Multimed Tools Appl (2017) 76:10389–10406
DOI 10.1007/s11042-016-3628-3

A novel approach for automatic text analysis and generation for the cultural heritage domain

Francesco Piccialli[1] · **Fiammetta Marulli**[1] · **Angelo Chianese**[1]

Received: 5 May 2016 / Revised: 15 May 2016 / Accepted: 17 May 2016 /
Published online: 8 June 2016

Abstract Knowledge is information that has been contextualised in a certain domain, to be used or applied. It 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. The natural language generation field consists in the creation of texts providing 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. This paper presents a research activity focused on 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. Furthermore, a real case study in the Cultural Heritage domain is described to demonstrate the effectiveness and the feasibility of the proposed model and approach.

Keywords Natural language generation · Cultural heritage · Text generation · Knowledge modeling

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.

✉ Francesco Piccialli
francesco.piccialli@unina.it

[1] University of Naples Federico II, Via Claudio, Naples, Italy

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 [22], but it is basically constituted by static components, called concepts or facts, and dynamic components, called skills, abilities, procedures, actions and so on. 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. 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. 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 [17], and what its components are [10]. The issue 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 [18]. 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 [9]. In [3, 4], 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 pre-compiled lexicon for English and Greek languages, and a flexible grammar, 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 this kind of systems and exploited approaches is the lack of a control strategy able to orchestrate and coordinate interventions of available knowledge resources into the steps of processes.

To face with these issues, this paper presents 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. The proposed multidimensional model enhances natural language generation processes, by strongly focusing on different textual generations, based on the same information sources. By exploiting paraphrases generation techniques, a target-driven approach is proposed and adopted. In addition, an information system prototype, characterized for Cultural Heritage domain and implementing the aforementioned workflow and approach, is presented. A real case study is described to demonstrate the effectiveness of the proposed model and approach. The rest of the paper is organized as follows: Section 2 presents the related work, Section 3 discusses the NLG techniques, Section 4 presents the proposed multi-dimensional knowledge model, Section 5 discusses the case study in the CH domain. Finally, Section 6 concludes the paper with some considerations.

2 Related work

In this section, a review of the basic concepts behind knowledge representation and the main types of knowledge representation models is presented.

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 [13], he stated that knowledge is the evidence of truth, which must have four properties [14]:

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) [11]. 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

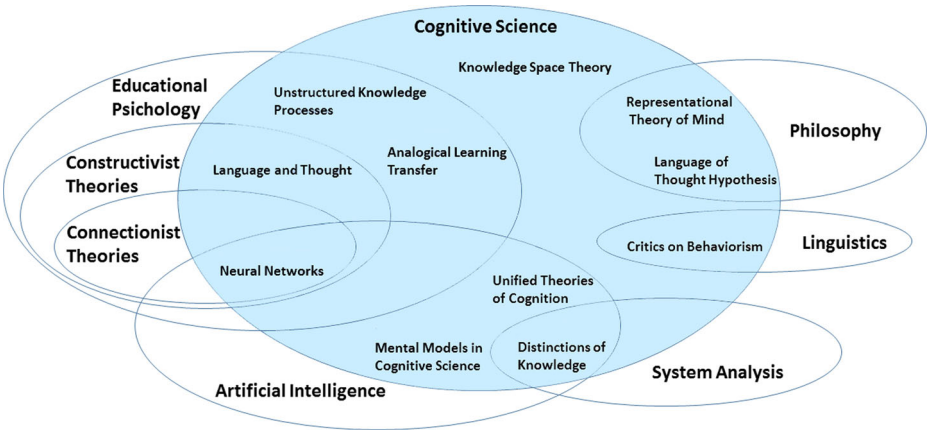


Fig. 1 Multiple approaches to knowledge representation from different disciplines

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. 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 different approaches to Knowledge Representation from different disciplines, as detailed in [22].

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:

1. Knowledge is composed of basic units, referred to as concepts. The approaches for representing those basic structures will be discussed in the following sections.
2. Concepts have associations or relations to other concepts. The debate on associations is about the representational aspects regarding to the following issues:

   What information should an association contain
   What elements should be used to describe such information i.e., type, directionality, name, intention, extension, among others.

3. Associations and concepts build dynamic structures which tend to become stable through time.

Community is agreed 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.

3 Natural language generation techniques

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. 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.

