computers to perform complex tasks, be it to assist humans or because humans cannot perform them. In this section it will be explained how knowledge can be used in computer systems by representing it through different knowledge representation models. Knowledge representation is deeply linked to learning and reasoning processes. In other words, in order to have any higher level cognitive process, knowledge must be generated, represented, and stored. The works of Newell (1972, 1982, 1986, 1994) and Anderson (1990, 2004) provide comprehensive explanations for the relations between these processes, as well as computer frameworks to emulate them. Both Newell's Unified theories of Cognition (1994) and Anderson's Adaptive Character of Thought (1990) theory have strongly influenced today's knowledge representation models in cognitive and computer sciences, examples include the components of the Cognitive Informatics Theoretical Framework (Wang, 2009). Models are representations of theories that allows us to run simulations and carry out tests that would render outputs predicted by the theory, therefore when we speak of knowledge representation models, we are referring to a particular way of representing knowledge that will allow the prediction of what a system knows and what is capable of with knowledge and reasoning mechanisms. Since most knowledge representation models have been designed to emulate the human brain and its cognitive processes, it is common to find knowledge representation models that focus on long term memory (LTM), short term memory (STM) or combine both types of memory (Newell, 1982).

Having computers that can achieve complex tasks such as driving a car require intelligence.

Intelligence involves cognitive processes like learning, understanding and reasoning, and as has been said before, all of these processes require knowledge to support or guide them. As Cognitive Informatics states if computers with cognitive capabilities are desired (Wang, 2003), then computerised knowledge representations are required.

To understand how generic knowledge can be represented in abstract systems we must also understand the types of possible representations, it is important to consider that these representations are descriptions of the types of knowledge; therefore they are usually akin to particular types of knowledge. A helpful metaphor is to picture types of knowledge as ideas and types of representations as languages, not all languages can express the same ideas with the same quality, there are words which can only be roughly translated.

2.3 Types of knowledge representation models

A distinction should be made between types of knowledge and types of knowledge representation models. Types of knowledge were described in the previous section as *declarative* vs. *procedural* and *structured* vs. *unstructured*. Types of models are the different ways each type of knowledge can be represented.

The types of representation models used for knowledge systems include distributed, symbolic, non-symbolic, declarative, probabilistic, ruled based, among others, each of them suited for a particular type of reasoning: inductive, deductive, analogy, abduction, etc (Russell & Norvig, 1995). The basic ideas behind each type of knowledge representation model will be described to better understand the complex approaches in current knowledge representation models. Since this is a vast field of research, the focus will be directed to monotonic non probabilistic knowledge representations models.

Symbolic systems are called that way because they use human understandable representations based on symbols as the basic representation unit, each symbols means something i.e., a word, a concept, a skill, a procedure, an idea. Nonsymbolic systems use machine understandable

representations based on the configuration of items, such as numbers, or nodes to represent an idea, a concept, a skill, a word, nonsymbolic systems are also known as distributed system. Symbolic systems include structures such as semantic networks, rule based systems and frames, whereas distributed systems include different types neural or probabilistic networks, for instance. As their names states, semantic networks are concept networks where concepts are represented as nodes and associations are represented as arcs, they can be defined as a graphical equivalent for propositional logic. This type of knowledge representation models relies strongly on similarity, contrast and closeness for conceptual representation or interpretation. In semantic networks, associations have a grade which represents knowledge or strength of the association; learning is represented by increasing the grade of the association or creating new associations between concepts.

Semantic networks are commonly used to model declarative knowledge both in structured and an unstructured way, but they are flexible enough to be used with procedural knowledge. When modelling structured knowledge the associations must be directed and have information of causality or hierarchy.

Ruled based systems are symbolic representation models focused in procedural knowledge, they are usually organised as a library of rules in the form of condition - action, e.g., *if answer is found then stop else keep looking*. Rule systems proved to be a powerful way of representing skills, learning and solving problems, rule based systems are frequently used when procedural knowledge is present. Rule systems might also be used for declarative knowledge generally with classification purposes, e.g., *if it barks then is dog else not dog*. The *else* component is not actually necessary, when there is no *else* component systems do nothing or go to the next rule.

A frame is a data-structure for representing a stereotyped situation. Frames can be considered as a type of semantic network which mixes declarative knowledge and structured procedural knowledge. Frames are different from other networks because they are capable of including procedures (fragments of code) within each symbol. This means that each symbol in the network is a frame which contains a procedure, which is called a ‘demon’, and a group of attributes for the description of the situation. The idea behind the frame is to directly emulate human memory which stores situations that mix procedural and declarative knowledge. When we find ourselves in a situation similar to one we have lived before,

we allude to the stereotype stored in our memory so we can know how to react to this new situation. This theory is an attempt at joining unifying several other approaches proposed by psychology, linguistics and Artificial Intelligence.

Very similar and contemporary theory to theory of frames is theory of scripts. Scripts are language oriented as their name suggests they resemble a long sentence that describes an action. Scripts are part of the description of a larger plan or goal, which can also be used to model networks similar to those of semantic networks. Script theory was originally oriented toward the understanding of human language and focusing on episodic memory.

Since scripts and frames have theories resemble so much they are both treated as part of a same sub-group of semantic networks.

Neural networks are the most popular type of distributed knowledge representation models, instead of using a symbol to represent a concept they use an activation pattern over and entire network. A simple way to understand how neural networks work, is by looking at the place from where the idea came, i.e., the human brain. Humans have a number of neurons connected in a highly complex structure, each time a person thinks thousands or millions of neurons in a localised part of the brain activate. This pattern of activation can be used then to identify a concept or an idea; hence if a tiny specific part of the concept is lost, is does not affect the general idea because what matters is the overall pattern. The pattern is strengthened each time we think about it, we refer to this as training of a network. Neural Networks emulate this cognitive process of mental reconstruction.

The combination of these inputs will activate an input layer and will generate a pattern of propagation until it reaches the last layer where it will return the result of a function which could be a concept. Even though neural networks are very flexible and robust for knowledge representation of certain structures, they cannot be used for vast amount of knowledge, since they become too complex for implementation over a small amount of time. The second reason why neural networks are not used as large scale knowledge representation models is that they must be trained so they can learn the patterns which will identify specific concepts; this means that knowledge must be previously modelled as training sets before it can fed unto the net, thus it becomes unpractical for average knowledge retrieval. Also it is worth mentioning that the black box nature of the neural networks

does not show to get to the knowledge, it only shows that some inputs will render this and that output,i.e., its representation is non-symbolic. The real advantage of neural networks are their capacity to emulate any function, this implies that the entire network will specialize in that particular function therefore it cannot specialize on everything. Among the common types of Neural Networks the following can be found: perceptrons which don not have hidden layers; Feed forward networks, back propagation and resilient propagation which are networks with the same structure but differ in the approach used to adjust the weights of the networks; Radio based function networks;

Hopfield networks, which are bidirectional associative networks; and self-organizing feature maps, which are a kind of network that does not require much training per se; among others (Rojas 1996, Kriesel, 2011). Neural networks indeed are of very different natures but in the end they are all based on connectionist theory and are inspired on biological neural networks, in particular the human, brain science.

Ontologies remain a debate issue in two aspects, first as to what is to be considered an ontology, and second how it should be used in computer science (Weller, 2007). Some authors argue that simple hierarchical relations in a structure is not enough as to call it an ontology (Gauch & et. al 2007), while others use these simple structures and argument they are (Weller, 2007). The most relevant insights in artificial intelligence as to how to define ontologies in computer systems are provided by Grubber: “An ontology is an explicit specification of a conceptualisation … A conceptualisation is an abstract, simplified view of the world that we want to represent… For AI systems, what ‘exists’ is that which can be represented.” (Gruber, 1993). Gruber also notes that “Ontologies are not about truth or beauty, they are agreements, made in a social context, to accomplish some objectives, it’s important to understand those objectives, and be guided by them.” (Gruber, 2003) However this definition has created a new debate since it also applies to folksonomies (Gruber, 2007), especially since ontologies and folksonomies (Medelyan & Legg 2008) became popular in the context of semantic web through RDF and OWL (McGuiness & Harmelen, 2004) specifications. Weller (2007) and Gruber (2007) present a deeper explanation of this debate as well as the differences and advantages of each of both folksonomies and ontologies. In practical sense ontology are flexible hierarchical structures that define in terms that a computer can

understand, the relations between its elements, a language often used for this purpose is first order logic. In reality, ontologies have been used mostly as enhanced controlled vocabularies with associated functionalities and categorisation. Computational implementations of ontologies tend to resemble taxonomies or concept networks (Helbig, 2003, Chen 2009), i.e., semantic networks with formal conceptual descriptions for their associations, and therefore can be considered symbolic systems. Some examples of Ontology include those defined as part of an interaction communication protocol in multi agent systems (FIPA, 2000), those built though ontology edition tools for ontology web language (OWL) like *protégé* which are used to build *the semantic net*, and project CYC.

All representation models presented satisfy the three basic characteristics cited before.

Both symbolic and distributed systems recognise a concept as a unit of knowledge, the main difference between them is that one approach models it as a symbol and the other as a pattern. Both approaches agree on the need for associations between concepts and both recognise that the configuration of the associations also represents knowledge. It should be noted that some symbolic models like ontologies include instances as another layer for representation of the embodiment of a concept, however not every models includes them and therefor even though they will be mentioned in future sections they will not be included within the basic characteristics that all knowledge representation models have in common.

With this we conclude a basic introduction of what knowledge is and how it is represented in computers, now we will analyse each of the basic units that compose knowledge: concepts, skills and associations.

2.4 Concepts, skills and their acquisition

We have already explained that knowledge is divided in two types: factual and procedural;

Roughly speaking factual knowledge in a higher cognitive dimension can represent concepts, and procedural knowledge in higher cognitive scale can be used to represent skills. As was mentioned in section 1, this does not

mean that any fact can be considered a concept or any procedure a skill, the inter-association between each of these components as well as the structures they build must also be considered. To get a deeper understanding of knowledge we now review each of these components in more depth.

2.5 Definition of concept

The definition of a concept is closely related to the discussion of knowledge, in fact most of the theories attempting to explain one also explain the other. The most traditional definitions of concepts are based on Aristotelian philosophy and can be considered as revisions and complements previous works in the same line, Representational Theory of the Mind (Hobbes, 1651) was the first formalisation of this philosophy and Language of Though hypothesis (Fodor, 2004) is the latest extension added to it.

