An iterative approach for lexicon characterization in juridical context

F. Amato, A. Mazzeo, S. Romano, S. Scippacercola[1]

Abstract In the juridical context, knowledge management applications have a central role. In order to improve the effectiveness of document management procedures, techniques for automatic comprehension of textual content are required. In this work, a methodology for semi-automatic derivation of knowledge from document collections is proposed. In order to extract relevant information from document text, a process integrating both statistical and lexical approaches is applied. Moreover, we propose a system for the evaluation of the extracted peculiar lexicon quality. The system is used for the processing of heterogeneous documents corpus issued by Italy's juridical domain.

Introduction

Actually, information tecnologies are applied to several service areas leading to a growth of services organizations. In this context, knowledge management dealing with acquiring, maintaining, and accessing knowledge within data can improve services furnishing. Often the competitiveness of an e-services organization depends heavily on how knowledge is maintained and accessed. Difficulties arise when the knowledge is contained in textual format (for example electronic or paper document) without any support so that the contents could be machine-readable and processable. In these cases, techniques for automatic comprehension of textual content are required. Many efforts are currently devoted to extract knowledge from texts in order to enhance some features provided by several systems. For example, in [2] is proposed a method for automatic detection of security access requirements for shared resources in a e-health system. Ontologies are developed to provide a machine-processable semantics of information sources that can be communicated between different agents (software and humans).

In general, knowledge is represented by a set of domain concepts and by the relationships between those concepts. Therefore, the automatic comprehension of textual contents involves several text-processing disciplines that work considering complex and strongly inter-dependent syntactic, semantic and pragmatic aspects.

[1] Università di Napoli "Federico II", Dipartimento di Informatica e Sistemistica, Napoli, Italia. flora.amato@unina.it, mazzeo@unina.it, sara.romano@unina.it, sergio.scippacercola@unina.it

In order to extract knowledge from textual documents, it is necessary to identify domain relevant terms (words), their meanings (i.e. concepts), and the relationships among them.

The activities of document processing and derivation of knowledge from text have as requirement the identification of the peculiar lexicon, which is a terminological vocabulary representative of the domain of interests.

The peculiar lexicon is a terminological vocabulary that contains the most significant and representative key-words which define the contents of the textual fragments and in general the whole domain whose corpus is a representative sample set. Once the peculiar lexicon has been extracted from documents, it provides the basis for the construction of the domain conceptual system, enabling semantic processing of the documents contents by working with the meanings of the resources.

Different kinds of text analysis methodologies are involved in the activity of knowledge extraction from texts. The state of the art in this field is related to techniques of NLP with cross-disciplinary perspectives including Statistical Linguistics [3, 4, 5, 11] and Computational Linguistics [9, 14], whose objective is the study and the analysis of natural language and its functioning through computational tools and models.

The detected concepts are coded by means of ontology and represent the starting point for semantic processing of document contents [9].

In this work, we propose a methodology for the semi-automatic derivation of documents content by means of techniques for domain-specific terms extraction for peculiar lexicon definition and techniques for domain relevant concepts identification that integrate both linguistics and statistics aspects for textual data interpretation. The paper is organized as follows: in the next paragraph the language characterization will be defined; in the 3rd we will explain what is the peculiar lexicon and what is intended for concept; in the 4th paragraph process for knowledge extraction from text will be described; in the 5th paragraph we describe our methodology for peculiar lexicon assessment.

Peculiar lexicon and concepts

It is possible to divide the knowledge extraction process into two macro-processes:

- peculiar lexicon extraction from text based on advanced terms extraction techniques;
- concepts identification based on recognition of specific relationship between the words belonging to the peculiar lexicon.

The peculiar lexicon is a terminological vocabulary. It contains the words that are representative for the domain of interests. Generally, not all the words are useful for characterizing the semantics of a documents corpus: this is the case of grammatical words, for example articles and prepositions, that, even forming the connective tissue of a text, represent *“noise”* since they are not carriers of meaningful contents.

Term-extraction involves a series of sub-tasks that affect different levels of analysis:

1. Text pre-processing: tokenization and normalization procedures;
2. Morpho-syntactic analysis: part-of-speech tagging, lemmatization, identification of phrase structures;
3. Relevant terms extraction;

Not only simple words but also complex words, which are syntagmatic combinations of terms, contribute to specific domain concepts definitions.

