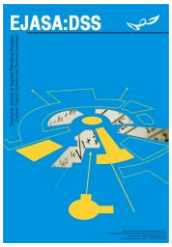

Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation
EJASA:DSS (2011). Vol 2, Issue 1, 54 – 64
e-ISSN 2037-3627, DOI 10.1285/i2037-3627v2n1p54

A METHOD FOR THE EVALUATION OF THE PECULIAR LEXICON SIGNIFICANCE

Flora Amato*, Antonino Mazzeo, Sergio Scippacercola

Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II", Italy

Received 23 February 2010; Accepted 13 April 2011
Available online 17 December 2011

***Abstract**: In this work, a methodology for semi-automatic derivation of knowledge from document collections is proposed. In order to extract relevant information from documents, a process integrating both statistical and lexical approaches is applied. We propose a strategy for the semantic evaluation of the index terms extracted in order to ensure a good correspondence between the information searched for and the information retrieved. Therefore, we propose a system for the peculiar lexicon extraction and assessment. The system can be used for defining an ontological model to be used in the semantic processing of a corpus of documents belonging to a specialist domain.*

***Keywords**: Peculiar lexicon, lexical information extraction, information retrieval techniques.*

1. Introduction

Knowledge Management deals with acquiring, maintaining, and accessing knowledge within data of an organization. Often the competitiveness among companies depends heavily on how they maintain and regulate access to their knowledge. Difficulties arise when knowledge is contained in a textual format (for example electronic or paper documents) and no support is available for codifying information in a machine-readable and processable way. In these cases, techniques for automatic processing of a textual content are required; in particular, ontologies may be used in order to provide a machine-processable semantics for the information sources, which can be further exploited for communication processes between different agents (software and humans).

*Corresponding Author. Email: flora.amato@unina.it

The proposed method for assessing the lexicon is directed towards defining and designing methodologies and techniques for semantic document-based processing, thus allowing, for example, research through content and the extraction of relevant information to be performed. Such functionalities may be used to implement suitable systems for information processing, aiming at giving support to many applications such as dematerialization processes, e-Government, web-service interoperability, the automatic treatment of structured information (questionnaires) and focus group analyses.

Nowadays, numerous systems for the management of knowledge are available, such as "Alfresco", "Cognition", as well as applications from the "Google" family. They implement procedures to search for concepts, overcoming the barriers of syntactic matches through keywords; however, they are designed for a general context, still do not prove satisfactory for specialized domains.

In fact, as far as we know, a generic strategy has not yet been developed for the retrieval of relevant concepts (i.e. those not based on matching among keywords) that is able to deliver results effectively and efficiently, without characterizing the search domain.

In a system dedicated to semantic processing, the formalization of the information embedded in the selected corpora should, therefore, not be based on generalisations, or provided by third-party information. A proper model of the target information must be built through the semi-automatic and automatic processing of the actual collection of documents that constitute the research domain.

By exploiting the information extracted with the application of the proposed method, we can build an ontological model of the domain of interest, making it possible to design advanced retrieval systems, specialized for the documents to be treated, and characterized by high performance, in terms of precision and recall.

In a system for the semantic processing of documents, knowledge may be represented by a set of domain concepts and by the relationships between these concepts. The automatic processing of textual contents involves several text-processing disciplines that work by considering complex and strongly inter-dependent syntactic, semantic and pragmatic aspects. 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.

Learning knowledge from texts includes a series of tasks starting from terminology extraction (for the identification of the relevant entities the domain concepts refer to) and leading to more complex ones, like the identification of taxonomic and non-taxonomic relationships, which aim at the identification of “Synsets” and/or conceptual taxonomies. Note that, according to lexical database WordNet, we refer to “Synset” as a set of terms that can be interchanged in a certain context [10]. 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 interest.