The Representational Theory of the Mind (RTM) states concepts and ideas as mental states with attributes sometimes defined as images, the Language of Thought (LOT) hypothesis states that thoughts are represented in a language which is supported by the principles of symbolic logic and computability. Reasoning can be formalised into symbols and characters;

hence it can be described and mechanised. In other words RTM states that concepts exist as mental objects with attributes, while LOT states that concepts are not images but words in a specific language of the mind subject to a unique syntax. A complete and practical definition of concept should be influenced by those two aspects, and therefore be as follows: A concept is considered as the representation of a mental object and a set of attributes, expressed through a specific language of the mind which lets it be represented through symbols or patterns which are computable. Such approach defines concepts as objects formed by a set of attributes, in the same atomic way as the Classic Theory of Concept Representation does (Osherson & Smith, 1981), but also considers descriptive capabilities of the role of a concept in the same as the approach of Concepts as Theory Dependent (Carey, 1985; Murphy and Medin, 1985; Keil,1987). This definition is useful for declarative knowledge since it can be easily included to most existing models and remains specific enough to be computationally implemented as will be shown in section 4.

2.6 Definition of skill

Philosophic views such as (Dummet 1993, Kenny 2010) propose that abilities and concepts are the same thing, however, these approaches have not been very popular in computer and cognitive sciences, because of studies made in learning theories from Cognitive Science provide a more practical and empirical approach which instead support the Aristotelian view of concepts. Skills are practical manifestations of procedural knowledge, the most popular definitions of skills used today are based on constructivist theories and variations of Bloom's Taxonomy of Skills, this comes as a historic consequence of research in education, were skills is a core interest in educational psychology. Therefore, it is then not strange that the most referenced theories for skill development are found in this social science.

Vygotsky's constructivist theory (Vygotsky, 1986) explains how skills are developed through a complex association process and upon construction of dynamic structures which can be traced through internal language or speech. Bloom's taxonomy for skills provides perhaps the most practical classification and enumeration of cognitive, social and physical skills. The combination of those works establishes enough theoretical insight to build more complex models for skill representation, such as those used in Cognitive Informatics for the Real Time Process Algebra (Wang, 2002), Newell's Soar cognitive architecture (Newell, 1990) and Anderson's ACT-R cognitive architecture (Anderson, 1994).

In Thought and Language, Vygostky (1986) explains several processes used to learn and create ideas. Ideas stated as concepts and skills dynamic in nature behave as processes incontinual development which go through three evolution stages starting at the basic stage of syncretism heaps, which are loosely coupled ideas through mental images, and concluding in formal abstract stable ideas, which are fully developed concepts and skills that manifest in language.

Benjamin Bloom (1956) developed a taxonomy for skills with a very practical approach, in which three domains are specified: cognitive, affective, and psychomotor. Each domain contains different layers depending on the complexity of the particular skill. Bloom's taxonomy is widely used, however, as with any other taxonomy, criticisms have been raised; Spencer Kagan (2008) made the following observations:

1. A given skill can have different degrees of complexity; hence a layer model might not provide an adequate representation.

2. Skill integration in complexity order does not always keep true.

These observations imply that if there is a hierarchy in skills it must be dynamic in nature and this characteristic must be taken into account when defining what a skill is. The idea of flexible structure can also be found in Vygotsky's theories. In the framework for Cognitive Informatics, Wang (2002) proposes an entire system for describing processes, according to what we now know of procedural knowledge we can use such system to define skills in computational terms, thus under this train of thought skills are pieces of computer code located in an action buffer, such processes are composed by sub-processes and are described using Real Time Process Algebra (RTPA). RTPA is oriented to a structured approach where a skill is not as flexible as Kagan's observations suggest, the types of data, processes, metaprocesses and operations between skills, should be included in a comprehensive definition

of skills.

Using constructivist theories as a basis, Bloom's taxonomies for organisation and the cognitive architectures for mappings to computational terms, a generic definition for skills in computer systems can be stated as: A cognitive process that interacts with one or more concepts as well as other skills through application and has a specific purpose which produces internal or external results. Skills have different degrees of complexity and may be integrated or composed by other skills. In contrast with concepts which are factual entities by nature, skills are process oriented, they are application/action related by nature and it is common to describe them using verbs.

2.7 Associations between concepts and skills

Of the three basic common characteristics of knowledge stated in section 1, perhaps the second characteristic: *Concepts have relations or associations to other concepts,* is the most agreed upon. Every theory and model reviewed so far agrees that associations are vital to knowledge (Hobbes,

1651, Fodors, 1975, Vygotsky, 1986, Bloom, 1956, Kagan, 2003, Newell, 1990, Anderson, 1994, Quillian, 1968, Wang, 2002, Helbig, 2003, among others); the differences appear when defining their properties and implications, these are better observed in cognitive or computer models, since more general theories tend to be vague in this regard and detailed specification is a requirement for computer models (Marr, 1982).

Most declarative knowledge representation models rely on propositional logic or its graphical equivalents in network representations e.g., Cyc (Read & Lenat, 2002), WordNet (Miller, 1990) , OAR (Wang, 2006), Multinet (Helbig, 2003) and Telos (Paquette, 1990) among others, the specific type of the network is determined by aspects such as directionality of associations (Helbig, 2003), the type of association (Wang, 2006), if the associations allows cycles, if they are hierarchical in nature (Paquette, 1990) or mixed and if there is a grouping or filtering scheme for them.

Traditional semantic networks only used presence or absence of associations; current semantic networks such as MultiNet or Object Attribute Relation OAR (Wang, 2007) provide deeper types of associations and integrate layers for knowledge composition. Examples of deeper type of association can be seen in MultiNet where associations are defined as a third type of node that contain procedural knowledge similar to Minsky frames, or OAR associations which are described as types of relations which can be grouped into several categories: Inheritances, Extension, Tailoring, Substitute, Composition, Decomposition, Aggregation and Specification. OAR categories are in fact operations for Concept Algebra (Wang, 2006), i.e., a mathematical way to describe how knowledge structures are integrated.

Concept algebra does not include procedural knowledge, for this reason RTPA has a different set of associations which describe a hierarchy for composition of processes; both real time process and concept algebras are integrated in a higher framework called system algebra (Wang, 2009).

Associations are important because they create the context and embody semantic meaning for each context, some authors refer to this as sense (Vygostky, 1986), others discriminate between intrinsic knowledge, i.e., knowledge inherent to that concept, and context knowledge i.e., knowledge inferred from the associations and other concepts surrounding the original concept (Helbig, 2008). Understanding these approaches we can then summarise that an association is a relation between two elements, which can

be skills or concepts that contain a particular function and a directionality that explains the nature of the relation.

Groups of associations are what create contexts and each of these contexts may provide a uniquely different sense to a concept or skill which should reflect upon interpretation and inference process.

2.8 A model for the representation of concepts and skills In different contexts

An important functionality for knowledge representation models is the capacity to represent multiple contexts in a single instantiation, as well as the impact that context changes have on a concept's meaning. Approaches such as micro-theories models used in Cyc contemplate this and have successfully managed to combine multiple facts of a subjective nature into a coherent knowledge base, however, Cyc requires understanding of its own native language which is based on predicate logic semantics for information modelling and for information extraction as well, this has proven a problem for most users (Lenat, 2006). Simpler graphical representations which retain this context flexibility and can be represented in computers present an attractive alternative for average users, such as domain experts not versed in CYC language. Graphical oriented models such as Multinet or OAR have been used for natural language processing and for knowledge composition and process specification respectively, but their focus is not to represent several contexts a time.

Multinet for example has specific context differentiation based on grammar attributes such as singular or plural elements, however, it does not have differentiators for the concepts meaning when the context changes. In these models when a new context is to be created only a small fraction of the information of concepts is reused and most of it has to be reinstantiated for each domain, this is a common trait of knowledge representation models that have instances as part of their model. OAR presents a similar situation since the context is defined as the relation between objects and its attributes in a given set (Wang, 2006). OAR is more flexible and does contemplate multiple contexts for the instantiations of the concepts, but not for the concepts themselves, which means that what are dynamic are not the

concepts themselves but the objects in regard to the context. The implication for this is that a concept will have several different instantiations depending on the context, however this issue does not represent the impact the context has on the formation of a concept as was described by Vygotsky (1986).

Chapter 3

Natural Language Generation Approaches and Tecniques

This chapter explores the current state of the task of generating easily understandable information from data for people using natural language, which is currently addressed by two independent research fields: the natural language generation field — and, more specifically, the data-to-text sub-field — and the linguistic descriptions of data field. Both approaches are explained in a detailed description including: (1) a methodological revision of both fields including basic concepts and definitions, models and evaluation procedures; (2) the most relevant systems, use cases and real applications described in the literature. Some reflections about the current state and future trends of each field are also provided, followed by several remarks that conclude by hinting at some potential points of mutual interest and convergence between both fields.

3.1 Natural Language Generation: an Introduction

The problem of automatic production of natural language texts becomes more and more salient with the constantly increasing demand for production of technical documents in multiple languages; intelligent help and tutoring systems which are sensitive to the user's knowledge; and hypertext which adapts according to the user's goals, interests and prior knowledge, as well as to the presentation context. This section will outline the problems, stages

and knowledge resources in natural language generation. Natural Language Generation (NLG) systems produce language output (ranging from a single sentence to an entire document) from computer-accessible data usually encoded in a knowledge or data base. Often the input to a generator is a high-level communicative goal to be achieved by the system (which acts as a speaker or writer). During the generation process, this high-level goal is refined into more concrete goals which give rise to the generated utterance. Consequently, language generation can be regarded as a goal-driven process which aims at adequate communication with the reader/hearer, rather than as a process aimed entirely at the production of linguistically well-formed output (ILC-CNR, 1996).

Nowadays, the task of generating easily understandable information for people using natural language is being addressed by two fields which, independently until now, have researched the processes this task involves from different perspectives: the natural language generation (NLG) field and the linguistic descriptions of data (LDD) field.

The natural language generation field consists in the creation of texts which provide information contained in other kind of sources (numerical data, graphics or even other texts), with the aim of making such texts indistinguishable, as far as possible, from those created by humans. On the other hand, the linguistic descriptions of data field, which arises as one of the many applications born from the fuzzy sets theory, provides summaries or descriptions from data sets using linguistic concepts defined as fuzzy sets and partitions, which deal with the imprecision and ambiguity of human language. The NLG field has been in development since the 1980s (although there are systems which date from even before this period, e.g. (Swartout, 1977), when the first applications which translated data into legible texts appeared (e.g., (Kittredge, 1986; Boyer, 1985). Since then, the complexity of the developed systems has increased notably and there are several techniques and methodolo-gies which guide the building of these solutions (Reiter, 2000; Mellish, 2006; Reiter, 2007). Even so, this research field is still open in many respects and there is no unique and well defined approach to address NLG problems.