These complex lexical expressions, which lead to a complete and autonomous sense, are very frequent when dealing with specialized domains. Phrase structures represent often specializations of more general concepts (like as the Italian expression ``imposta di bollo" -- duty stamp -- that is a specialization of ``imposta" -- duty -).

Loosing the overall sense of these sequences during text analyses, may lead to lexical item dispersion: for this reason, it is necessary to process complex expressions as autonomous units of analysis [4].

Relevant concepts identification firstly requires the ability to recognize the entities within the text structure which refer to concepts and in the second place the ability to identify the constraints to which entities are subjected and the properties characterizing them [7].

A concept can be defined as a mental representation whose definition should ideally include [6]:

1. an intentional meaning, defined by the set of intrinsic properties that are necessary and sufficient to characterize concepts and to make it possible to distinguish them from other concepts;
2. an extensional meaning, defined by all the referential entities to which intrinsic properties of concepts are applied;
3. a lexical expression used to refer to entities to which concepts apply and to refer to concepts themselves.

While operating in specialized domains, the extensional meanings of concepts are simple enough to be managed, since lexicons are more specialized and full of technical terms within the intentional meanings of domain concepts. During interpretations of the document contents, which is dependent by authors and readers shared domain competences and knowledge, the process of coding/decoding concepts from the words can be reached without (or in the worst case, with a reduced) ambiguity.

Extracting the semantic content from text

In order to identify the most significant words in a text both linguistic and statistical approaches are used in a deeply integrated way. The former goes into the linguistic structures of the text by analyzing the meanings of words; the latter, instead, provides quantitative representations of the identified phenomena.

In particular, the extraction of peculiar lexicons process is given by the integration of:

1. Endogenous (corpus based) strategies, like the extraction of the TF-IDF index (Term Frequency Inverse Document Frequency), by which it is possible to extract the most relevant lexical forms, representing the topics of the documents. It is classically used for identifying index terms, and it is based on the principle that, for every document, the most relevant words occur many times within a single document, but in a small number of the total documents.
2. Exogenous (external) strategies, like as the comparison of the corpus with domains sub-languages (list of words that certainly belong to the issued domain). The comparison is applied for recognition of shared words, and for the identification of the lexical items, which are over or under used with respect to sub-languages of references usually provided by domain experts.

The first strategy enables the extraction of statistically significant lexical items, whose semantic specificity is evaluated with regard to the topics dealt in the corpus under examination.

Domain terms behave differently since they can occur at a high or low rate of frequency or have a wider or narrower distribution within the corpus. The best strategy to single out domain terms within a document collection is to resort to the second strategy, which is based on exogenous resources, such as general or specialized lexical external lists. This strategy enables the extraction of peculiar lexical items, where this peculiarity is evaluated with regard to the specific sublanguage to which the corpus under examination pertains (in this case, the legal language). By comparing the vocabulary of the corpus under examination to a domain lexical list (such as JurWordNet [8] or any other domain lexical database) it is possible to identify those terms that surely pertain to the specific sublanguage [1]. It is then clearly important to opt for appropriate strategies capable of describing the relevance of the words in a document collection in terms of discriminating power and semantic representativeness and peculiarity with respect to a sublanguage.

The idea of integration of statistical and lexical approaches rises from Lame [10], which has shown that a purely statistical approach produces high values of semantic precision with respect to the corpus contents but poor values of word recall with respect to the domain language. Statistical indexes, which were classically used to identify index terms, cannot be used to distinguish domain terms from non-domain terms since they do not always correspond with domain terms. Therefore, in order to extract the peculiar words from a document collection with respect to the specific domain of interest, Lame suggests the use of exogenous resources, like lexical external resources that enable useful comparisons with general or specialized domain terms. Therefore, index terms do not always correspond with domain terms. Vice versa, domain terms do not always correspond with lexical items having the highest lexicometric values.

In order to define the peculiar lexicon that better represents the domain of interest, our strategy uses a hybrid method, that integrates both linguistic and the statistical approaches. It is based on the Luhn's law [12] that is based on the following consideration: if we order the words in the text by frequency, and consider the distribution of the frequency of the ordered words (Figure 1), the index terms between the two cut-offs have the highest discriminating capacity.

We can consider two cut-offs dividing the distribution of the word frequencies into three main sections. The lowest cut-off separates all the words having a high frequency, which are not significant for document characterization (such as generic or common words). On the contrary, the highest cut-off separates rare words, which cannot be considered significant enough to be inserted in the peculiar lexicon, because they are present only in few documents. Conventionally the two cut-offs are set arbitrarily.

