# Semantic Annotation to Support Decision-Making

Francesca Parisi
Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica
Università della Calabria
Rende, Italy
francesca.parisi@dimes.unical.it

***Abstract*—Domain-specific organization management processes require particular operations and information exchange. A large amount of information concerning these tasks has to be reported and capitalized. Usually, people working in big organizations use Information Systems (IS) and other collaborative software producing a large number of information sources (data, documents, e-mails, images, etc.). The methodology presented in this paper concentrates on the contribution of semantic knowledge within textual information to improve processes carried out in domain-specific organizations. For example, activity logs, technical reports or contracts represent a part of structured textual knowledge. E-mails or texts describing activities are a part of unstructured (or semi-structured) knowledge and often contain important information to support decision-makers. Therefore, people involved in these processes need to analyze such documents to enrich their knowledge about said processes. Hence, the present methodology concerns the use of textual analysis approaches in order to evaluate their contribution to expert's activities. The experts are the human resources with specific competences involved in processes that require particular analyses. The methodology proposed is focused on an automatic semantic annotation technique considering the most important entities and their significance within specific domains. It works on the different kinds of textual knowledge (structured and unstructured) and allows to construct the most representative document classification label. In particular, the proposed method uses results of semantic annotation techniques to optimize document classification and, consequently, to support the decision-making process. This methodological proposal is the result of a work experience within a company operating in the energy sector.**

***Keywords-semantic annotation; document classification; corpus annotation; decision-making support.***

## I. INTRODUCTION

The proposal put forth in this paper was developed and refined during a case study carried out in the energy field and aims at analyzing the processes of any type of domain-specific company (e.g., companies working in the energy, manufacturing, industrial sector etc.) that needs to organize their working group, maximize task efficiency and minimize production costs. This kind of company has to solve many types of alert situations where it is necessary to be ready to act promptly to avoid negative consequences for both people and the environment.

Usually, this kind of company has to manage important problems (manual intervention, ordinary and extraordinary maintenance of technological machines) where an optimal resources intervention schedule is necessary. The experts need to be aware of all the information regarding activity sequence and the specific competences of the people involved in the process. The experts have to take decisions in a very short period of time and an error could bring negative consequences for company activities. It is necessary to explicate all knowledge typologies (structured, unstructured and semi-structured knowledge) to make information available when a particular situation is verified.

One of the most important activities for the experts is to build the entire history for each event. In domain-specific organizations, this need is stronger since there are many situations in which people have to promptly take a decision. When the experts have to analyze particular situations they need all the information about process activities and the people that are working on them. Often, a large part of information is stored in documents or reports and experts spend a lot of time looking for the relevant parts.

In particular, we consider the experts' need to rebuild the history of events in order to analyze or to find similar situations that have already been solved. This means that they have to search for both the same subject and a similar event that occurred for similar machines. This aspect is important in deciding on the significance for each domain entity to extract from texts.

The methodology proposed in this paper presents an approach to improve the search of relevant documents involved in the management of alert situations through textual semantic annotation techniques. It goes on to explain how this improvement can support decision makers and enrich the global performance of processes. It illustrates how an optimal document organization can support the experts during their analysis and how semantic annotation techniques could explicate a relevant part of important process knowledge contained in the text. The aim is to present a methodological approach usable for all kinds of texts and specific domains where analysis of the entities' semantic relatedness plays an important role.

This paper is structured as follows. Section II presents the state of art of related works. Section III describes the context and the problem statement of the methodological proposal. Section IV illustrates the aim and steps of the proposed methodology. Section V discusses future work and presents concluding observations.

## II. STATE OF ART

Text linguistic annotation consists in coding the linguistic information associated to textual data. In computational linguistics, text annotation has an important role that has been consolidated over the years. It allows computers or machines to extract, interpret and explore the linguistic structure of texts and gives an added value to single terms. From a linguistic point of view, textual data are arranged in different levels through a hierarchical organization that is often partially defined. As a consequence, text annotation is a delicate process as it gives different interpretations of the phenomena that have to be annotated based on annotation levels applied to the texts [1].