The linguistic descriptions (or summaries) of data aim to obtain informative, brief and concise descriptions from numeric datasets and cover a group of soft computing-based techniques, such as linguistic variables or fuzzy quan-tifiers and operators. It is a young field when compared to the

NLG domain, whose solutions provide information in the form of linguistic terms. Specifically, although preliminary ideas appeared early in the 1980s (Yager, 1982; Yager, 1990), it started to develop in the second half of the 1990s, when the advances in the field of fuzzy sets (namely computing with words (Zadeh, 1996) and the computational theory of perceptions (Zadeh, 2000; Zadeh, 2001) provided new potential applications in the descriptive side of data mining. Due to its short career and its formal background, many approaches in this field are on the theoretical side, although in some cases practical examples and real life based problems are given. More recently, the use of hy-brid approaches which employ LDD techniques together with NLG systems to provide solutions to real life problems has emerged (Ramos-Soto, 2015).

3.2 From Knowledge to Text

Natural language generation (NLG) is described by (Bateman, 2001) in as the branch of natural language pro-cessing which deals with the problem of how texts in human natural language can be automatically created by a machine. This may be seen as the inverse of the problems addressed by natural language understanding but, actually, the NLG field emerges from a very different set of motives and objectives, both theoretical and practical. In this sense, on the theoretical side it explores how language is grounded in non-linguistic information and how it is produced. From a practical point of view, NLG tries to provide solutions for text generation problems in real life application contexts.

The demand of natural language texts which provide all kinds of information is currently increasing. Thus, it is likely that NLG will be a key information technology in the future (a good indicator of this is the considerable number of NLG companies which have emerged in recent years). As a consequence, many NLG systems have found a practical use, while the demand of real life applications is having a growing impact in the approaches and questions contemplated in the NLG field. Examples of well established NLG applications include the generation of weather reports from meteorological data in several languages, the creation of custom letters which answer customers' questions, the generation of reports about the state

of neonatal babies from intensive care data, and the generation of project management and air quality reports.

Bateman also states that, usually, it is hard for a casual user to distinguish between hand made texts, texts built using simple techniques or a complete natural language generation using NLG technology. This is, in fact, what any NLG solution should achieve in order to be considered successful. It should be simply a perfect text production which ideally fulfills the necessities and the knowledge of the reader/listener. This duality directly translates into two quite different research issues within NLG:

- producing texts which are humanlike,
- producing comprehensible texts to fulfill certain needs.

The fact that auser is incapable of distinguishing between texts however they are produced is also a problem for the research and development of NLG in the sense that it implies that the required effort to build a successful NLG system is hard to be perceived by users. Since users are not frequently aware of it until something goes wrong, there is little appreciation of the possibilities and complexities of a full natural language generation. In fact, users and application developers who could see the utility of providing automatically produced flexible texts in natural language are not aware of the complexity it might imply, the available range of technological solutions and the effort level required to create scalable solutions.

In this sense, the complete range of possible applications has not been broadly explored. Given this potential as well as the wide range of interests involved, it should not come as a surprise that NLG has experienced a fast growth since the 1990s. This makes providing an exhaustive revision of the field rather complicated. Until the end of the 1980s it was almost possible for a revision to enumerate the most significative systems in NLG. This, however, is not currently feasible: the most extensive list of NLG systems is [20], which currently contains near 400 systems and is regularly updated as new systems appear.

It must also be noted that NLG can be divided into several sub-fields depending on the type of communicative tasks they perform and the kind of input they receive (e.g., NLG in interactive systems, narrative NLG or data-to-text NLG, among others). Although many of the concepts and ideas in this discussion are made on a general sense, for this review we are focusing mainly on data-to-text, which strongly resembles the linguistic descriptions of data field. Furthermore, data-to-text has allowed the emergence of the

most successful applied NLG systems and is the most commercially-oriented NLG sub-field.

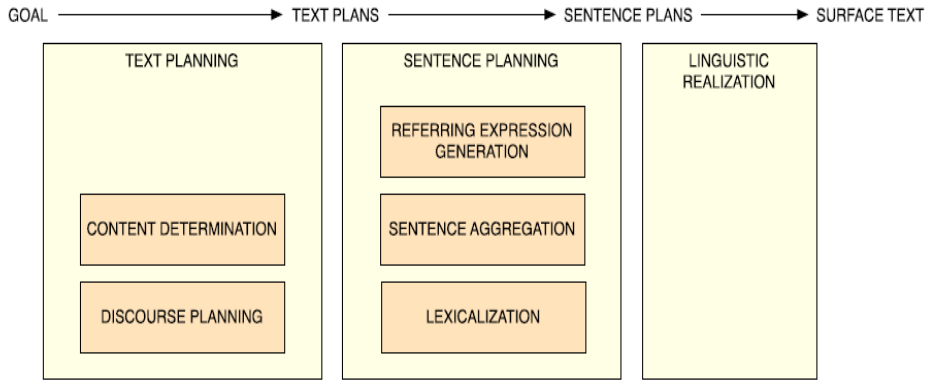

Figure 2: A general schema for Natural Language Generation Process

3.3 Design of a NLG system

The design of NLG systems is an open field where a broad consensus does not exist. Instead, there is a diversity of architectures and implementations which depend on the developer and the problem for which the NLG system is created. In this sense, it is hard to identify common elements and to provide a complete abstraction which is applicable to most NLG systems.
However, there does exist a certain agreement about the tasks that a NLG system usually performs. However, there does exist a certain agreement about the tasks that a NLG system usually performs. (Reiter, 1997) argues that, in general terms, the main task of a natural language generation system can be characterized as the conversion of some input data into an output text. However, as in most computational processes, this task can be splitinto a number of substages or modules which then can be further specified. In this context they present a sequential pipeline architecture for NLG divided into general three stages (Fig.2):

- *Text planning*
- *Document planning*
- *Surface realization*

This architecture is then further decomposed into six basic activities (Fig.2):

- **Content determination**. It is the process of deciding which information shall be communicated in the text. It can be perceived as the creation of a set of messages from the system input. Those messages are the data objectsused in the subsequent tasks. In general terms, the message creation process consists in filtering and summarizing the input data. The messages are expressed in some kind of formal language which labels and distinguishes the entities, concepts and relations determined by the application domain.

- **Discourse planning**. It is the process by which the set of messages to be verbalized is given an order and structure. A good structuring can make a text much easier to read. In the general architecture, text planning combines the tasks of content determination and discourse planning. This reflects the fact that in many real applications it is hard to separate these activities.

- **Sentence aggregation.** This process groups several messages together in a sentence. This task is not always neces-sary (each message can be expressed in a separate sentence), but in many cases a good aggregation significantly improves the fluidity and readability of a text.

- **Lexicalization.** In this process it is decided which words and specific expressions must be used to express the concepts and relationships of the domain that appear in the messages. In many cases this task can be performed trivially, assigning a unique word or phrase to each concept or relationship. In others, however, the fluidity can be improved allowing the system to vary the words used to express the concepts and relationships.

- **Referring expression generation**. This task selects words or expressions which identify entities from the domain. Although this task seems similar to the previous one, in this case the referring expression generation is charac-terized as a discrimination activity, in which the system needs to provide enough information to differentiate one domain entity from the rest. In the general architecture, sentence planning combines the sentence aggregation, lexicalization and referring expression generation processes.

- **Linguistic realization.** This task, which directly matches the one defined in the general architecture, applies gram-matical rules to

produce a text which is syntactically, morphologically and orthographically correct.

Although, in general, these six tasks are considered as essential in a complete NLG system, the way in which they are structured allows many variants, depending on the specific language generation problem and its associated complexity. This, in fact, implies that a NLG system does not necessarily need to be composed of these six modules, since in many cases some of these activities can merge into a single module or are not needed if the language generation complexity is low. For instance, template-based NLG addresses several of these tasks at once, although this usually comes at the cost of flexibility due to the use of relatively fixed templates. An interesting discussion about the use of standard and template-based approaches is given by (Van Deemter, 2005), where the authors suggest that there is no such a gap between both approaches.
While the model provided by (Reiter, 1997) can be considered the de facto standard classically, other authors have also explored and reviewed the complexity and variety of tasks and architectures in NLG. In this sense, (Mellish, 2005) shows that:

i) there is a very broad variety of tasks;

ii) most NLG systems adopt some of these tasks, but not all;

iii) the architectures of such systems often do not follow the pipeline described in (Reiter, 1997).

In order to respond to this reality, in (Mellish, 2005) the RAGS framework is proposed; it relaxes the "architectural" requirement to a point where it is sufficiently inclusive of actual systems to be relevant, yet still sufficiently restrictive to be useful.
In such a perspective, a characterization at a quite abstract level for the data types, functional modules and protocols for manipulating and communicating data that most modular NLG systems seem to embody, is performed. For this, the RAGS proposal considers the following elements:

- A high-level specification of the key (linguistic) data types that NLG systems manipulate internally. This uses abstract type definitions to give a formal characterization independent of any particular implementation strategy;

- A low-level reference implementation specifying the details of a data model flexible enough to support NLG systems.

- A precise XML specification for the data types, providing a standard "off-line" representation for storage and communication of data between components.

- A generic view of how processing modules can interact and combine to make a complete NLG system, using data formats "native" to their particular programming languages which are faithful to the high-and low-level models and exploiting agreed instantiations of the high-level data types.

- Several sample implementations to show how the development of a range of concrete architectures can be achieved.

3.4 Knowledge Sources

In order to make these complex choices, language generators need various knowledge resources, as listed below:

- **discourse history** - information about what has been presented so far. For instance, if a system maintains a list of previous explanations, then it can use this information to avoid repetitions, refer to already presented facts or draw parallels.
- **domain knowledge** - taxonomy and knowledge of the domain to which the content of the generated utterance pertains.
- **user model** - specification of the user's domain knowledge, plans, goals, beliefs, and interests.
- **grammar** - a grammar of the target language which is used to generate linguistically correct utterances. Some of the grammars which have been used successfully in various NLG systems are:
    - *unification grammars--Functional Unification Grammar, Functional Unification Formalism,*
    - *Phrase Structure Grammars--Referent Grammar (GPSG with built-in referents), Augmented Phrase Structure Grammar;*
    - *systemic grammar;*
    - *Tree-Adjoining Grammar;*
    - *Generalised Augmented Transition Network Grammar.*

- **lexicon** - a lexicon entry for each word, containing typical information like part of speech, inflection class, etc.

The formalism used to represent the input semantics also affects the generator's algorithms and its output. For instance, some surface realisation components expect a hierarchically structured input, while others use non-hierarchical representations. The latter solve the more general task where the message is almost free from any language commitments and the selection of all syntactically prominent elements is made both from conceptual and linguistic perspectives. Examples of different input formalisms are: hierarchy of logical forms, functional representation, predicate calculus, conceptual graphs.

3.5 Text Summarization

With the proliferation of online textual resources, an increasingly pressing need has arisen to improve online access to textual information. This requirement has been partly addressed through the development of tools aiming at the automatic selection of document fragments which are best suited to provide a summary of the document with possible reference to the user's interests. Text summarization has thus rapidly become a very topical research area.
Most of the work on summarization carried out to date is geared towards the extraction of significant text fragments from a document and can be classified into two broad categories:

- **domain dependent approaches** where a priori knowledge of the discourse domain and text structure (e.g. weather, financial, medical) is exploited to achieve high quality summaries;
- **domain independent approaches** where a statistical (e.g. vector space indexing models) as well as linguistic techniques (e.g. lexical cohesion) are employed to identify key passages and sentences of the document.