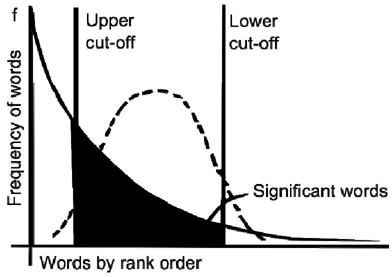

Figure 1. Luhn's law

Peculiar lexicon Assessment

Our approach aims at determining the position of the two cut-offs, in order to increase the meaningfulness of the extracted peculiar terms. This approach is based on a iterative method that refines cut-off positions depending on the computed distance between the document and the lexicon extracted. The proposed methodology is enacted following the steps depicted in Figure 2.

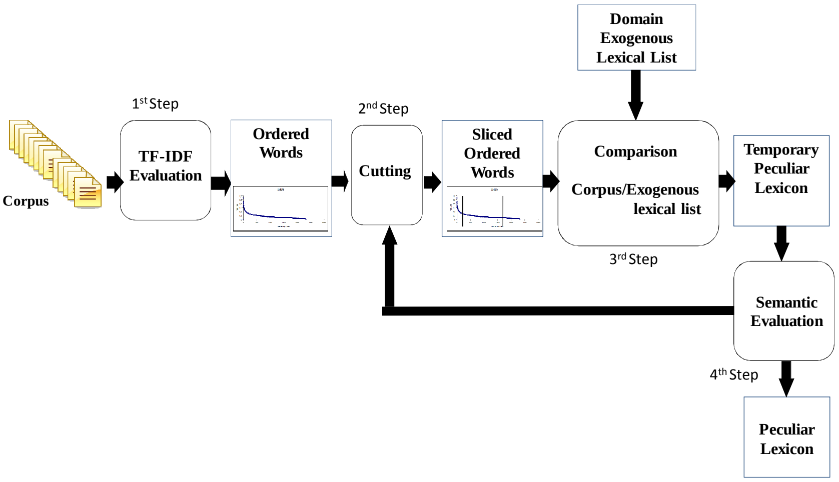

Figure 2. Iterative Processing for identification of Peculiar Lexicon

In the first step, the TF-IDF is computed and we sort the index terms list in decreasing order.

In the second step, the index terms, in the list, are filtered selecting that lemmas included between two cut-offs.

The filtered list, in the third step, is compared with a reference vocabulary in order to discard terms that don't belong to the domain. From this step, a temporary *peculiar lexical list* is obtained.

In the fourth step, the *semantic distance* among the documents and the temporary peculiar lexicon is evaluated using a distance measure, based on χ^2 statistical measure, and the cut-off positions are assessed consequently, enlarging the range of selected words if the distance is below some tolerance values, narrowing vice versa. The tolerance value is empirically defined by the help of domain experts.

The evaluation of the *semantic distance*, in the assessment algorithm devised, is based on four criteria:

(I) The decrease of the χ^2 distance among all the documents, the corpus, the peculiar lexical items;

(II) The increase of the cover rate of each document and the corpus;

(III) The increase of the cover rate of each document and the peculiar lexical items;

(IV) The χ^2 distance among the corpus, the peculiar lexical items derived by exogenous method and the peculiar lexical items by using the proposed method. Lower values of χ^2 distance imply better result.

The algorithm is iterated until a satisfying result is obtained (*peculiar lexical items*).

For example, we consider the similarity analysis performed on a corpus of heterogeneous documents (Tabb. 1, 2, 3) issued by our running example in Notary domain. We execute, therefore, the extraction of a list of relevant words through the TFIDF index and the progressive skimming of the list obtained by comparing it with two different lexicons: firstly a general lexicon for the Italian language and in second place the lexical database of JurWordNet in order to extract a more and more specialized lexicon. After the first iteration (Table 1), the document *Doc1* is the worst semantically represented (I criterion). This is confirmed by the low cover rates (second and third criterion) in Table 2. In the same example, the document *Doc11* is instead the best semantically represented according to second and third criterion (Tab. 1, Tab. 2). In the shown application the fourth criterion is fully confirmed (Tab. 3) as the χ^2 distance between the corpus and the lexicon extracted is lower than the χ^2 distance between the corpus and the lexicon extracted by means of exogenous method.