3.1 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. Authors in [23] argue 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 split into a number of sub stages or modules which then can be further specified. In this context they present a sequential pipeline architecture for NLG divided into general three stages (see Fig. 2):

1. Text planning
2. Document planning
3. Surface realization

This architecture is then further decomposed into six basic activities (see 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

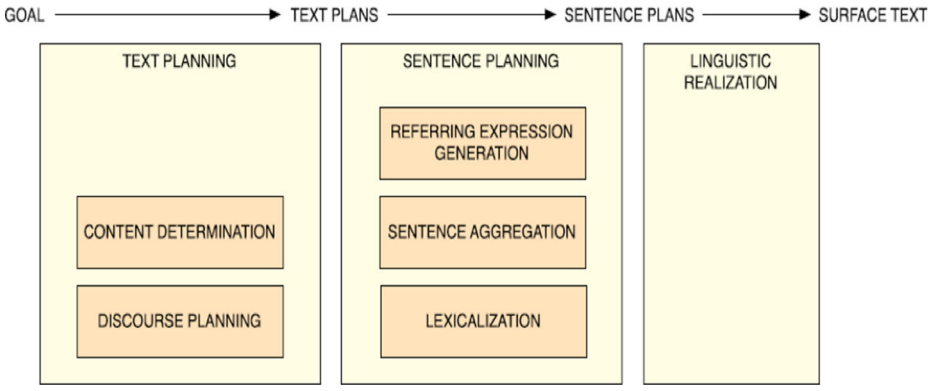


Fig. 2 A general schema for natural language generation process

system input. Those messages are the data objects used in the subsequent tasks. In general terms, the message creation process consists in filtering and summarizing the input data.

- **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.
- **Sentence aggregation**: This process groups several messages together in a sentence. This task is not always necessary (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.
- **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 characterized as a discrimination activity, in which the system needs to provide enough information to differentiate one domain entity from the rest.
- **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.

3.2 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 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, and so on.

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.

4 A multidimensional representation model for knowledge supporting user profiling and domain driven text generation

Natural Language Generation (NLG) Systems, applied in CH domain, are investigated in [4]. 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 [12], 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

After evaluating current literature and existing systems, in the Natural Language Generation field, we just conclude that only few systems are available for automatic generation of long and fluent textual descriptions, nonetheless a fewer number effectively considers a deep characterization for target users, target domain and target applications. As a matter of fact, only a simple user model is taken into account during textual verbalization of knowledge structures, thus only providing the opportunity to select facts that would be included into the descriptions and some preferred referring expressions (in the most cases, simple alternative micro-plans are available, as active and passive form of the same sentence, e.g.). But no trace is available of a deeper exploitation of users' profiling information or the specific domain or target application. The proposed model aims to keep together all this aspects of knowledge sources, that could be conveniently exploited to generate better textual descriptions, in terms of expressiveness and personalization, taking also into account the background and the age, as an example, of the target audience.

4.2 Problem formulation and proposed solution formulation

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 enjoyment)

thus 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:

– Identification and Formalization of a representation model for knowledge able to support user-driven and domain-driven automatic text analysis and generation. In particular, a reinforcement of Textual Entailment Recognition and Paraphrasing Generation Processes
– Automatic Annotation of Knowledge and Linguistic Resources (in a User and Specific Domain Perspective) by Textual Big Data Acquisition and Processing:
– Lexical resources;
– Users' folksonomies and Taxonomies of Users Common Linguistic Deviations (in wide spread syntax mistakes (solecisms), barbarisms (forcing usage of foreign terms in the current language), linguistic blunders, etc.).

4.3 The multidimensional knowledge representation model

Many 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 define a High Level Abstract Conceptual Model, composed by a set of entities (called Hard Bricks and Software Bricks, as described in the following of the section), a set of properties and a basic set of relations among the entities. In order to be machine readable and interoperable, the abstract conceptual model, is remapped on SKOS (Simple Knowledge Organization System), an RDF Schema Vocabulary. So, the model is composed at a glance by An Abstract Conceptual Level, describing Concepts, Properties and Relationships, remapped over an RDF Schema (adopting SKOS (Simple Knowledge Organization System) Vocabulary). An RDFXML was adopted to express and serialize the SKOS graph as an XML document.

4.3.1 Model components

The basics constitutive elements of the proposed model are listed below:

– A set of conceptual bricks CB = {HB,SB}:

HardBricks HB = {Artefact (AF), Artefact Plan (AFP), Knowledge Dimension (KD), Target Requirements Set (TRS)}; → mapped over SKOS <Concept> category HardBricks represents the main conceptual entities for the model.