The proposed methodology uses semantic annotation techniques to extract concepts from the specific domain texts where the general meaning within a specific domain is important to explicate linguistic entities.

This section shows a short overview of the related work. In particular, let us refer to the semantic annotation techniques applied in different domains, for different aims and on different types of texts.

Due to the large amount of literature concerning Natural Language Processing (NLP) techniques, this section is focused on a representative set for the presented work. The aim is showing how NLP is used in different specific domains, mainly referring to the medical diagnosis and business processes. Concerning the methodology presented in this paper, this section aims at illustrating how it was defined and also how, with examples of NLP applications in other domains, the guidelines and criteria to follow were identified.

For example, in the medical domain NLP techniques have had a large application in structuring clinical information and making available codified diagnosis information so as to understand a natural specific domain language used in the text [2]. In another application in the medical domain, an annotated corpus (PhenoCHF) was created to better understand the medical sub-domain of congestive heart failure (CHF) [3]. The corpus focuses on the identification of phenotype information for a specific clinical sub-domain, congestive heart failure (CHF) and the annotation scheme, whose design was guided by a domain expert, includes both entities and relations pertinent to CHF. Extracting phenotype information can have a major impact on our deeper understanding of disease etiology, treatment and prevention. This is a case in which a corpus is annotated in order to complete the domain-specific vocabulary and to understand a possible evolution of the phenomena.

In general, there are a large number of NLP medical domain applications as the language and the context are more difficult to understand; there are linguistic and contextual factors to consider and language is subject to continuous evolution. Such as in the domain presented in this paper, the medical domain is an interesting similar application that offers many points for reflection to define the following methodology. The medical domain is similar to the domain considered in this paper for the importance of entities' semantic relatedness for analyzing diagnostic processes.

In [4], a platform named MeTAE (Medical Texts Annotation and Exploration) to extract medical entities and relationships from medical texts is presented. Determining the categories of the medical entities identified in the text is difficult. Such as in the domain presented in this paper, one of the most important obstacles is the high terminological variation to express the same concept (ex. Swine influenza = swine flu = pig flu). The semantic annotation application is also used to understand medical problems and to maintain the problems lists accurate and up-to-date. Indeed, in [5] an NLP application to extract medical problems from narrative text clinical documents is presented. The algorithm has extracted 80 different medical problems selected for their frequency of use. Within a set of documents, it identifies for each document a series of problems treated, selecting the medical entities from the sentences. This methodology also works on negation detection to explicate both what a patient has and what a patient does not have. The same disease could be part of a positive sentence (ex. “The patient is known for diabetes mellitus”) but also in the negative sentence (ex. “No diabetes is reported in the patient’s history”) so the application may be able to recognize specific linguistic context.

As mentioned above, the medical domain offers an important starting point to reflect on the diagnostic processes analyzed in this paper. The diagnostics process for technological machines could be compared to the patients’ disease diagnosis in the medical domain. The problems list is important also in the technical domain (such as the energy domain) and the same concept can be expressed in different linguistic forms. It is not sufficient to identify the specific entities but it is necessary to analyze their context in the sentences. This application on the texts in the specific domain could be an important starting point for improving technical vocabulary that often, mainly in the free texts (such e-mails, narrative description etc.), is not used in their typical forms. In addition, these NLP techniques in the medical domain play an important role also in clinical decision making support such as explained in [6].

As already noted, NLP techniques have a wide variety of domain applications. In the business intelligence domain, the enterprise can use these textual techniques to capture information and opinions contained in different kinds of sources. An experiment to identify lexical, morpho-syntactic, and sentiment-based features derived from web sources is presented in [7]. The aim was to collect all opinions about the company and analyze what has been said about company. To do that, business analysts need to analyze textual sources and accordingly make decisions about business actions. The experiments use NLP techniques to identify if the sentences refer to positive or negative opinions. These techniques are inserted in the business lifecycle as a strategic monitoring phase.