Table 1. The χ^2 distance among the documents, the corpus and the peculiar lexical items (example in Notary domain).

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11
Corpus	15,47	2,61	3,88	4,88	3,23	4,34	5,71	5,28	6,20	4,61	2,36
Peculiar lexicon	27,25	13,18	15,15	16,14	13,57	15,75	16,80	16,49	17,02	15,48	13,40

Table 2. Cover rates of each document, the corpus and the lexical peculiar index (example in Notary domain). In the table, the acronym CRC stands for “Cover rate respect to corpus” while CRPL stands for “Cover rate respect to lexical peculiar index”.

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11
CRC	6,022	34,017	19,5	14,35	23,51	16,4	14,41	14,03	12,428	16	43,11
CRPL	2,02	36,364	10,1	8,081	26,77	7,071	8,586	7,071	8,5859	11,1	31,82

Table 3. The χ^2 distance among the corpus, the peculiar lexical items (313 lemmas) by using the proposed method and the peculiar lexical items (198 lemmas) derived by exogenous method (example in Notary domain). In the table, the acronym PLP stands for “Peculiar lexicon by the proposed method” while PLE stands for “Peculiar lexicon by the exogenous method”

	PLP	PLE
Corpus	2,98	4,63

Conclusion

In this work, we have described a strategy for refinement of the peculiar lexicon associated to a corpus belonging to a specialistic domain.

The proposed strategy is the starting point for the definition of a lexicon to be used in a system for the management of documents belonging to specialized domain. The restricted area of specialization reduces the intrinsic semantic ambiguity of the words, related at the generalist domain, allowing a more accurate semantic processing.

For the moment, the strategy is used by a corpus of documents belonging to juridical domain: future effort will be devoted to extend experimental results to other corpora, in order to validate the proposed approach.

References

1. Amato F., Canonico R., Mazzeo A., Penta A., Picariello A. (2008), Semi Automatic Extraction of a Peculiar Vocabulary in Notary Domain. Book of short papers MTISD 2008, Methods, Models and Information Technologies for Decision Support Systems, Ed. Università del Sannio, Lecce, pp. 313-316.
2. Amato F., Casola V., Mazzeo A., Romano S. (2010), A semantic based methodology to classify and protect sensitive data in medical records", Proceedings of sixth international conference on information assurance and security (IAS), Atlanta, USA, August 23-24.
3. Balbi S., Bolasco S., Verde R., (2002) Text Mining on elementary forms in complex lexical structures", JADT 2002. Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles, Saint-Malo, IRISA, pp. 89-100.
4. Bolasco S. (2004), L'analisi statistica dei dati testuali: intrecci problematici e prospettive. In: Bolasco S., Cutillo E. A. Applicazioni di analisi statistica dei dati testuali, Roma, Casa Editrice Università La Sapienza.
5. Bolasco S., Pavone P. (2007), Automatic dictionary and rule-based systems for extracting information from text, Classification and Data Analysis, Book of Short Paper, Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, 255-258.
6. Buitelaar, P., Cimiano, P., and Magnini, B., editors (2005). Ontology Learning from Text: Methods, Evaluation and Applications, volume 123 of Frontiers in Arti_cial Intelligence and Applications.IOS Press, Amsterdam.
7. Dell'Orletta F., Lenci A., Montemagni S., Marchi S., Pirrelli V., Venturi G. (2008) Acquiring Legal Ontologies from Domain-specific Texts. In Proceeding of LangTech 2008, Rome, February 2008, 28-29, CD-ROM.
8. Gangemi A., Sagri MT., Tiscornia D., (2005) A constructive framework for legal ontologies, Law and the Semantic Web, Springer.
9. Giovannetti E., Marchi S., Montemagni S., Bartolini R. (2008) Ontology Learning and Semantic Annotation: a Necessary Symbiosis. LREC 2008: Proceedings of the Sixth International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), Marrakech, Morocco, May 26-1 June 20.
10. Lame G. (2005), Using nlp techniques to identify legal ontology components: concept and relations, Lecture notes in Computer Science, vol. 3369: 169-184.
11. Lebart L., Salem A., Berry L. (1998), Exploring Textual Data, Kluwer Academic Publishers, Dordrecht, NL.
12. Luhn H. P. (1958) The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2:157-165.
13. Robertson S. (2004), Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation, 60(5):503–520.
14. Salton G. (1989), Automatic Text processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison Wesley.