SoftBricks (SB)= {res_id, res_name, res_date, res_author, res_uri, res_tag}

SoftBricks represent properties and tags for HB. They are mapped over SKOS labels and notation syntax. Property *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

    Meronymy 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

    annotating HB for NLP Process

4.3.2 The model structure

HardBricks and SoftBricks interact by defined relations to compose the whole model structure, as described in the following sections. The whole model offers a view as a whole of the entire Artefact Creation Workflow, in which all the available knowledge resources can be conveniently exploited for a complex and tailored textual generation. The following list provides the high level composition for this processing pipeline, based on the proposed model structure. At the highest level we have an HardBrick, called Artefact.

– HB_AF: Artefact: a container element bridging Target Requirements Set with Knowledge Resources; it is composed of:

  A set of properties
  An Artefact plan
  A Target Requirements Set

– HB_AFP: Artefact Plan: a collection of Knowledge Dimensions;
– HB_TRS: Target Requirements Set: a set of requirements specified to customize text generation process and adopted resource.

Its composition depends on the semantic annotation process for knowledge resources; typical elements are:

– Target language
– Target domain
– Target user
– Target application

4.3.3 Basic knowledge dimensions

The basic knowledge dimensions proposed in the model, are represented by Domain Knowledge, Basic Language Lexicon and Basic Grammatical.

These three dimensions are detailed below:

1. 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)

2. 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.
3. 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)

4.3.4 Enrichment knowledge dimensions

The enrichment knowledge dimensions proposed in the model, are represented by Specific Domain Lexicon, Target Audience Model, Target Application and Basic Grammatical.

These three dimensions are detailed below:

1. HB_KD4: Domain Lexicon
   - Aim: describing dictionaries or semantic lexicons for specialist and technical terms for considered domain

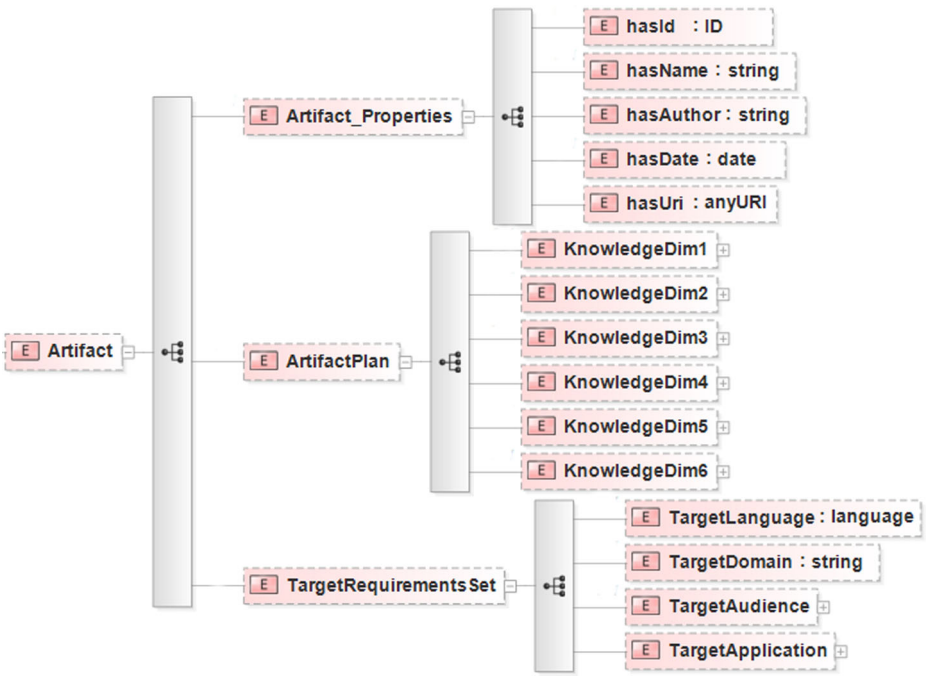


Fig. 3 Multidimensional model

- Author: domain experts
- Short Description: Specific Vocabularies or Semantic Lexicon for interest Domain)

2. HB_KD5: Target Audience (User) Model

   - Aim: taking into account more meaningful features for target audience characterization: age, interest or skills 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).

3. HB_KD6: Target Application

   - Aim: taking into account more constrained features for target application: length (for users enjoyment), time duration (for Text-To-Speech 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 shows a graphical view for the proposed model.