The peculiar lexicon is a terminological vocabulary that contains the most significant and representative keywords, which define the contents of the processed documents and in general the whole domain whose corpus is a representative set. Once the peculiar lexicon has been extracted from documents, it provides the basis for the construction of the domain’s conceptual system, enabling the 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 Natural

Language Processing (NLP) with cross-disciplinary perspectives including Statistical Linguistics [12][13][6][2][3][5] and Computational Linguistics [17][4][10], 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 ontologies, exploited for further semantic processing of document contents [10].
In this work, we propose a methodology for the semi-automatic derivation of document content by means of techniques for domain-specific terms extraction for peculiar lexicon definition and techniques for domain-relevant concept identification, that integrate both linguistic and statistical aspects for textual data interpretation.
The paper is organized as follows: in the next paragraph the language characterization will be illustrated; in the 3rd we will introduce the notion of peculiar lexicon and concept; in the 4th paragraph we will describe the process of knowledge extraction from the text, and in the 5th paragraph our methodology for peculiar lexicon assessment is defined.

2. Language characterization

Language is a code provided with a set of signs and rules for the correct use of the signs themselves. The linguistic sign has the important property of being a "dual-side" entity: one side is the so-called "signifier", that is the physically perceptible part of the sign (i.e. the sequence of sounds or graphemes composing the sign); the other side is the so-called "signified", that is the conceptual meaning transmitted by the sign itself. Therefore, the linguistic sign is nothing but an arbitrary conventional correspondence between signifier and signified. The notion of the linguistic sign often coincides with the concept of "word". A lexicon contains the set of language signs (words) while a grammar contains the finite set of rules among language signs. Note that, by abusing the notation, we generally consider a sign in a text as a sequence of characters, that can constitute a separator or a word or term.
Semantics is the part of linguistics dealing with the meaning of words, sentences and texts. The role of semantics is to clarify the relationships between signifier and signified, by determining the correct interpretation of the linguistic signs (words). Giving a representation to a word's meaning is not an easy task because the Words acquire sense depending on different kinds of syntactical and semantic relationships within the textual structure. Moreover, the meaning of words varies depending on the communicative scenario, the competences and knowledge of the interlocutors, the domain to which the text belongs, etc. The description of the semantic content of textual data implies three different sets of problems to be dealt with:

- the identification of the structure of textual data by means of the recognition of the different lexical units (words);
- the identification of the relevant textual data by indexing and extracting the relevant terminology with regard to the topic dealt with in the document and to the specific relative domain;
- the identification of the semantic relationships that the words have with other words.

The conceptual contents of a document are transmitted by words that make up the frame of a text and, more in general, the frame of a specific knowledge-domain, which they semantically characterize. To make this semantic characterization clear we first need to identify the textual lexical units and to resort to corpus-based strategies (as the computation of the TF-IDF index

(Term Frequency Inverse Document Frequency [16]) or to external lexical resources (such as domain lexical repertoires) in order to extract the relevant and peculiar textual data transmitting the semantic contents.