Considerably less effort has been devoted to ``text condensation" treatments where NLP approaches to text analysis and generation are used to deliver summary information of the basis of interpreted text.

3.5.1 Domain Dependent Approaches

Several domain dependent approaches to summarization use Information Extraction techniques, in order to identify the most important information within a document. Work in this area includes also techinques for Report Generation and Event Summarization from specialized databases.

3.5.2 Domain Independent Approaches

Most domain-independent approaches use statistical techniques often in combination with robust/shallow language technologies to extract salient document fragments. The statistical techniques used are similar to those employed in Information Retrieval and include: vector space models, term frequency and inverted document frequency. The language technologies employed vary from lexical cohesion techniques to robust anaphora resolution.

3.6 Role of Lexical Semantics

In many text extraction approaches, the essential step in abridging a text is to select a portion of the text which is most representative in that it contains as many of the key concepts defining the text as possible (textual relevance). This selection must also take into consideration the degree oftextual connectivity among sentences so as to minimize the danger of producing summaries which contain poorly linked sentences. Good lexical semantic information can help achieve better results in the assessment of textual relevance and connectivity. For example, computing lexical cohesion for all pair-wise sentence combinations in a text provides an effective way of assessing textual relevance and connectivity in parallel [Hoe91]. A simple way of computing lexical cohesion for a pair of sentences is to count non-stop (e.g. closed class) words which occur in both the sentences. Sentences which contain a greater number of shared non-stop words are more likely to provide a better abridgement of the original text for two reasons:

- the more often a word with high informational content occurs in a text, the more topical and germane to the text the word is likely to be, and
- the greater the times two sentences share a word, the more connected they are likely to be.

The assessment of lexical cohesion between text units can be improved and enriched by using semantic relations such as synonymy, hyp(er)onymy as well as semantic annotations such as subject domains in addition to simple orthographic identity. Related areas of research are: Information Retrieval, Information Extraction and Text Classification.

3.7 Paraphrasing and textual entailment

As widely discussed in (Malakasiotis, 2011), in recent years, significant effort has been devoted to research on paraphrasing and textual entailment (Androutsopoulos and Malakasiotis, 2010;).
Paraphrasing methods recognize, generate, or extract (e.g., from corpora) paraphrases, meaning phrases, sentences, or longer texts that convey the same, or almost the same information.
For example, (1.1) – (1.3) are examples of paraphrases.

(1.1) Leo Tolstoy wrote “War and Peace”.

(1.2) “War and Piece” was written by Leo Tolstoy.

(1.3) Leo Tolstoy is the writer of “War and Peace”.

Paraphrasing methods may also operate on templates of natural language expressions, like (1.4) – (1.6), where the slots X and Y can be filled in with arbitrary phrases; e.g.,

X = “Jules Verne” and Y = “Around the World in Eighty Days”.

(1.4) X wrote Y.

(1.5) Y was written by X

(1.6) X is the writer of Y

Textual entailment methods, on the other hand, recognize, generate, or extract pairs <T ; H> of natural language expressions, such that a human who reads (and trusts) T would infer that H is most likely also true.

For example, (1.7) textually entails (1.8), but (1.9) does not textually entail (1.10).1

(1.7) The drugs that slow down Alzheimer's disease work best the earlier you administer them.

(1.8) Alzheimer's disease can be slowed down using drugs.

(1.9) Drew Walker, Tayside's public health director, said: "It is important to stress that this is not a confirmed case of rabies."

(1.10) A case of rabies was confirmed.

As in paraphrasing, textual entailment methods may also operate on templates.
The natural language expressions that paraphrasing and textual entailment methods consider are not always statements. In fact, many of these methods were developed having question answering (QA) systems in mind. In QA systems for document collections, a question may be phrased differently than in a document that contains the answer, and taking such variations into account can improve system performance significantly. Paraphrasing and textual entailment methods are also useful in several other natural language processing applications, including for example text summarization, especially multi-document summarization, sentence compression, information extraction systems, machine translation, and natural language generation. Among other possible applications, paraphrasing and textual entailment methods can be employed to simplify texts, and to automatically score student answers.

Chapter 4

A Multidimensional Representation Model for Knowledge supporting UserProfiling and Domain Driven Text Generation

Natural Language Generation (NLG) Systems, applied in CH domain, are investigated in (Androstopulous, 2013). They are employed in order to build structured textual descriptions, based on cultural objects ontologies as lexical vocabulary and documents plan to establish the phrasing structures. The authors propose Natural OWL (Galanis, 2008), an effective working implementation of a NLG engine, able to automatically generate simpler or more complex textual descriptions in two different languages, English or Greek. System feeds with a lexical ontology, a micro-plan for text structure and users' profile information. Entities vocabularies are fixed for all type of users and the profiling information are used to modify some text features, as length. So, the general appearance of the textual description keeps quite unchanged but such a system represents an example of authoring system in the CH domain.

4.1 General Aims of the proposed solution

Given a Domain of Interest (e.g. C.H.) we need to represent the related knowledge in a double way:

- A *machine readable* one (for automatic computation)
- A *human readable* one (for human enjoiment)

Providing the opportunity to transform one into the other, automatically:

- without information loss (from text to knowledge synthesis)
- Taking into account the diversity of:
  - target HUMAN users (user-profiling) (structuring (verbalizing) knowledge for multiple textual profiled descriptions generation)
  - target languages (machine translation is not a one to one process (e.g., problem of linguistic blunders))
  - language rapid metamorphosis (linguistic deviations, idiomatic sentences, neologisms, standard de facto but not de iure in the official language)

The proposed solution aims to face the following problems, which can be summarized as follows:

1. Identification and Formalization of a representation model for knowlege able to support user-driven and domain-driven automatic text analysis and generation
   - Reinforcement of *Textual Entailment Recognition* and *Paraphrasing Generation* Processes

2. Automatic Annotation of Knowledge and Linguistic Resources (in a User and Specific Domain Perspective) by Textual BigData Acquisition and Processing:

- Lexical resources;
- Domain and Linguistic Ontologies;
- Users' folksonomies and Taxonomies of Users' Common Linguistic Deviations (extremized in wide spread syntax mistakes (solecisms), barbarisms (forcing usage of foreign terms in the current language), linguistic blunders, etc..).

4.2 NLG Tasks supported by the proposed model

The proposed model for knowledge representation aims to support NLG tasks by attempting to catch some key aspects involved into NLG subtasks. The questions to which this model try to answer are summarized in the following list:

- What about INPUT Knowledge Sources Organization and Selection Strategies?
    - Desiderata for output text:
        - How can we represent it?
        - How and in which step should we introduce it in the NLG process?
        - Which further resources are needed?
        - How and where can we retrieve them?
        - Which strategies to select the most compliant sources for output desiderata satisfaction?

- Document Construction Strategies? How to drive the process?
    - Many systems perform these tasks simultaneously because often rhetorical goals determine what is relevant.
    - Most text planners have hierarchically-organized plans and apply decomposition in a top-down fashion following a planning techniques. S
    - ome planning approaches rely on previously selected content - an assumption which has proved to be inadequate for some tasks

- Multiple diversified output preserving semantic equivalence
    - Paraphrasis generation

- Terms Replacement (based on Synonim and Hypernym substitution, very basic and weak strategy)
- Referring Expressions Selection (e.g., X wrote Y, X is the author of Y, Y was written by X, etc..) (more sophisticated, needs strategies for suitable selection)

4.3 The Multidimensional Knowledge Representation Model

Many different dimensions of knowledge have to be taken into account for a text generation with established quality properties. A multidimensional model for representing knowledge underlying text analysis and text generation is an effort to describe and keep together Knowledge resources needed to catch most of the expected and desired features for a textual output.
The proposed model is composed by:

- An Abstract Conceptual Level, describing Concepts, Properties and Relationships, remapped over an RDF Schema (adopting SKOS (Simple Knowledge Organization System) Vocabulary);
- RDF/XML was adopted to express (serialize) the RDF graph as an XML document.

The consitutive elements of the proposed model are:

- A set of conceptual bricks *CB = {HB,SB}*:
  - HardBricks HB= {Artifact (AF), Artifact Plan (AFP), Knowledge Dimension (KD), Target Requirements Set (TRS)}; →SKOS <Concept>
    - Main entities
  - SoftBricks (SB)= {res_id, res__name, res_date, res_author, res_uri, res_tag}
    - Properties and tags for HB (SKOS labels and notation)
    - *res_id* is a mandatory and unique value property
- A set of relationships R = {*R1, R2, R3} : CB → CB* :
  - Hierarchical composition (SKOS collection) *R1: HB → HB*

        - Meronym relation (part of) expressing composition of higher Level HBs of lower Level HBs;
    - Association (SKOS related) *R2: SB → HB*
        - linking properties SB to HB;
    - Annotation (SKOS notation) *R3: SB → HB*
        - annoting HB for NLP Process

4.4 The Model Structure

Model Structure is described below:

- HB_AF: Artifact: a container element bridging Target Requirements Set with Knowledge Resources; it is composed of:
    - A set of properties
    - An Artifact plan
    - A Target Requirements Set
- HB_AFP: Artifact Plan: a collection of Knowledge Dimensions;
- HB_TRS: Target Requirements Set: a set of requirements specified to customize text generation process and adopted resources:
- Its composition depends on the semantic annotation process for knowlege resources; typical elements are:
    - *Target language*
    - *Target domain*
    - *Target user*
    - *Target application*

4.5 Six Knowledge Dimensions

HB_KD1: Domain Knowledge

- Aim → modeling specific domain entities, properties and domain relevant relations
- Author: domain experts

- Short Description: often referred as domain ontology, it Includes domain template and domain instances (assertional knowledge)

HB_KD2: Basic Language Lexicon

- Aim → describing general dictionaries or semantic lexicons for reference language
- Author: language experts
- Short Description: General and Basic Vocabularies or Semantic Lexicon for interest Domain)
- Addon: linguistic blunders taxonomies for intermediate translations.

HB_KD3: Grammars for Text Coherent Planning

- Aim → mapping domain knowledge relations to extended referring expression, also providing annotated variations for the same relation
- Author: language/communication experts
- Short Description: The Grammar Structures and Rules underlying Text Composition and Alternative Expression Evaluation (Paraphrases Selection)

- **HB_KD4: Domain Lexicon**

- Aim → describing dictionaries or semantic lexicons for specialist and techical terms for considered domain
- Author: domain experts
- Short Description: Specific Vocabularies or Semantic Lexicon for interest Domain)

HB_KD5: Target Audience (User) Model

- Aim → taking into account more meaningful fetaures for target audience characterization: age, interest or skils toward specific domain
- Author: communication experts
- Short Description: user's affiliation level towards the domain is crucial for lexicon selection; age features can influence the grammar structure selection (referring expressions).