5 A cultural heritage application

In this section we illustrate a case study in the CH domain, in order to demonstrate effectiveness for the proposed multi-dimensional model of knowledge. We will explain the authoring platform FEDRO [19], as part of an intelligent infrastructure developed into DATABENC District [2, 6, 7], to support cultural exhibition of talking artworks [8], among which that one called *The Beauty or the Truth*, exhibiting sculptures and held in the Southern Italy.

5.1 Fedro platform system architecture

A general overview of FEDRO platform architecture and processing flow is shown in Fig. 4. 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:

- Text analysis: typical text analysis and summarization techniques are applied to input documents corpora; terms and sentences are extracted and disambiguated by the support of lexical and domain ontologies The output is represented by lists of relevant terms and sentences.
- 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 sentences taxonomies are employed to select new simplified sentences, according to semantic similarity criteria.

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

Furthermore, annotating resources, according to various users' behaviour is a quite difficult and time consuming activity. In such a perspective, in the proposed approach was exploited users' activities on social networks, which provides a large amount of desired information. In Fig. 5, the adopted processing work-flow for Big Data from social network activities (Twitter) is showed. As stated in [5], the combination of Big Data technologies and traditional machine learning algorithms has generated new and interesting challenges in other areas as social media and social network. In order to create users' profiled vocabularies, a case study, was considered. It aimed to perform a type of analysis, mainly focused on the messages exchanged by people from Campania region, in Italy, in order to analyze messages and opinion evenly related to the cited exhibition. Filtering criteria, such as the source location of emitting devices and the original language selected by users in their message settings, were applied in order to apply a first level filtering on the incoming Big and Noisy Social Pulses. At a glance, the service architecture system implementing the users' vocabulary mining process is depicted in Fig. 5.