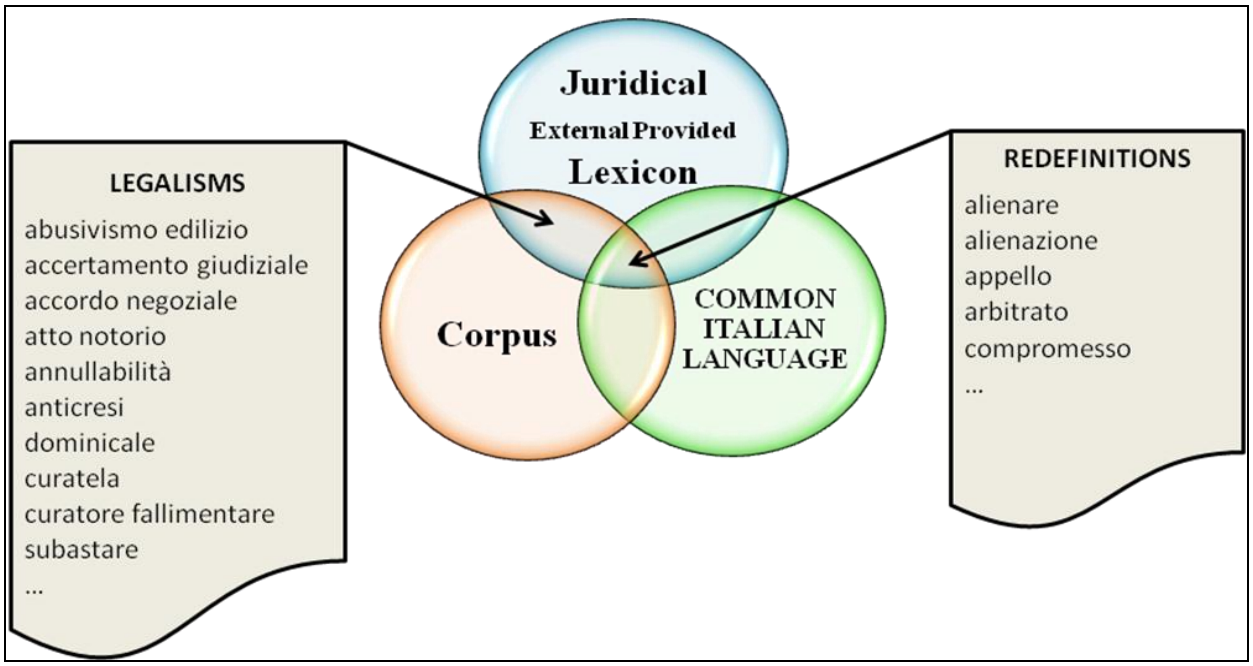

Figure 1. Lexical resources for seeking legal terms.

We aim at defining a specialized language used for providing a definite, technical and precise vocabulary, able to cope with the specific needs of a particular domain.
This vocabulary defines new words or gives other meanings to words already existing in the standard language (this is called redefinition). Many terms belonging to general domains, in fact, may be assimilated to specialized terms since they label objects, facts or behaviours that are characteristic of a specialist domain. An example is provided in Figure 1.

3. Peculiar lexicon and concepts definition

The above considerations on language characterization lead to defining processes and techniques that should be used for knowledge extraction from texts.
It is possible to divide the knowledge extraction process into two macro-activities:

- peculiar lexicon extraction from a text based on advanced term extraction techniques;
- concept's identification based on the recognition of a specific relationship among 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 interest. Generally, not all the words are useful for characterizing the semantics of a document corpus: this is the case of grammatical words, for example articles and prepositions, that, even if 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 [5]:

1. text pre-processing: tokenization and normalization procedures;
2. morph-syntactic analysis: part-of-speech tagging, lemmatization, identification of phrase structures;
3. relevant term extraction;
4. concept identification.

In these steps, particular attention is paid to the structure identification phase (Index 2). In fact, not only simple words but also complex words, which are syntagmatic combinations of terms, contribute to specific domain concepts definitions.
It is common to find sequences of words that are semantically-linked and co-occurring regularly, because of their intrinsic sense of words, which make them conceptually associated.
These complex lexical expressions, which lead to a complete and autonomous sense, are very frequent when dealing with specialized domains. Phrase structures often represent specializations of more general concepts (such as the Italian expression ``imposta di bollo" -- stamp duty -- that is a specialization of ``imposta" -- duty -).
Losing 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 [5].

The step of Relevant Concept Identification requires the ability to *(i)* recognize the entities, within the text structure, which can be referred to concepts and *(ii)* identify the constraints and the properties characterizing such entities [8].
A concept can be defined as a mental representation whose definition should ideally include [7]:

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 extentional 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 to, or to refer to concepts themselves.

While operating in specialized domains, the extentional 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 are dependent on the author's and reader's shared domain competences and knowledge, the process of coding/decoding concepts from the words can be reached without (or in the worst case, with reduced) ambiguity.

4. Extracting the semantic content from the text

In order to identify the most significant words in a text, both linguistic and statistical approaches are used in a highly 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:

- Endogenous (corpus based) strategies, like the extraction of the TF-IDF index (Term Frequency Inverse Document Frequency [5]), 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 total documents.