Another example of NLP techniques applied to Business Intelligence is described in [8]. A system which allows extracting relevant information to be fed into models for analysis of financial and operational risk and other business intelligence applications is described.

Let us consider that NLP techniques and information extraction from texts have a strategic role in making textual knowledge available. This kind of knowledge is often important to understand domain evolution (in terms of vocabulary used, opinions, contexts), support the decision-makers that need to have available and formalized all kinds of knowledge, and also to predict future actions and strategies.

These examples are aimed at giving a general idea on NLP applications inserted in specific domains and useful for specific processes and objectives.

The proposal presented was inspired by this kind of experiences but with the consequence to optimize classification documents crucial for the diagnostic processes carry out in a specific domain. The methodology proposed in this paper refers to specific domains in which textual information has to be inserted in diagnostic process as the guidelines to promptly find relevant documents. It considers not only the domain specific entities present in the text and their relevance but also the relatedness within them (e.g., temperature – pressure, plant n°1- plant n°2, etc.). In this way, people involved in diagnostic processes can find the correct information and the information related whit their initial searches in the shortest time.

## III. METHODOLOGICAL PROPOSAL

### A. Context

The work proposed in this paper has an important application in domain-specific organizations. The proposed methodology can be applied in domain-specific organizations due to the important role that technological capital plays for these enterprises. In particular, this assumption is based on the diagnostic process provided in companies operating in the energy sector.

The present methodological proposal concentrates on diagnostic problem processes that require a strong expertise exchange within the experts involved and where there is a large part of textual knowledge to explicate. The diagnostic activities represent all operations carried out by experts to determinate the main causes and the nature of defects verified on technological machines or other support. Let us consider the diagnostic processes where there are two main typologies of human resources that communicate with each other: the technicians for the manual intervention and maintenance, and the experts with specific competences that have to analyze monitoring data and parameters (Fig. 1).

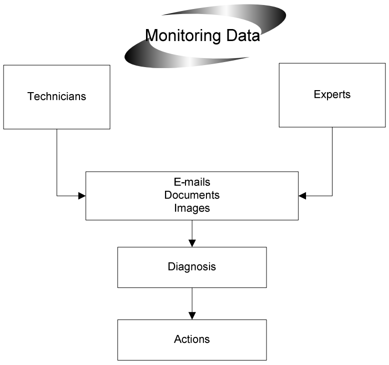

Fig. 1: Diagnostic process

With their expertise exchange, they produce unstructured information sources (e.g., e-mails, graphs, images) that have to be analyzed for formulating precise diagnosis quickly. In order to carry out these analyses and search for solutions, an optimal text classification could be an important aspect to improve complex diagnostic processes. The assumption is that document classification plays a fundamental role in experts' activities while searching for relevant information that sometimes is unstructured and difficult to obtain.

### B. Problems Statement

This section illustrates the problem statement and outlines the kind of processes in which the proposed methodology can be applied. Let us consider the diagnostic process to identify problems on technological machines and equipment or the maintenance process carried out to repair defects or malfunctions. They are two different kinds of processes; however, both impose the analytical phases.

Let us take, for example, the problems that may occur in a power plant or nuclear power plant where an in depth analysis is needed and a solution must be found quickly.

All pieces of these plants are controlled with specific monitoring tools and all particular situations are reported in specific documents (e.g., maintenance report, check-ups). The experts' activities concern the examination of monitoring values and the diagnostic problems after consulting with other actors involved in the process. When the experts verify alert situations in the parameters concerning the technological pieces of the energy plants, they begin an information exchange using specific information support (databases, e-mail exchange, documents, images etc.) and creating different kinds of information sources.

The methodology presented here concentrates solely on the textual information concerning particular events which is sometimes unstructured and difficult to find.