5.2 Fedro platform implementation details

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

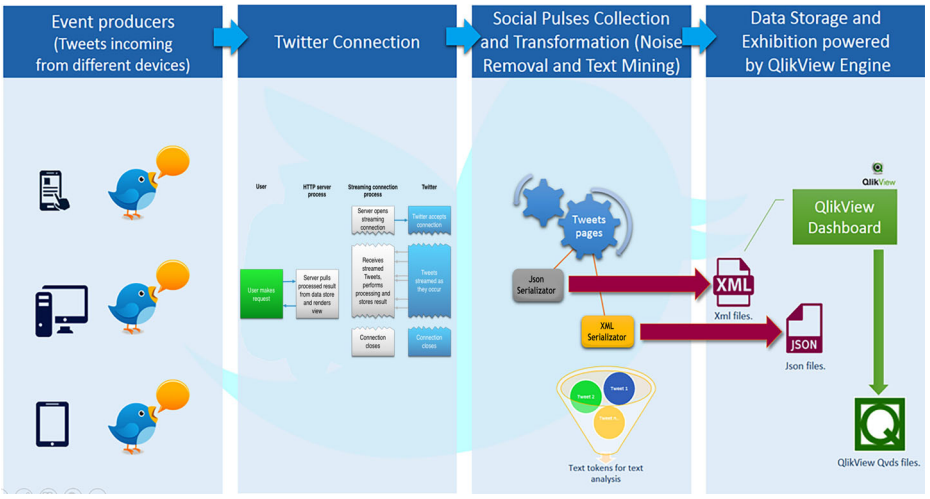


Fig. 4 Fedro underlying knowledge resources pre processing

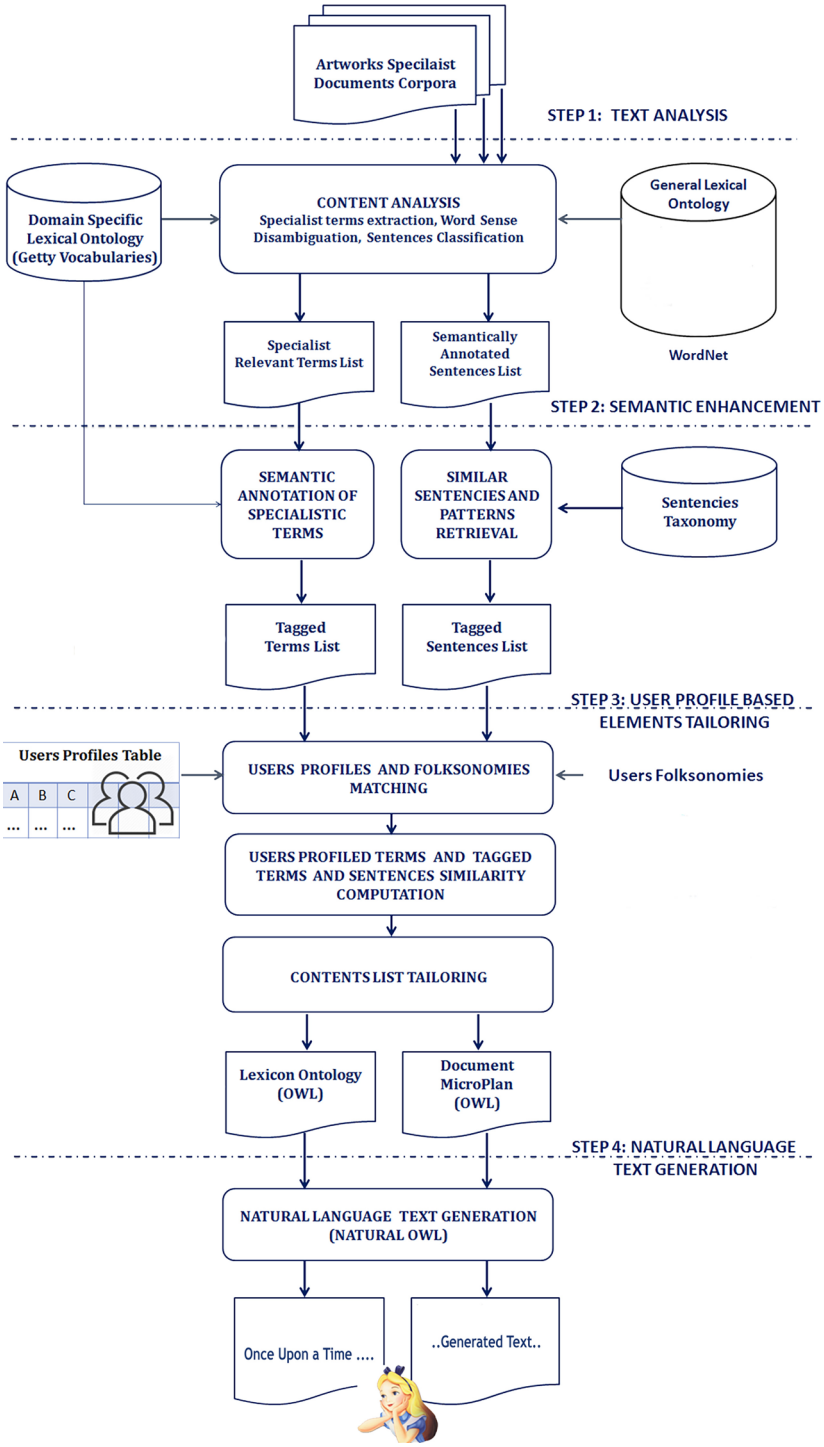


Fig. 5 FEDRO system general architecture and processing flow

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 [21] framework and integrated with Java components by Jython API [16]. As large lexical databases, WordNet [24] and MultiWordNet [20] were employed for English and Italian languages, respectively. The Getty Vocabularies [1], 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 [15]. To generate new textual descriptions in natural language, the Natural OWL [4, 12] framework was employed. This system offers a native support for English and Greek languages. So, it was extended to support Italian language.

5.3 The case study and preliminary results: *The Beauty or the Truth* art exhibition

Fedro was employed to generate textual descriptions for different type of users. Just by exploiting of specific grammars, as suggested by experts working in the education and communication fields, textual artworks descriptions for children audience, are generated in the form of little story or fables, telling about the artworks, its author and the place guesting the exhibition. In such a way, a description is generated by exploiting the same information core used for other users' type descriptions but adopting different referring expressions. Table 1 shows a sample text automatically generated by adopting the described approach. Left and right columns show, respectively, the original complex text, provided or generated by/for a domain expert and the platform generated description, presented as a fable.

5.4 Results analysis

During the art exhibition *The Beauty or the Truth*, over than 250 sculptures were exhibited for about 7 months (see Fig. 6); different schoolchildren visits were scheduled in 15

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, portrayed 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'Angi, dominating Southern Italy about in XIII century.

Fig. 6 The Beauty or the Truth art exhibition rooms

different days, and each day a different group of 10 artworks fables was proposed by exploiting a mobile application. 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 about 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 *The Beauty or the Truth* (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.