- Exogenous (external) strategies, such as the comparison of the corpus with the domain's sub-languages (list of words that certainly belong to the resulting domain). The comparison is performed for the 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 then evaluated with regard to the topics dealt with in the corpus under examination. An example is given in Table 1 whose lexical items have been extracted by computing the TFIDF index on a corpus composed of a set of heterogeneous documents issued by Italy's Ministry of Finance. Table 1 contains a list of significant terms having a high and low rate of TFIDF. The example shows that domain terms (for our case, legal terms) may present a high or low rate of TFIDF: statistical indexes, classically used to identify index terms, cannot be used to distinguish domain terms from non-domain terms. The statistical approach, in fact, allows the extraction of lexical profiles semantically specific to the corpus analyzed, thus producing high values of precision [17] with respect to the corpus contents but poor values of lexical recall with respect to the domain language.

Table 1. Some lexical items and respective TF-IDF rate.

Lexical item	TF-IDF rate
Enti previdenziali	0.5503
Iva	0.5503
Beneficiarie	0.4260
Immobili	0.4062
valutazione	0.2451
Redditi	0.2441
Società	0.2426
imposta di registro	0.2403
trasferimento	0.1614
apertura della successione	0.1600
Collegio	0.1600
Imprenditoria giovanile	0.1600
a seguito della cessione	0.0224
a tal fine	0.0224
finanza	0.0174
ministero	0.0174

Domain terms, in fact, can occur at a high or low rate of frequency or have a wider or narrower distribution within the corpus. The best strategy for extracting 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, whose peculiarity is evaluated with regard to the specific sublanguage to which the corpus under examination pertains. By comparing the vocabulary of the corpus under examination to a domain lexical list (such as JurWordNet[9] or any other domain lexical database) it is possible to identify the terms that certainly pertain to the specific sublanguage [1].
In order to perform further semantic processing of the text, it is important to adopt appropriate strategies able to indicate the relevance of the words in a document collection in terms of discriminating power, semantic representativeness, and peculiarity with respect to a target sublanguage.
The idea of integration of statistical and lexical approaches arises from Lame [11], who has shown that lexical items with the highest lexicometric values that were classically used to identify index terms, cannot be used to distinguish domain terms from non-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 external lexical vocabularies, enabling useful comparisons with general or specialized domain terms.
Therefore, 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 statistical approaches. For this aim, we exploit Luhn's Law [14] 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 2), the index terms between the two cut-offs have the highest discriminant capacity.
We can consider two cut-offs by dividing the distribution of the words frequencies into three main sections. The highest 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 lowest 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.

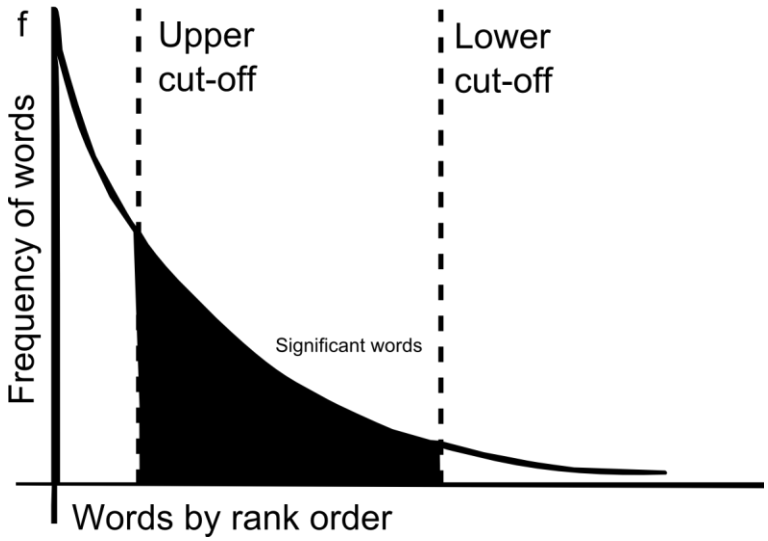

Figure 2. Luhn's Law.

5. 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 an iterative method that refines cut-off positions depending on the computed distance between the documents and the lexicon extracted. The proposed methodology is enacted following the steps shown in Figure 3. Unlike the original idea proposed by Luhn, we exploit the TF-IDF distribution, by resorting to terms on the basis of such an index.