HB_KD6: Target Application

- Aim → taking into account more constrained fetaures for target application: lenght (for user's enjoyment), time duration (for TextToSpeech application), memory usage (mobile device applications), etc..
- Author: technology experts

- Short Description: length and memory usage can significantly impact over the enjoyment or usefulness of text in constrained application contexts.

Figure 3 and Figure 4 show an graphical view for the proposed model.

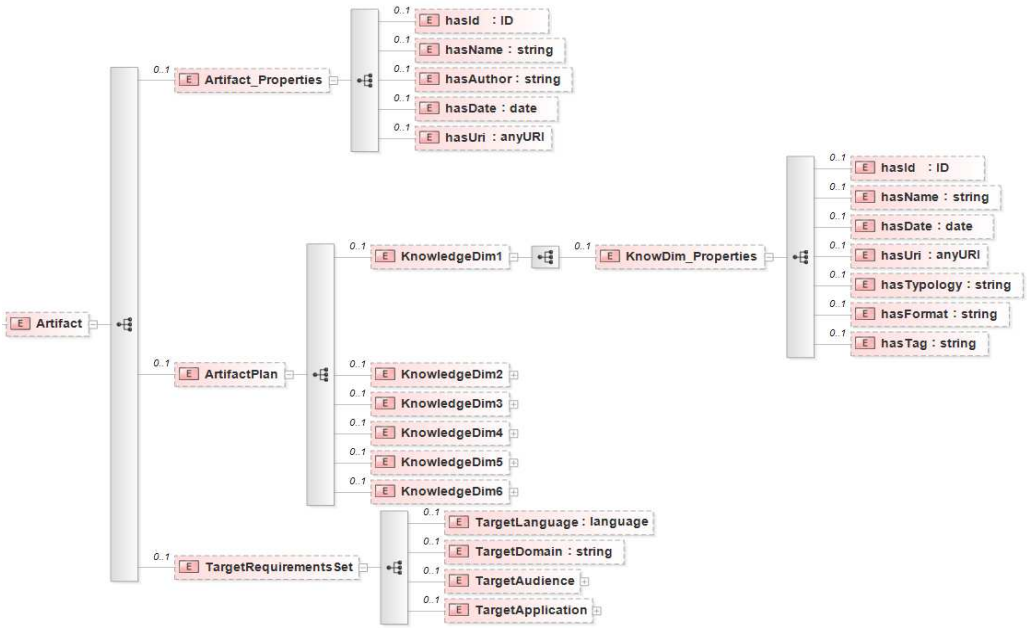

Figure 3: Multidimensional model 1

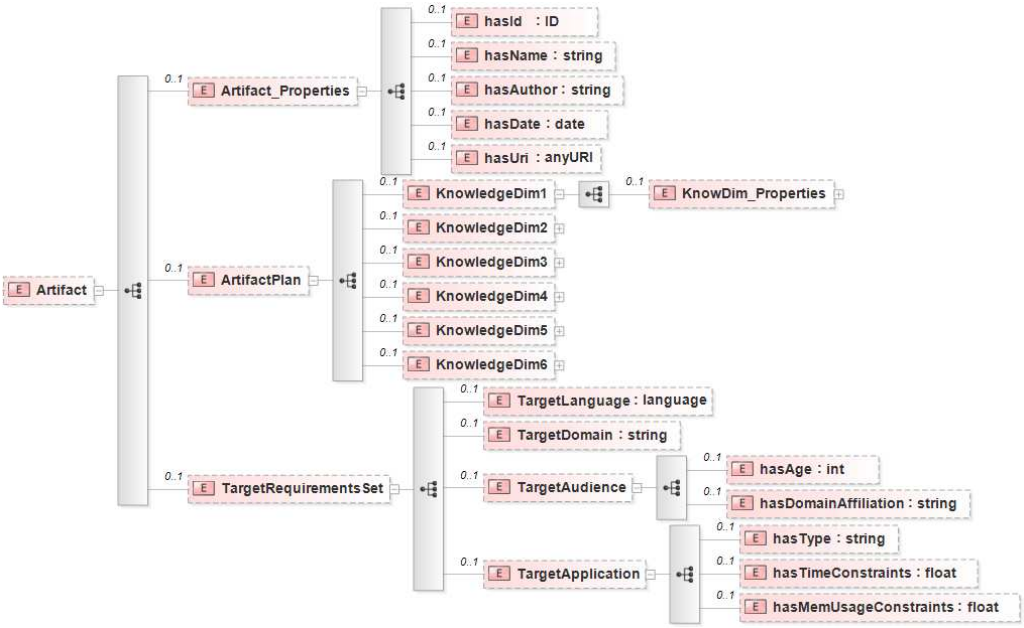

Figure 4: Multidimensional model: a detailed view

Chapter 5

Cultural Heritage Applications: a Case Study

Cultural Heritage has got great importance in recent years, in order to preserve countries history and traditions and to support social and economic improvements. Typical IoT smart technologies represent an effective mean to support understanding of Cultural Heritage, by their capability to involve different users and to catch their explicit and implicit preferences, behaviors and contributions.

In this chapter we illustrate a Case of Study in the CH domain, in order to demonstrate effectiveness for the proposed multidimendional model of knowledge, illustrated in the previous chapter. We will explain the authoring platform FEDRO (Marulli, 2015), as part of an intelligent infrastructure developed into DATABENC District, to support cultural exhibition of "talking" artworks, among which that one called "Il Bello o il Vero", exhibiting sculptures and held in the Southern Italy, in 2015. FEDRO is a prototypal version of a software system for acquiring data from domain experts in the form of scientific catalogue sheets (in compliance with the reference model for Fine Arts and Cultural Goods dictation from Italian Ministry (MIBACT)), and generating automatically textual and users profiled artworks biographies. Such biographies can be employed to feed a smart app for guiding visitors during the exhibition or as material for alimenting holographic projections reproducing the human presence with a natural language interaction (Marulli, 2016; Vallifuoco, 2016). A preliminary experimentation revealed a tangible improvement in the users' experience appreciation during the visit. Quality estimations of generated output were also computed exploiting users' feedbacks, collected through a manual questionnaire, subscribed at the end of their visit.

5.1 Fedro platform System Architecture

A general overview of FEDRO platform architecture and processing flow is shown in Fig. 6. Its users are mainly domain experts, enabled to to fill in original complex artworks textual descriptions (documents corpora) by a friendly GUI. They can select the target audience and language (currently, English and Italian) and new profiled descriptions are provided as output.

Additional process inputs are users' profiles tables, lexical dictionaries and domain ontologies, user generated terms taxonomies (folksonomies), sentences taxonomies (containing the phrasal structures and language rules needed during the customized text generation step).

At a glance, the processing flow is composed of the following four steps:

1. **Text analysis**: typical text analysis and summarization techniques are applied to input documents corpora; terms and sentencies are

extracted and disambiguated by the support of lexical and domain ontologies The output is represented by lists of relevant terms and sentences.

2. **Semantic enhancement:** lists of terms and sentences are semantically enriched and expanded. Terms are annotated by a detailed description and a list of synonymous, each one provided with a label indicating the most appropriate lexical forms for each type of user. Domain ontologies (for specialist terms), Linked Open Resource Archives and sentencies taxonomies are employed to select new simplified sentences, according to semantic similarity criteria.

3. **User Profile Based Elements Tailoring:** Annotated terms and sentences are tailored according to users' profiles. When a user's profile is selected, terms and annotations matching the label profile are selected. Prebuilt users' folksonomies, when available, are consulted to refine terms and sentences with those ones more familiar to user's class.

4. **Natural Language Text Generation:** The filtered list of terms (user's vocabulary) and sentences (micro-plan text structure) are provided as ontologies to the NLG engine, finally producing the expected textual description, in the selected language.

5.2 Fedro platform System Architecture

FEDRO platform was basically implemented in Java technology, according to a MVC architectural pattern. It is characterized by a layered and multi-tier structure. The View Layer is represented by a friendly web user interface for filling in complex descriptions and desired target text features. The Control layer is a collection of Java servlets, involved in the dispatching and coordination phases of requests among Model modules. The Model layer, is the core of the authoring system consisting in a set of services, responsible for workflow orchestration of data source interactions and processing tasks. Text analysis is performed by a Python module implemented by the Natural Language Toolkit (NLTK, 2015) framework

and integrated with Java components by Jython API (JYT, 2015). As large lexical databases, WordNet (WDNET, 2015) and MultiWordNet (MWN, 2015) were employed for English and Italian languages, respectively. The Getty Vocabularies (AAT, 2015), available as LOD, were integrated as specific art domain ontology. Users folksonomies were integrated in the aspect of profiled users' lexical ontologies. Ontologies were managed by using API Jena (Jena, 2015). To generate new textual descriptions in natural language, the Natural OWL (Androutsopoulos, 2013; Galanis, 2008) framework was employed. This system offers a native support for English and Greek languages. So, it was extended to support Italian language.

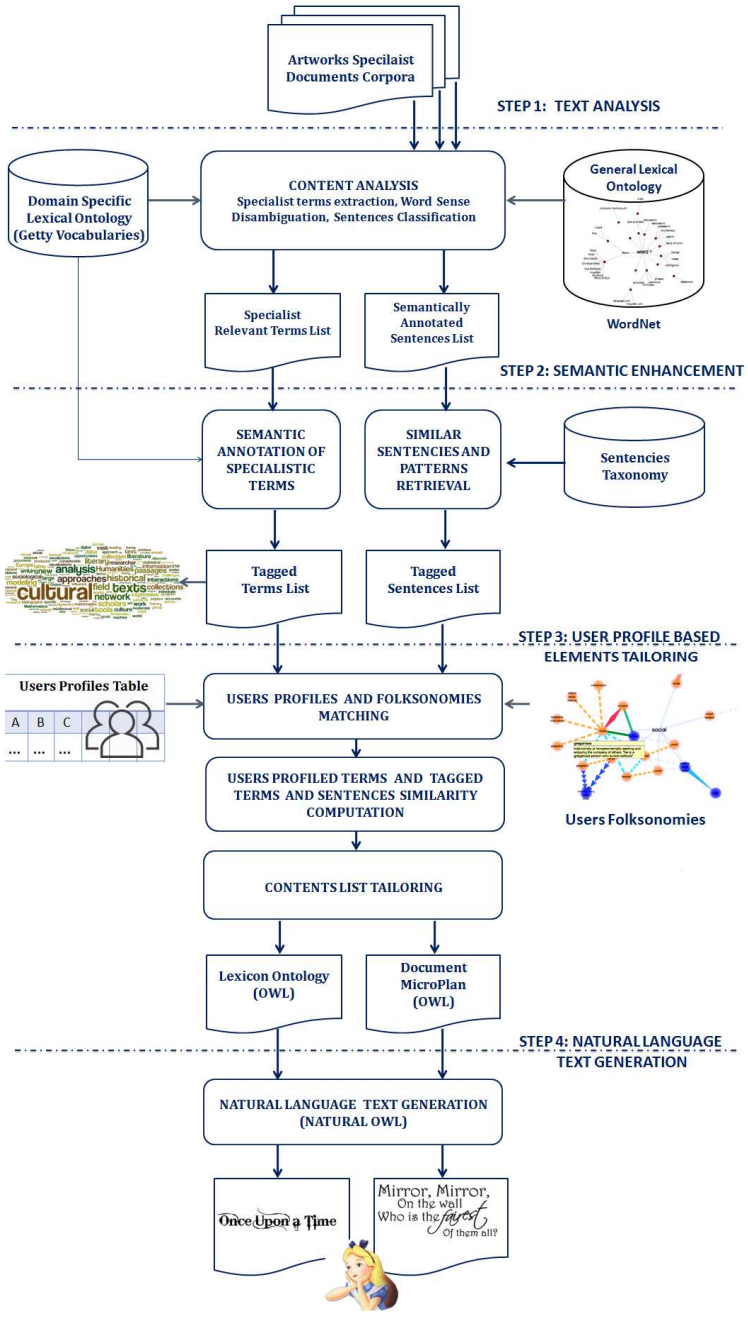

Figure 5: FEDRO system general architecture and processing flow.