The semantic annotation approach presented in this paper gives a significance value to each domain-specific entity. In this way, the document classification labels will be formed of relevant domain entities and will better represent the texts' content. The general concept consists in assigning a significance indicator to each entity, consequently building the classification labels considering these criteria. The aim is to create a representative label with the entities based on a relevant importance indicator and not only on frequency.

Let us remember that the methodological aim is to create document classification labels to allow finding relevant documents for alert situations that could be verified in a specific domain. This relevance indicator could be established a-priori considering specific domain language and the specific aims, which will be explained in the next section.

To sum up, the assumed diagnostic processes are structured as follows:

1) *All pieces of energy plants or technological machines are always monitored. People with special competences have the role to verify potential anomalies.*

2) *Monitoring tools show the anomalies in the values of parameters or that machines do not function properly.*

3) *The experts begin the information exchange using specific information support. They produce specific documents such as reports or write e-mails to consult other actors involved in the process.*

ISBN: 978-1-61208-457-2

*4) This textual information is capitalized in specific document bases that should facilitate the users' search for relevant documents. In particular, documents are organized according to specific classification schema that has to explicit the real content of the texts.*

*5) When a particular situation is verified, the experts have to search relevant documents to rebuild the history of events.*

Therefore, the proposal outlined here, has the aim to improve document classification in order to find relevant documents in less time. In order to do so, the methodology proposes to use semantic annotation techniques to explicate the real contents of the texts with a domain-specific language.

## IV. Methodological Approach

Let us imagine a sample of documents (structured or unstructured texts) that need to be classified based on their content, and suppose that such documents are composed of terms used in a specific domain. Let us consider that a part of these documents already have a classification and are already organized with the appropriate metadata. Let us consider also that the terminology that has to be found in the documents is composed of single and composite terms and an associated importance indicator has to be determined for each entity.

In this paper, the list of domain vocabulary is referred to as "domain-specific terminology" and it is searched for in each text included in the corpus.

The paper proposes an application of semantic annotation techniques to detect representative features for each document and optimize their classification where their consultation is a strategic part of an important process. It works on domain-specific language contained in structured or unstructured texts and aims to identify the most representative features based on significance indicators. These indicators will allow working with a reduced number of features considering the terms' distributional semantics in analyzed context [9] and their semantic similarity [10]. Examples of methods for features reduction are presented in [11].

Specific semantic annotation rules are used to identify the composite terms and suggest how this annotation can facilitate document classification and consequently contribute to improving processes aimed at decision making. In this way, textual knowledge extracted through semantic annotation techniques could be considered an important support during processes that require all kinds of knowledge available and formalized. For this reason, the methodology is applied when textual information and document classification play an important role in the processes.

For the diagnostic processes presented, the specific documents and textual information produced contain a great deal of expertise exchange and fundamental rules to solve particular situations.

### A. Text pre-processing

In a first phase, it is necessary to identify a "bag of words" characterizing the domain-specific terminology. Let us consider that any structured language is used and any domain ontology is built for the analyzed domain. Let us suppose that a specific enterprise has a proper vocabulary (ex. particular name for machines and pieces) that sometimes is difficult to explicate for the domain inexpert.

The specific domain entities are identified considering the specific terms used within specific enterprises to carry out their activities. For example, the specific machine codes, place with a specific definition ("place name + plant number"), specific machines' piece names ("piece name + piece number"), etc. For each of these entities a significance indicator is determined for building a classification label. It is possible to associate a different significance indicator to the terms and build a classification label according to the final goal (e.g., it could be important to rebuild the history of the parameters or places but parameters are more important than places).

The corpus is annotated according to Part-of-speech techniques and specific annotation rules are created in order to find all the elements in the domain-specific terminology list. In particular, in a first phase the methodological proposal consists in applying semantic annotation techniques using Part-of-speech tagging tools.