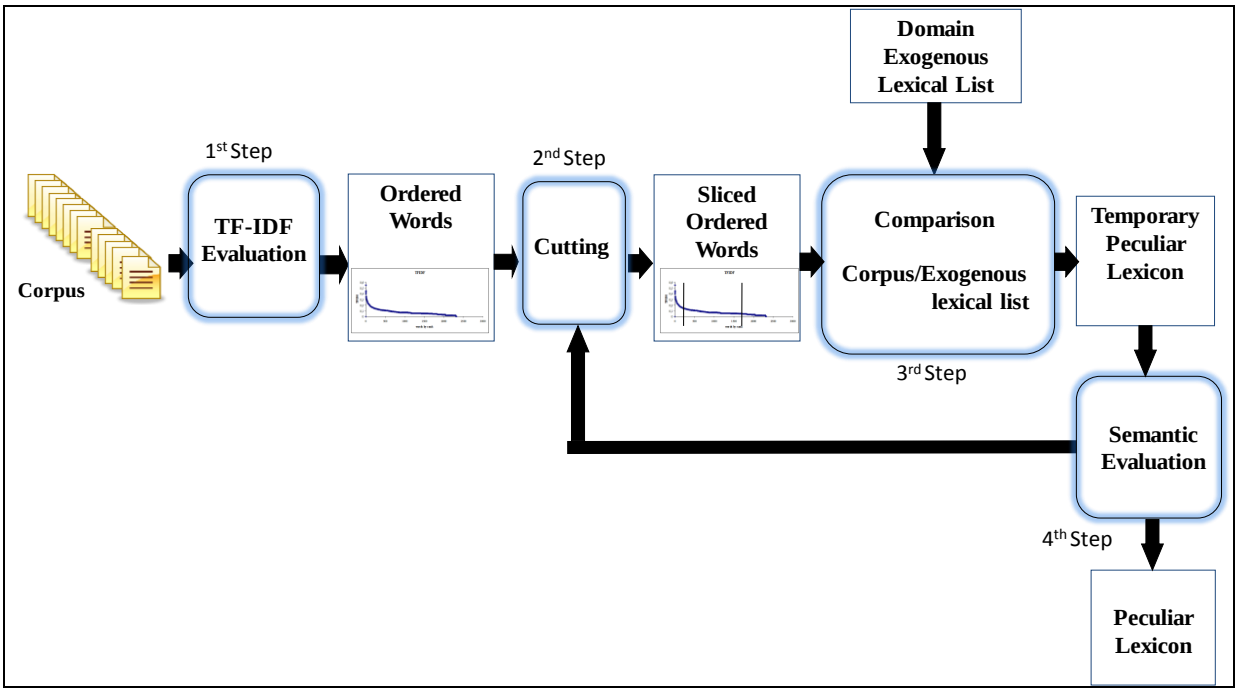

Figure 3. Iterative Processing for identification of Peculiar Lexicon.

In the first step, the TF-IDF is computed, and term list arranged in decreasing order. In the second step the index terms, in the list, are filtered by selecting the lemmas to be included between two cut-offs. In the first iteration the two cut-offs are arbitrarily set in order to include 75% of the terms.

The filtered list, in the third step, is compared with a reference vocabulary in order to discard the terms that do not 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 by using a distance measure, based on the χ^2 statistical measure, and the cut-off positions are assessed consequently, by enlarging the range of selected words, whenever the distance is lower than the defined tolerance value and by narrowing otherwise. The tolerance value is empirically defined by the help of domain experts.

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

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

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

(III) The increase in 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 a better result.

The algorithm is iterated until a satisfying result is obtained, and the *peculiar lexical items,* extracted in the last iteration, is outputted.

For example, we consider the similarity analysis performed on a corpus of heterogeneous documents (Tables 2, 3, 4) issued by our running example in the notary domain. We execute, therefore, the extraction of a list of relevant words through the TF-IDF 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 an ever more specialized lexicon. After the first iteration (Table 2), it is possible to note that the document *Doc1* results to be the worst semantically represented (I criterion).