5.3 Case Study and Preliminary Results: Il Bello o il Vero Exhibition

In Table 1, left and right columns show, respectively, the original complex text, provided by domain expert and the platform generated fable description.

Table 1: A comparison between input text and output simplified textual descriptions.

Input: Technical description (Domain Expert)	Output: Simplified fable description (Schoolchildren)
Carlotta D'Asburgo A Miramare is a model in gypsum and it was realized around 1914 by the sculptor Francesco Jerace. He was born in Polistena in 1853 and he died in Napoli in 1937. It comes from the collezione privata. The plaster model by Francesco Jerace represents The Empress of Mexico Charlotte of Habsburg in Miramare, where the marble was exhibited for the first time in 1999 at the Museo Civico di Castelnuovo. Charlotte is shown seated in front of the castle of Miramare in Trieste, with an eye toward the sea in expectation of the return of melancholy consort Maximilian of Hapsburg. Daughter of Leopold of Belgium, becomes, after the shooting of her husband, the heroine of a nineteenth-century romantic tradition of the last chapter.	Once upon a time, in a country named Italy, there was a man, whose name was Francesco Jerace. This man worked as a sculptor. A sculptor is an artist who is very able in working stones in beautiful shapes. What you are now looking at is named "Carlotta D'Asburgo A Miramare" Empress of Mexico, portraited when she looked out the balcony of her castle of Miramare, in Trieste, waiting for the return of her husband. This sculpture was made in 1914, in white gypsum and it is stored in another famous Castle, in Naples, in the Southern Italy. This castle is used as a museum. Its name is "Civic Museum of Castelnuovo", built in 1266. Local people call it as Maschio Angioino, from the name of French King Carlo d'Angiò, dominating Southern Italy about in XIII century.

Over than 200 sculptures were exhibited for about 7 months; different schoolchildren visits were scheduled in 15 different days, and each day a different group of 10 artworks fables was proposed by exploiting a mobile app. An appreciation questionnaire was submitted at the end of the visits, asking to assign a quality score in the range 1 – 4 (very much, enough, low, absolutely not) to specify the appreciation level in the visiting experience. Some of measured features were the comprehension and recording level, the clarity and the pleasantness of the proposed narrations. An overall improvement in the comprehension and appreciation level in the exhibition experience was recorded, but more robust and unbiased tests and metrics have to be performed to assess and improve the effectiveness of the proposed approach.

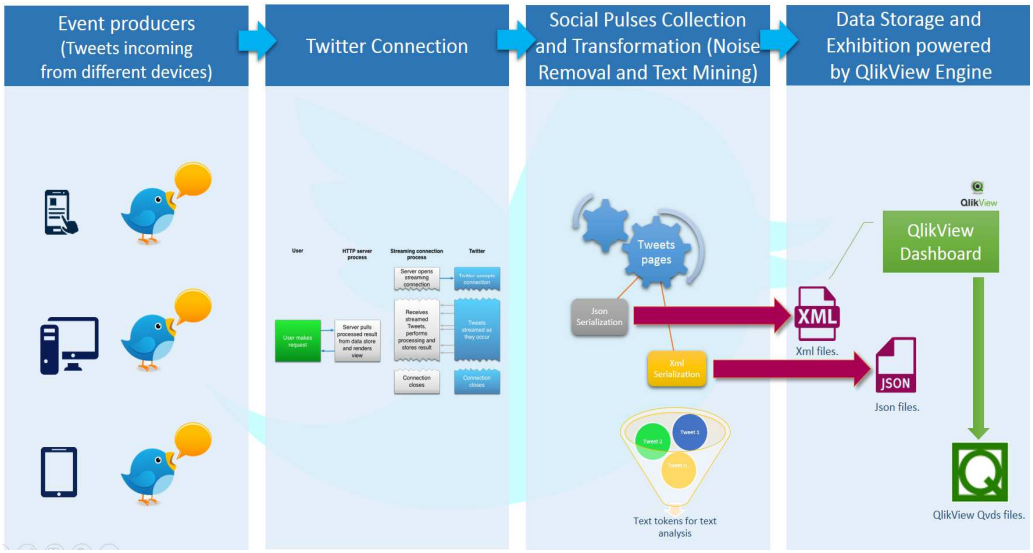

Figure 6: Fedro underlying knowledge resources pre processing

5.4 Textual generation using Natural OWL

In the following Figures (7, 8, 9) are showed three results obtained by processing the same set of input information introduced in Fedro. In this case, the Natural Owl tool (Androustopulous, 2013) was employed, after a customization process to extend the system for supporting lexicon and grammatic for Italian language textual generation.
It was extended for supporting Italian Language (native support for English and Greek);

- Inputs and Knowledge Resources employed:
  - A domain ontology populated, first manually and then automatically, by CH experts
  - A lexicon for italian language, manually defined (finite lexicon)
  - A grammar for text structure, as suggested by the native tool, filled in by human operators

- Preferences for user profiling (manual selection of the facts to be included, their priority and sequence order, a little selection for referencing expression (just reversing roles for entities (from active form to passive one).

- Three types of target users:
  - *CH expert, generic user (tourist), schoolchild*

- No semantic annotations for lexicon or grammar
- More than 500 artworks and cultural sites description were automatically produced for a real sculpture exhibition “Il Bello o il Vero” (http://www.ilbellooilvero.it), supported by Databenc District (http://www.databenc.it).

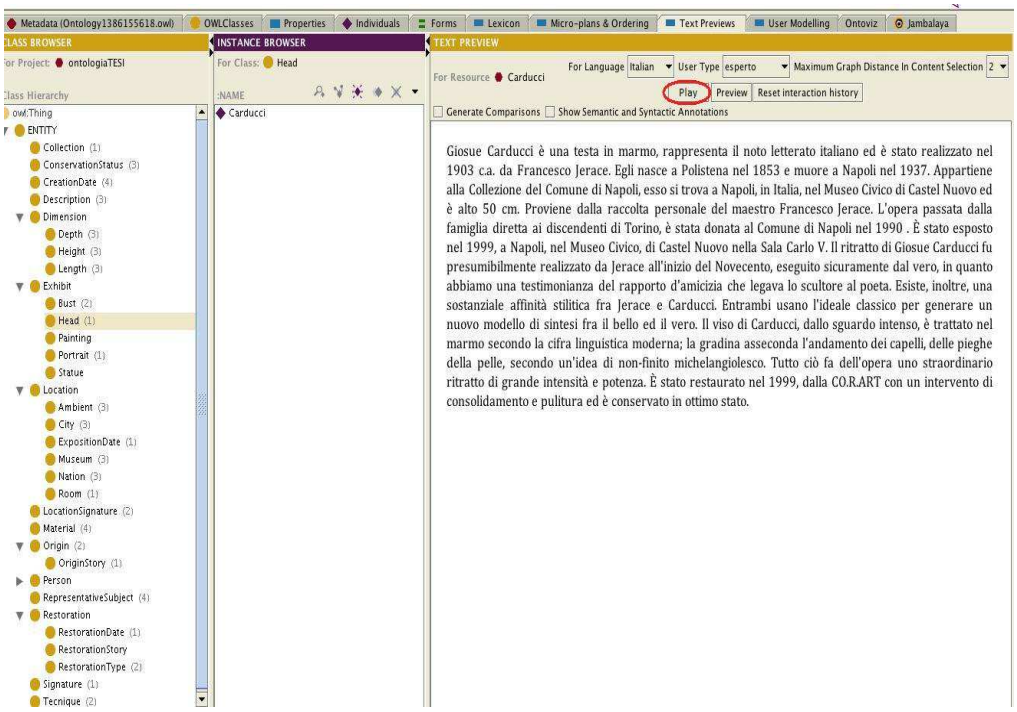

Figure 7: An example of text generation in Natural OWL (1)

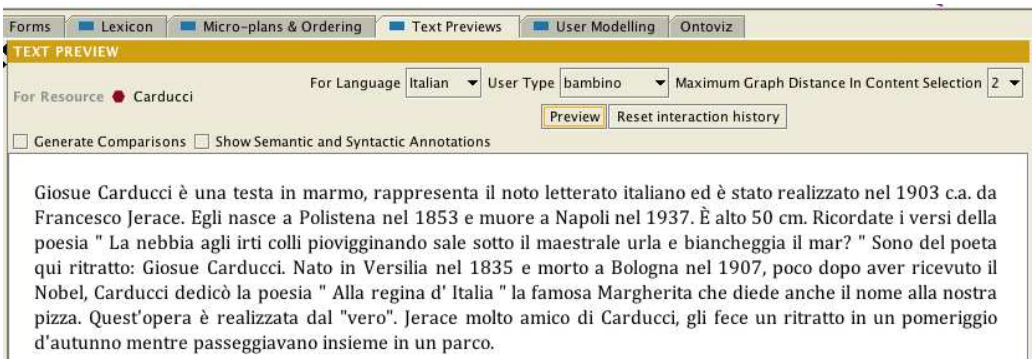

Figure 8: : An example of text generation in Natural OWL (2)