Part-of-speech (POS) tagging is one of the most popular and thoroughly researched tasks in the field of natural language processing, particularly since it is a prerequisite for a wide variety of more complex tasks [12].

This analysis gives the grammatical category for each term contained in the texts based on the text language. Subsequently, the methodology provides the construction of a set of semantic annotation rules to identify word features of the specific domain.

Each term contained in the domain-specific terminology is searched for in the documents regardless if they are unstructured or structured. Hence, the proposed method allows the possibility to classify all kinds of textual information involved in processes with an important analytical phase. For each document, a word vector characterizing the texts' content is identified.

### B. Features Selection

The identified terms represent the features for each document that are subsequently used to form the classification label formed by terms having greater significance. A selection from the word vector associated to each document is carried out according to domain and entities relatedness. Let us consider finding a list of terms associated to a document; and let us suppose that the list of terms to find in the documents is divided into categories (e.g., parameters, places, machines etc.). According to the method presented, if there is more than one term belonging to the same category in the analyzed text, the most frequent term along with the one with the greatest semantic relatedness are chosen to represent the text content.

Indeed, the classification classes that the method presented here aims to create are based on the texts' content and related tags. The approach provides a textual analysis (in particular semantic annotation) to explicate the real content of documents involved in analytical phases of processes that require prompt actions. The application of this work could be an automatic and dynamic classification schema capable of identifying in real time the correct label of documents. This approach could improve classification schema precision, minimizing search time and facilitating the access to relevant documents. The results obtained could benefit decision support makers because of their need to have knowledge available of past and connected events.

The classification labels built with this methodological approach could help event trend analysis and explicate the important knowledge contained in the texts.

### *C. Example Case Study*

What follows is an illustration of an example case study in which the proposed methodology is being applied. A corpus of e-mails exchanged among the different actors involved in diagnostic or maintenance processes in a domain-specific organization has been analyzed. In particular, approximately 2000 e-mail conversations are considered; the POS tagging has been applied and the specific domain entities are in part identified.

This organization, just like many others, has a technological capital that has to be constantly monitored to repair defects and avoid serious damages.

Each e-mail text has to be classified based on its content so as to facilitate future access to information. In these unstructured texts, they explain particular situations that require a diagnostic analysis and how they solved past problems; this however, is not the only content. As already mentioned, the experts control the values of parameters of the technological machines through specific monitoring tools. When they verify a particular situation that requires a discussion, they start an information exchange via e-mail to find a correct diagnosis. During this phase they have to study the history of the event in question searching in previous documents or e-mails related to this event.

This expertise exchange is capitalized in a document base and each conversation (a set of e-mails that should be referred to the same subject) is classified with a label and metadata. Sometimes the e-mail content was not compatible with the metadata associated to the specific conversation because relevant elements could appear successively.

These e-mail texts already have a classification schema based on the e-mail subject or experts’ personal evaluation that often do not represent the real content of the messages. Let us take into consideration the particular situations described in the texts with the widest variety of terms possible that could be found within them.

Consequently, the actors involved spend a lot of time searching for pertinent documents and a reduction in the time spent could bring an important improvement to the general diagnostic process. In particular, they have to find similar events in the past related on the same machine, all events concerning the particular piece analyzed or similar events concerning the same piece but belonging to another machine with the same technological structure.

Considering this example case study, the methodology proposed aims at creating a classification schema that allows to explicate the real contents and that considers all important elements in the text. For this reason, it is extremely important to determine the list of domain-specific terminology and their significance: to capture as many domain-specific terms as possible in the text Compound terms are identified subsequent to POS tagging and the parsing techniques [13] and through the automatic rules used in GATE (General Architecture for Text Engineering) software with the JAPE language [14]. With this platform it is possible to build specific rules starting from general categorization obtained with POS tagging. The JAPE language allows to determine a grammar for GATE to annotate not only the named entities but also the domain-specific items.

The extracted terms represent the candidates in setting classification labels. The label associated to each text will be formed considering the term frequency but also the importance indicators associated to each term.