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

	Doc1	**Doc2**	**Doc3**	**Doc4**	**Doc5**	**Doc6**	**Doc7**	**Doc8**	**Doc9**	**Doc10**	**Doc11**	**Corpus**	**Peculiar lexicon**
Doc1	0.00	15.53	16.71	17.66	16.06	17.47	19.01	18.09	19.75	17.57	16.12	15.47	27.25
Doc2	15.53	0.00	3.28	4.65	0.76	3.92	5.74	4.87	6.35	4.38	2.48	2.61	13.18
Doc3	16.71	3.28	0.00	5.39	3.71	4.75	6.83	5.96	7.11	5.36	3.53	3.88	15.15
Doc4	17.66	4.65	5.39	0.00	5.09	5.89	7.94	7.03	7.85	6.16	4.90	4.88	16.14
Doc5	16.06	0.76	3.71	5.09	0.00	4.39	6.34	5.19	6.78	4.96	2.85	3.23	13.57
Doc6	17.47	3.92	4.75	5.89	4.39	0.00	7.34	6.70	7.64	5.92	4.16	4.34	15.75
Doc7	19.01	5.74	6.83	7.94	6.34	7.34	0.00	8.60	9.22	7.65	6.09	5.71	16.80
Doc8	18.09	4.87	5.96	7.03	5.19	6.70	8.60	0.00	8.83	7.00	5.18	5.28	16.49
Doc9	19.75	6.35	7.11	7.85	6.78	7.64	9.22	8.83	0.00	7.84	6.54	6.20	17.02
Doc10	17.57	4.38	5.36	6.16	4.96	5.92	7.65	7.00	7.84	0.00	4.64	4.61	15.48
Doc11	16.12	2.48	3.53	4.90	2.85	4.16	6.09	5.18	6.54	4.64	0.00	2.36	13.40
Corpus	15.47	2.61	3.88	4.88	3.23	4.34	5.71	5.28	6.20	4.61	2.36	0.00	11.70
Peculiar lexicon	27.25	13.18	15.15	16.14	13.57	15.75	16.80	16.49	17.02	15.48	13.40	11.70	0.00

This is confirmed by the low cover rates (second and third criterion) in Table 3. In the same example, document *Doc11* is instead the best semantically represented according to the second and third criterion (Table 2, Table 3).

Table 3. Cover rates of each document, the corpus and the lexical peculiar index (example in the notary domain).

	Doc1	**Doc2**	**Doc3**	**Doc4**	**Doc5**	**Doc6**	**Doc7**	**Doc8**	**Doc9**	**Doc10**	**Doc11**
Cover rate respect to corpus	6.02	34.02	19.50	14.35	23.51	16.40	14.41	14.03	12.43	16.00	43.11
Cover rate respect to lexical peculiar index	2.02	36.36	10.10	8.08	26.77	7.07	8.59	7.07	8.59	11.10	31.82

Moreover, it is possible to note that, for the application we have shown, the fourth criterion is fully confirmed (Table 4).

Table 4. The χ^2 distance among the corpus and the peculiar lexical items (313 lemmas) by using the proposed method and the peculiar lexical items (198 lemmas) derived by the exogenous method (example in the notary domain).

	Corpus	Peculiar lexicon by the proposed method	Peculiar lexicon by the exogenous method
Corpus	0.00	2.98	4.63
Peculiar lexicon by the proposed method	2.98	0.00	6.43
Peculiar lexicon by the exogenous method	4.63	6.43	0.00

6. Conclusion

In this work we have presented a strategy for refining the peculiar lexicon extracted from a corpus belonging to a specialist domain.
The proposed strategy is the starting point for defining an ontological model to be used in a system for the management of documents belonging to a specialized domain, suitable for various applications, such as e-Government, web services interoperability, semi-structured document processing, focus group and discourse flow analysis[15].
This work proposes a methodology for a first characterization of the domain of interest, through the selection of a peculiar lexicon, based on the iterative refinement of a list of terms extracted on the basis of their associated TF-IDF index and the computed χ^2 distance between this lexicon and the corpus terms. The restricted area of specialization reduces the intrinsic semantic ambiguity of the words, relating to the generic domain, thereby allowing a more accurate semantic processing.
The strategy is applied on a corpus of documents belonging to a juridical domain whereas, in the future, efforts 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. *In MTISD 2008: Methods, Models and Information Technologies for Decision Support Systems*, Ed. Università del Sannio, Lecce, 18 - 20 September 2008, 313-316.