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 *The Beauty or the Truth* 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, 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

Table 2 Scoring results from appreciation interview

ID	Question	Average rating
SIM01	It was easy to interact with the exhibit .	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

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).

6 Conclusions and future directions

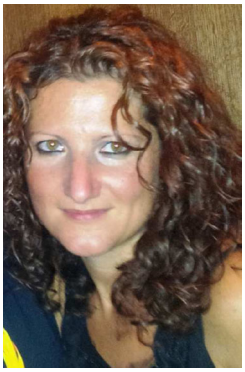
Knowledge represents the basic core of our Cultural Heritage and Natural Language provides us with prime versatile means of construing experience at multiple levels of organization, storing and exchanging knowledge and information encoded as linguistic meaning. Nowadays, the task of generating easily 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. This paper 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 work-flow, producing strongly customized textual descriptions. 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 work-flow. 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.

References

1. AAT, Getty Vocabularies, 2015, http://www.getty.edu/research/tools/
2. Amato F, Chianese A, Mazzeo A, Moscato V, Picariello A, Piccialli F (2013) The talking museum project. Procedia Comput Sci 21:114–121
3. Androutsopoulos I, 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
4. Androutsopoulos I, Lampouras G, Galanis D (2013) Generating natural language descriptions from OWL ontologies: the naturalOWL system. J Artif Intell Res 48:671–715
5. Bello-Orgaz G, Jung JJ, Camacho D (2016) Social big data: recent achievements and new challenges. Inf Fusion 28:45–59
6. Chianese A, Piccialli F, Valente I (2015) Smart environments and cultural heritage: a novel approach to create intelligent cultural spaces. J Locat Based Serv:209–334
7. Chianese A, Piccialli F (2016) A smart system to manage the context evolution in the cultural heritage domain. Comput Electr Eng. doi:10.1016/j.compeleceng.2016.02.008
8. Chianese A, Marulli F, Piccialli F, Benedusi P, Jung JE (2016) An associative engines based approach supporting collaborative analytics in the internet of cultural things, future generation computing systems. Elsevier. doi:10.1016/j.future.2016.04.015
9. EAGLES Project, Natural Language Generation, http://www.ilc.cnr.it/EAGLES96/rep2/node35.htm, 1996
10. Feldman R (2002) Epistemology. Prentice Hall
11. Fodors JA (1975) The language of thought. Harvard University Press, Cambridge, p 214
12. Galanis D, Karakatsiotis, Androutsopoulos G (2008) How to install NaturalOWL, http://www.ling.helsinki.fi/kit/2008s/clt310gen/docs/NaturalOWL-README.pdf
13. Hobbes T (1651) Leviathan. Clarendon Press, Oxford
14. Hobbes T (1969) Elements of law, natural and political. Routledge, p 186
15. Jena, Apache JENA API, 2015, https://jena.apache.org/
16. JYT, Jython: Python for the Java Platform, 2015, http://www.jython.org/
17. LoBue P, Yates A (2012) 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, pp 329–334
18. Malakasiotis P (2011) Paraphrase and textual entailment recognition and generation. PhD thesis, Department of Informatics, Athens University of Economics and Business
19. Marulli F (2015) 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
20. MWN, MultiWordNet, 2015, http://multiwordnet.fbk.eu/
21. NLTK, Natural Language Toolkit, 2015, http://www.nltk.org/
22. Ramirez C, Valdes B A general knowledge representation model of concepts. In: Ramirez C (ed) Advances in knowledge representation. ISBN: 978-953-51-0597-8, InTech
23. Reiter E, Dale R (1997) Building applied natural language generation systems. Nat Lang Eng 3:57–87
24. WDNET, WordNet, a lexical database for English, 2015, https://wordnet.princeton.edu/

Francesco Piccialli is a researcher a in University of Naples "Federico II", Italy. His research topics are Smart Environment design, Data Mining on Internet of Thing systems with application in the Cultural Heritage domain.



Fiammata Marulli is a Ph.D student in Computer Science in University of Naples Federico II, Italy. Her research topics are text analysis, ontology-based techniques, Business Intelligence tools and applications.

Angelo Chianese is a Full Professor in University of Naples "Federico II" Italy. His research topics are web semantic techniques, multimedia recommender systems, Internet of Things with application in the Cultural Heritage domain. He is the president of the High Technology District for Cultural Heritage in Campania Region, Italy.