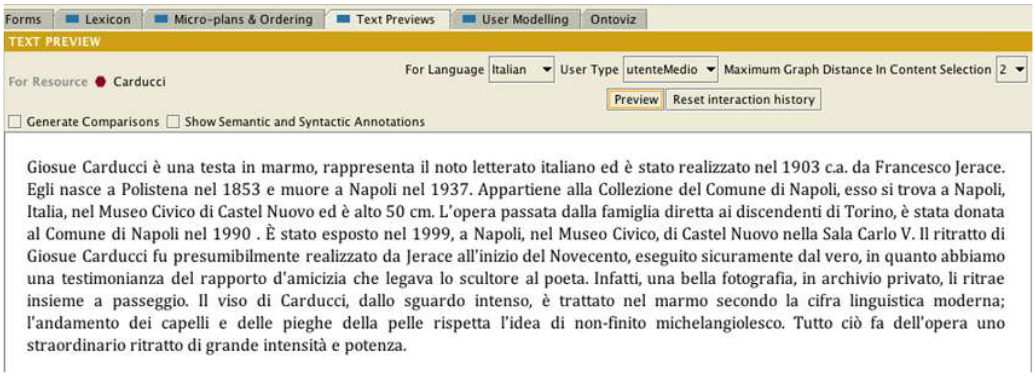

Figure 9: : An example of text generation in Natural OWL (3)

5.5 Comparison between Fedro and Natural Owl Textual generation

Analyzing the obtained results, we can observe the following if comparing Fedro platform with the Natural OWL system:

- Fedro provides an NLG Process Enhancement by exploiting domain and user profiled Knowledge Resources;

- Its working Conditions are represented by a set of Knowledge Resources semantically annotated;

- It leverages authors and communication experts from inserting manually grammars, user profiling rules and domain information, for every new entry
- It exploits:
    - User Contents Generation (UCG) mechanisms to automatically annotate Knowledge Resources

- Social Network Activities Monitoring, Extraction, Textual Features Analysis (Twitter, Facebook, Blogs, etc..)
- Folksonomies Exploration and Exploitation

- Authoritative Specialist Domain Knowledge (Domain Lexicons, e.g. Getty Art & Architecture Thesaurus (AAT))

- It relies on Paraphrasing Generation strategies

5.6 Results Analysis

Because of the lack of a standard ideal model of output, initially, the similarity between segments of text was measured by applying lexical matching techniques, good for finding semantically identical matches. Basing on experience, a semantic compliance threshold was set to a value of 85%. A test plan, performed on a 150 generated texts sample, produced a recall value of ~ 70%. Interesting but less unbiased indications about the effectiveness of the proposed approach, were provided by users' feedback at the end of their visit in the "Il Bello o il Vero" (http://www.ilbellooilvero.it) exhibition. An appreciation questionnaire was submitted at the end of the visits, asking to assign a quality score in the range 1 – 4 (very much, enough, low, absolutely not) to specify the appreciation level in the visiting experience. Some of measured features were the comprehension and recording level, the clarity and the pleasantness of the proposed narrations. An overall improvement in the comprehension and appreciation level in the exhibition experience was recorded, but more robust and unbiased tests and metrics have to be performed to assess and improve the effectiveness of the proposed approach.

5.7 Ouput Evaluation Metrics: Usability, Enjoyment and Naturalness Estimation

A number of trials have been performed to assess the behaviour, the users' enjoyment and, consequently, the usability and the utility of the proposed

application. A sample of about 100 visitors were logged during one of the events organized for celebrating the return to its original location for the Il Bello o il Vero exhibition. These participants were engaged at the entrance of the exhibition, before starting the visit and were given a 10-minute presentation about the infrastructure. According to the usability dimensions for a mobile application, as proposed by the literature in (Baharuddin, 2013), we investigated three of these dimensions to have an overall estimation for the proposed approach. We considered the following dimensions: simplicity (SIM), usefulness (USN) and enjoyment (satisfaction) (ENJ). For a better investigation, we added a further dimension, the naturalness of interaction (NAT).

Participants were asked to fill in a post-visit questionnaire. These questionnaires stimulated users to express their level of agreement with a set of statements, using a 10-point Likert scale, or to make choices between proposed options.

Table 2 summarizes results extracted from the users' answers, showing the most relevant questions related to the four dimensions of the usability considered and their average ratings.

The overall degree of satisfaction manifested by participants towards the proposed infrastructure was positive with an average rating of 8.86 (ENJ08).

Furthermore, the overall degree of perceived naturalness in the proposed interaction modality (NAT04) and the expected waiting time in the performing interaction (NAT03) were positive with an average rating of 7.89 (NAT03) and 8.45 (NAT04), respectively.

Multimedia features such as image-galleries (ENJ03), texts (ENJ04) and the quality for audio responses (ENJ05), were rated 7.44, 7.06 and 7.52, respectively. As for the usefulness dimension, users agreed that the application was useful overall (USN01, 7.83), facilitating to a certain degree the acquisition of a better knowledge (USN02, 7.66) and a deeper insight (USN03, 7.89) on the artwork on display.

Additionally, the analysis of the ease of use dimension pointed out that participants found the information access about the artworks quite easy (SIM01, 8.56) as well as the multimedia content browsing (SIM02, 7.81).

ID	Description	Value
SIM01	It was easy to interact with the exhibit artworks.	8.56
SIM02	It was easy to obtain useful multimedia contents.	7.81
SIM03	It was easy to navigate among the mobile App functionalities.	8.02
USN01	The infrastructure was overall useful during the visit.	7.83
USN02	Using the infrastructure was useful to gain knowledge about the exhibit artworks.	7.66
USN03	Using the infrastructure was useful to get a deeper insight on the museum themes.	7.89
ENJ01	I appreciate the mobile Assistant App GUI.	8.32
ENJ02	I appreciate the artworks detection metaphor.	8.45
ENJ03	I appreciate the image galleries.	7.44
ENJ04	I appreciate reading cultural information about exhibit artworks.	7.06
ENJ05	The quality of the sound was high.	7.52
ENJ06	Using the infrastructure contributed to increase my will to visit other art exhibitions.	8.09
ENJ07	Using the infrastructure positively contributed to the enjoyment of my visit.	8.87
ENJ08	I overall appreciated the infrastructure and the proposed approach.	8.86
NAT01	I appreciate listening cultural information about exhibit artworks.	8.98

NAT02	I appreciate the clearness of the spoken dialogue.	8.32
NAT03	The waiting time in the performing interaction attended my expectations.	7.89
NAT04	I appreciate the naturalness of the interaction with the environment	8.45

Table 2: Scoring Results from appreciation interview

Chapter 6

Conclusions and Future Directions

The NLP and NLG referring community as a whole is involved in a distributed effort to design and build resources for developing and evaluating solutions to new and existing NLP and NLG tasks.

Knowledge represents the basic core of our Cultural Heritage and Natural Language provides us with prime versatile means of construing experience at multiple levels of organization, storing and exchanging knowledge and information encoded as linguistic meaning.

Nowadays, the task of generating easily understandable information for people using natural language is being addressed by two fields which, independently until now, have researched the processes this task involves from different perspectives: the natural language generation (NLG) field and the knowledge and information extraction and retrieval (IER) field. The natural language generation field consists in the creation of texts which provide information contained in other kind of sources (numerical data, graphics, taxonomies and ontologies or even other texts), with the aim of making such texts indistinguishable, as far as possible, from those created by humans. The linguistic verbalization of segmented data is a young field still in its early stages, which has a solid formal base and whose real potential is still waiting to be uncovered. However, although nowadays there are relevant research results in this domain, most of them (theoretical ones aside) present simple use cases whose application in real problems seems somehow limited, since the complexity of descriptions for real problems in

terms of natural language is in general higher than what quantified sentences and the most complex linguistic descriptions currently provide.

To face with these issues, this doctoral thesis shows the research activity conducted with the aim of exploring and scientifically describing knowledge structure and organization involved in textual resources generation. Thus, a novel multidimensional model for the representation of conceptual knowledge, is proposed, in order to support and drive an effective and feasible processing workflow, producing strongly customized textual descriptions.

By exploiting paraphrases generation techniques and target users', applications and domains characterizations, a target-driven approach is proposed to generate automatically multiple instances for a textual description, sharing the same information core but differencing in the lexical and expressive form. A very extended case study, in the Cultural Heritage domain, is described to demonstrate the effectiveness and the feasibility of the proposed model and approach, thus providing the means for comparing and positioning this contribution with current state and future directions.

As further contribution of this work, an authoring platform supporting IoT smart applications in the CH domain was introduced. Most valuable contribution of this work should be identified in the novel proposed approach, mashing up top level information retrieval and text analysis strategies, with semantic processes, involving lexical and domain ontologies and users generated contents (by UGC systems).

The final aim is to automatically generate customized artworks descriptions (artworks descriptions in the case study provided) for different type of users and different type of target applications, feeding smart IoT cultural applications. Approaching people in a right and customized way could significantly enhance people's awareness about the effective value of their resources, thus creating social and economic opportunities for wellfare. In this perspective, from current literature, no other contributions are strongly focused on this issue or implement similar approaches for the same aim. A further novelty aspect is the communication strategy, based on the choice to generate simplified descriptions in the shape of fables, in order to make culture and art environment more charming for children audiences. Current version of the platform prototype is able to generate two different types of profiled textual artworks biographies (general descriptions and short fables). The adopted approach in the platform design promises to be scalable

and flexible enough to support extensions for other types of users. New and different lexical ontologies can be built and easily integrated in the system.

As future work, an interesting possibility is the exploitation of different top level text analysis and semantic based strategies and interactive users' experiences and evaluations, to improve the quality of generated textual descriptions. Finally, a related open issue, object of future investigations, is the absence of a standard human or automatic evaluation metrics to establish a text quality baseline.

At current time two basic metrics were adopted to quantify, at an early stage, results and improvements obtained by adopting this target driven approach against a flat generation approach. The first one was based on the evalution of the subjective pleasantness and the enjoyment degree recorded by users during a cultural exhibit visit, assisted by smart technologies provided by IoT service infrastructures. During the visit, users can enjoy descriptions of artworks by the means of a smart application and expressing, at the end of their tour, a feedback, by esplicitly scoring the provided textual resources, in terms of expressive clearness and compliance with their expectations. As a more objective metric, a reverse process was applied to automatically generated textual descriptions. It was aimed to to verify and assess a correct categorization and user profiling reverse extraction and the tracebility and coverance levels against the specified users' desiderata.

As conclusive observations, we can state that further refinements can be projected and applied to the proposed knowledge model and the related target-driven generation workflow. A more refined design, supporting Big Data and Business Intelligence processing system, could enhance the opportunity for a better exploitation of User Generated Contents, such providing a more precise semantic annotation for knowledge resources and a wide range of source resources to be exploited in the text construction processes. Reducing the gap between producers and consumers of textual resources, could only bring benefits in all technological solutions, supporting a more effective and natural interactions between human beings and machines. Finally, far from the pretension to be an exhaustive solution for the complex problem of a high quality customized text generation, this contribution has the value to pave the way for a not yet explored approach in facing these research area, thus underlining the importance of considering the various facet of knowledge influencing a good communication product, just like a text, as a whole and not as a fragmented contribution. The results

of a pilot generation study showed this model is feasible and the results immediately useful.