[2]. Balbi, S., Di Meglio, E. (2004). A Text Mining Strategy based on Local Contexts of Words. In *JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelles*, Presses Universitaire de Louvain, Louvain-la-Neuve, Belgique, 10 - 12 Mars 2004, 1, 79-87.

[3]. Balbi, S., Bolasco, S., Verde, R. (2002). Text Mining on elementary forms in complex lexical structures. In *JADT 2002: 6es Journées internationales d'Analyse statistique des Données Textuelles*. Saint-Malo: IRISA-INRIA, 13- 15 mars 2002, 89-100.

[4]. Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V., Soria, C. (2004). Automatic Classification and Analysis of Provisions in Italian Legal Texts: A Case Study. In WORM 2004: The *Second International Workshop on Regulatory Ontologies 3292/2004*, Springer-Verlag , Berlin Heidelberg, 26 October 2004, 593-604.

[5]. Bolasco, S. (2004). L'analisi statistica dei dati testuali: intrecci problematici e prospettive. In *Applicazioni di analisi statistica dei dati testuali*, eds. Bolasco S., Cutillo E. A. Roma: Casa Editrice Università La Sapienza, 9-26.

[6]. Bolasco, S., Pavone, P. (2007). Automatic dictionary and rule-based systems for extracting information from text, Classification and Data Analysis. In: *7° Meeting of the Classification and Data Analysis Group of the Italian Statistical Society*, Catania, Italy, 9-11 September 2009, 255-258.

[7]. Buitelaar, P., Cimiano, P., and Magnini, B., editors (2005). Ontology Learning from Text: Methods, Evaluation and Applications. In *Volume 123 of Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press, 569-572.

[8]. Buitelaar, P., Cimiano, P., and Magnini, B., G. (2008). Acquiring Legal Ontologies from Domain-specific Texts. In *LangTech 2008*, Rome, 28-29 February 2008, 28-29.

[9]. Gangemi, A., Sagri, MT., Tiscornia, D. (2003). A constructive framework for legal ontologies. In *Law and the Semantic Web*, eds. V.R. Benjamins, P. Casanovas, J. Breuker, A. Gangemi, Berlin-Heidelberg: Springer, 97-124.

[10]. Giovannetti, E., Marchi, S., Montemagni, S., Bartolini, R. (2008). Ontology Learning and Semantic Annotation: a Necessary Symbiosis. In *LREC 2008: Sixth International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA)*. Marrakech, Morocco, 26-30 May 2008, 2079-2085.

[11]. Lame, G. (2005). Using NLP techniques to identify legal ontology components: concept and relations. *Lecture notes in Computer Science vol. 3369*. Berlin Heidelberg: Springer 169-184.

[12]. Lebart, L. (1995). Analyse statistique des données textuelle: quelques problèmes actuel et futures. In JADT : *International Conference Journées d'Analyse statistique des Données Textuelles*, CISU, Rome, 9-11 June 2010, 45-55.

[13]. Lebart, L., Salem, A., Berry, L. (1998). *Exploring Textual Data. Dordrecht*, NL: Kluwer Academic Publishers.

[14]. Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2,157-165.

[15]. Nitti, M., Ciavolino, E., Salvatore, S., Gennaro, A. (2010). Analyzing psychotherapy process as intersubjective sensemaking: An approach based on discourse analysis and neural networks. *Psychotherapy Research*, 20(5), 546 -563.

[16]. Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5),503-520.

[17]. Salton, G. (1989). *Automatic Text processing: The Transformation, Analysis and Retrieval of Information by Computer*. Boston: Addison Wesley.