To sum up, the main methodological steps are structured as follows:

*1) Identifying a process where documents search and consulting have a crucial role.*

*2) Determining the aims for semantic annotation and the relevant aspects to search in the documents*

*3) Structuring the set of specific domain items (with typical vocabulary used in the specific company) to search in the texts and determining for each of these a significance indicator for selecting the representative features.*

*4) Applying the POS tagging technique.*

*5) Building the specific rules to annotate the texts and find specific domain entities.*

*6) Verifying for each document the vector of words associated and selecting the entities with the greatest significance indicator.*

*7) Determining for each document the classification label sorted by their significance indicator and their respective class.*

## V. CONCLUSIONS AND PERSPECTIVES

This paper has outlined the most important aspects of semantic annotation techniques applied in a specific domain. In particular, it has presented the use of semantic annotation to support decision making within processes that require important analytical phases and where document consulting plays an important role in specific processes. The presented annotation method aims to improve document classification in order to help experts who need to find relevant information in a relatively short period of time. It has explained how this could be an important contribution in improving the global process that is being analyzed. In particular, it has based its assumption on the real need identified after an experience in a domain-specific organization.

In this work it would be worthwhile to use specific optimization algorithms in order to evaluate the classification schema found through semantic annotation techniques. The futures goals will be to test the classification quality using semi-supervised classification algorithms. In particular, the aim is to represent each document in the n-dimensional space where n is the dimension of bag of words. For each text, a vector of n elements will be created. It represents the absence or presence in the text of each word contained in the bag of words. Subsequently, the proposal is to verify the classification with a semi-supervised algorithm.

## REFERENCES

[1] A. Lenci, S. Montemagni, and V. Pirrelli, “Testo e computer: elementi di linguistica computazionale”, Carocci, May 2005, ISBN:88-430-3425-1.

[2] C. Friedman, H. George, "Natural language processing and its future in medicine", Academic Medicine 1999, 74.8, pp. 890-

895.

[3] N. Alnazzawi, P. Thompson, and S. Ananiadou, "Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature", LOUHI 2014, pp. 69-74.

[4] A.B. Abacha, P. Zweigenbaum, “Automatic extraction of semantic relations between medical entities: a rule based approach”, J. Biomedical Semantics 2011, 2(S-5), S4.

[5] S. Meystre, P.J. Haug, “Natural language processing to extract medical problems from electronic clinical documents: performance evaluation”, Journal of biomedical informatics 2006, 39(6), pp. 589-599.

[6] C. Demner-Fushman, W. W. Chapman, and C.J. McDonald, “What can natural language processing do for clinical decision support?”, Journal of biomedical informatics 2009, 42(5), pp. 760-772.

[7] H. Saggion, A. Funk, “Extracting opinions and facts for business intelligence”, RNTI Journal 2009, E (17), pp. 119-146.

[8] D. Maynard, H. Saggion, M. Yankova, K. Bontcheva, and W. Peters, “Natural language technology for information integration in business intelligence”, Business Information Systems, Springer Berlin Heidelberg, January 2007, pp. 366-380.

[9] A. Lenci, "Distributional semantics in linguistic and cognitive research", Italian Journal of Linguistics 20.1, 2008, pp. 1-31.

[10] T. Cohen, D. Widdows, “Empirical distributional semantics: Methods and biomedical applications”, Journal of biomedical informatics, 42(2), 2009, pp. 390 -405.

[11] Y. Yang, J.O. Pedersen, “A comparative study on feature selection in text categorization”, ICML 1997, Vol. 97, pp. 412-420.

[12] H. Van Halteren, ed. “Syntactic word class tagging”, Kluwer 1999.

[13] G. Kennedy, “An introduction to corpus linguistics”, Routledge 2014.

[14] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham, "Evolving GATE to meet new challenges in language engineering", Natural Language Engineering 2004, 10.3-4, pp. 349-373.