Bibliography

Androutsopoulos, I., V. Kokkinaki, V., Dimitromanolaki, A., Calder, J., Oberlander, J., Not, E. 2001. Generating multilingual personalized descriptions of museum exhibits – the m-piro project, in: Proceedings of the 29th Conference on Computer Applications and Quantitative Methods in Archaeology, Gotland, Sweden, 2001.

Androutsopoulos, I. and Malakasiotis, P. 2010. "A Survey of Paraphrasing and Textual Entailment Methods". Journal of Artificial Intelligence Research, 38:135-187.

Androutsopoulos, I., Lampouras, G., Galanis, D. 2013. “Generating Natural Language Descriptions from OWL Ontologies: the NaturalOWL System”, 2013, Journal of Artificial Intelligence Research 48, pp.671-715.

Balbi, S. 2012. Beyond the curse of multidimensionality: high dimensional clustering in text mining, 2012, Statistica Applicata, Vol. 22, F. 1, pag.53-63.

Bordoni, L., Ardissono, L., Semeraro, G. et al., "The contribution of AI to enhance understanding of Cultural Heritage", 2013, Intelligenza Artificiale 7 (2013) 101–112. DOI 10.3233/IA-130052. IOS Press. 101

Calzolari N., Lenci A. 2004. Linguistica Computazionale - Strumenti e risorse per il Trattamento Automatico della Lingua. Mondo Digitale, n. 2 - giugno 2004, 56-69.

Cantador, I., Konstas, I., Jose, J.M. 2011. Categorising social tags to improve folksonomy-based recommendations. Web Semant. 9, 1 (March 2011), 1-15.

Chomsky, N. 1967. A Review of B. F. Skinner’s Verbal Behavior. In L. A. J. A. M. S. Miron. (Ed.), Readings in the psychology of language (pp. 142 - 143). Englewood Cliffs, N.J.: Prentice-Hall psychology.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F.M., 2013, Recognizing Textual Entailment: Models and Applications. Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 23), 2013, xx+200 pp;

Crisp-bright, A. K. 2010a. The Effects of Domain and Type of Knowledge on Category-Based Inductive Reasoning. Memory, 67-72.

Crisp-bright, A. K. 2010b. Knowledge Selection in Category-Based Inductive Reasoning. Cognitive Science. Durham University.

Eysenck, M. W., & Keane, M. T. 2010. Cognitive Psychology A Student's Handbook, (6th ed.). East Sussex: Psychology Press. Retrieved September 3, 2010.

Feldman, R. 2002. Epistemology. Prentice Hall.

Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. Bradford Books.

Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C. and Silvello, G. eds., 2016. Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings (Vol. 9626). Springer.

Fodors, J. A. 1975. The Language of Thought (p. 214). Cambridge: Harvard University Press.

Galanis, D., Karakatsiotis, Androutsopoulos, G. 2008. "How to install NaturalOWL", http://www.ling.helsinki.fi/kit/2008s/clt310gen/docs/NaturalOWL-README.pdf.

Guha, R.V. and Lenat, D.B. 1990. Cyc: a mid-term report. AI Magazine, 11(3).

Hobbes, T. 1651. Leviathan. Oxford: Clarendon Press.

Hobbes, T. 1969. Elements of Law, Natural and Political (p. 186). Routledge.

Kaushik, A. and Naithani, S., 2016. A Comprehensive Study of Text Mining Approach. International Journal of Computer Science and Network Security (IJCSNS), 16(2), p.69.

LoBue, P. and Yates, A., 2011, June. Types of common-sense knowledge needed for recognizing textual entailment. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 329-334). Association for Computational Linguistics.

Malakasiotis, P. 2011. "Paraphrase and textual entailment recognition and generation" (in English), PhD thesis, Department of Informatics, Athens University of Economics and Business, 2011.

Marulli, F., Pareschi, R., Baldacci, D. 2016. The Internet of Speaking Things and its Applications to Cultural Heritage. In Proceedings of the International Conference on Internet of Things and Big Data (IOTBD2016), SCITEPRESS, 2016.

Ramirez, C., Valdes, B. (2012). A General Knowledge Representation Model of Concepts, Advances in Knowledge Representation, Dr. Carlos Ramirez (Ed.), ISBN: 978-953-51-0597-8, InTech.

Ramos-Soto, A., Bugarín, A., Barro, S. 2016. On the role of linguistic descriptions of data in the building of natural language generation systems, Fuzzy Sets and Systems, Volume 285, 15 February 2016, Pages 31-51, ISSN 0165-0114.

Semeraro, G., Lops, P., De Gemmis, M., Musto, C., Narducci, F. 2012. A folksonomy-based recommender system for personalized access to digital artworks. J. Comput. Cult. Herit. 5, 3, Article 11 (October 2012), 22 pages.

Stork, D. G. 1999. The OpenMind Initiative. IEEE Expert Systems and Their Applications, 14(3):19–20.

Vallifuoco, L., Marulli, F. 2016. The Imitation Game in Cultural Heritage: A Human-like Interaction Driven Approach for Supporting Art Recreation. In 5th EAI International Conference: ArtsIT, Interactivity & Game Creation (ArtsIT2016), Springer, 2016.

Vygotsky, L. (1986). Thought and Language. (A. Kozulin, Ed.). New York, USA: MIT Press.

W.R. Swartout, A digitalis therapy advisor with explanations, Tech. rep., Massachusetts Institute of Technology, Cambridge, MA, USA, 1977.

R. Kittredge, A. Polguère, E. Goldberg, Synthesizing weather forecasts from formated data, in: Proceedings of the 11th Conferenceon Computational Linguistics, COLING ’86, Association for Computational Linguistics, Stroudsburg, PA, USA, 1986, pp.563–565.

M. Boyer, G. Lapalme, Generating paraphrases from meaning-text semantic networks, Comput. Intell. 1(1) (1985) 103–117, http://dx.doi.org/10.1111/j.1467-8640.1985.tb00063.x.

E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, 2000.

C. Mellish, D. Scott, L. Cahill, D. Paiva, R. Evans, M. Reape, A reference architecture for natural language generation systems, Nat. Lang. Eng. 12 (2006) 1–34, http://dx.doi.org/10.1017/S1351324906004104.

E. Reiter, An architecture for data-to-text systems, in: S. Busemann (Ed.), Proceedings of the 11th European Workshop on Natural Language Generation, 2007, pp.97–104.

R.R. Yager, A new approach to the summarization of data, Inf. Sci. 28(1) (1982) 69–86, http://dx.doi.org/10.1016/0020-0255(82)90033-0.

R.R. Yager, K.M. Ford, A.J. Cañas, An approach to the linguistic summarization of data, in: B. Bouchon-Meunier, R.R. Yager, L.A. Zadeh (Eds.), Uncertainty in Knowledge Bases, Proceedings of the 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 90, Paris, France, July 2–6, 1990, in: Lecture Notes in Computer Science, vol.521, Springer, 1990, pp.456–468.

L.A. Zadeh, Fuzzy logic = computing with words, IEEE Trans. Fuzzy Syst. 4(2) (1996) 103–111.

L.A. Zadeh, From computing with numbers to computing with words: from manipulation of measurements to manipulation of perceptions, in: Intelligent Systems and Soft Computing: Prospects, Tools and Applications, Springer-Verlag, 2000, pp.3–40.

L.A. Zadeh, A new direction in AI— toward a computational theory of perceptions, in: B. Reusch (Ed.), Computational Intelligence. Theory and Applications, in: Lecture Notes in Computer Science, vol.2206, Springer, Berlin, Heidelberg, 2001, p.628.

A. Ramos-Soto, A. Bugarín, S. Barro, J. Taboada, Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data, IEEE Trans. Fuzzy Syst. 23(1) (2015) 44–57, http://dx.doi.org/10.1109/TFUZZ.2014.2328011.

J.A. Bateman, 2001, Natural language generation: an introduction and open-ended review of the state of the art, http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/info-pages/nlg/ATG01/node1.html, 2001.

E. Goldberg, N. Driedger, R. Kittredge, 1994, Using natural-language processing to produce weather forecasts, IEEE Expert 9(2) (1994) 45–53, http://dx.doi.org/10.1109/64.294135.

J. Coch, Interactive generation and knowledge administration in multimeteo, in: Proceedings of the Ninth International Workshop on Natural Language Generation, Niagara-on-the-lake, Ontario, Canada, 1998, pp.300–303, software demonstration.

J. Coch, R. David, J. Magnoler, 1995, Quality test for a mail generation system, in: Proceedings of Linguistic Engineering ’95, Montpellier, France, 1995, pp.435–443.

F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, C. Sykes, 2009, Automatic generation of textual summaries from neonatal intensive care data, Artif. Intell. 173(7–8) (2009) 789–816, http://dx.doi.org/10.1016/j.artint.2008.12.002.

M. White, T. Caldwell, Exemplars: a practical, extensible framework for dynamic text generation, in: 9th INLG, Niagara-on-the-Lake, Ontario, 1998, pp.266–275.

S. Busemann, H. Horacek, 1997, Generating air-quality reports from environmental data, in: S. Busemann, T. Becker, W. Finkler (Eds.), DFKI Workshop on Natural Language Generation, 1997, DFKI Document D-97-06.

J.A. Bateman, M. Zock, Nlg systems wiki [cited March 2015], http://www.nlg-wiki.org/systems/.

E. Reiter, R. Dale, 1997, Building applied natural language generation systems, Nat. Lang. Eng. 3 (1997) 57–87.

K. Van Deemter, E. Krahmer, M. Theune, 2005, Real versus template-based natural language generation: a false opposition?, Comput. Linguist. 31(1) (2005) 15–24, http://dx.doi.org/10.1162/0891201053630291.

EAGLES Project, 1996, Natural Language Generation, http://www.ilc.cnr.it/EAGLES96/rep2/node35.html

AAT, Getty Vocabularies, 2015, http://www.getty.edu/research/tools/

Jena, Apache JENA API, 2015, https://jena.apache.org/

JYT, Jython: Python for the Java Platform, 2015, http://www.jython.org/

MWN, MultiWordNet, 2015, http://multiwordnet.fbk.eu/

NLTK, Natural Language Toolkit, 2015, http://www.nltk.org/

WDNET, WordNet, a lexical database for English, 2015,https://wordnet.princeton.edu/

ILC-CNR (1996), EAGLES Prj, http://www.ilc.cnr.it/EAGLES96/rep2/node35.html

Italian NLP Research Group- ILC- CNR Pisa**,** 2015, T2K toolkit

Stanford Computational Linguistic Group,
https://linguistics.stanford.edu/research/computational-linguistics