Semi automatic Extraction of a Peculiar Vocabulary in Notary Domain

Flora Amato, Rosanna Canonico, Antonino Mazzeo, Antonio Penta and Antonio Picariello

Università di Napoli "Federico II", Italy
Dipartimento di Informatica e Sistemistica, via Claudio 21, 80125, Naples
flora.amato@unina.it, rosannacanonico@libero.it, mazzeo@unina.it, a.penta@unina.it, picus@unina.it

Abstract: *The bureaucratic domain and the notary one, in particular, are characterized by a huge amount of unstructured information. In order to opportunely manage the knowledge contained within these documents for structuring, indexing and retrieval purposes, a suitable semantic-lexical approach requires a domain vocabulary useful for a quick identification of relevant information.*
In this paper we provide a description of a system for semi-automatic extraction of a terminological vocabulary, representative of the notary domain, based on the analysis and processing of a significant collection of notary documents. In addition, the extracted peculiar lexicon will provide the basis for the construction of the domain conceptual system, able to perform semantic processing of the document contents

Keywords: Peculiar Lexicon, Peculiar Vocabulary, NLP, Legal Information System.

1. Introduction

The exponential growth of digital documents has currently pointed out the need of an easier access to the contained knowledge.

Classic Information Retrieval (IR) systems generally fail to bring users to the desired documents which best satisfy their information needs, thus creating problems of information overload: as a matter of fact, the quantity of information to be handled is enormous, consequently, to access this kind of information is more and more difficult. Therefore, the possibility to make searches based on textual data provided with explicit semantic content would surely open the doors to a more intelligent information retrieval and the textual documents could really become machine understandable as well as machine readable.

Nowadays the research is engaging in the development and in the implementation of methods and technologies for the syntactic and semantic processing of textual documents in order to ensure an effective management and an intelligent sharing of knowledge.

An intelligent management of information requires some fundamental steps:

i. Specification of the macro and micro structure of the text;
ii. Indexing and extraction of the peculiar terminology;
iii. Construction of a terminological and conceptual knowledge base.

A text, considered as stream of characters, words and sentences, is a huge source of information, but a text is also a complex entity where data are correlate according to multiple levels of organizations. The comprehension of the structure of the text means accessing to its semantic contents.

It is finally evident the need of an infrastructure composed by integrated linguistic and statistic resources able to i) transform the implicitly knowledge contained in textual documents into explicitly structured knowledge; ii) correlate, in the field of specific domains, the meanings of textual data in order to give a representation of their semantic potential; iii) ensure a semantic retrieval allowing us the access to the documental base not by queries based on key-words but by contents.

The paper is organized as follows. Section 2 will present a system architecture for extracting a peculiar in the notary domain; section 3 describes how to produce suitable RDF triples for storing and managing the extracted information: eventually, some concluding remarks will be given in section 4.

2. Semi-automatic extraction of a peculiar lexicon for notary domain

In this section we provide the description of a system for semi-automatic extraction of a terminological vocabulary representative of notary domain, starting from a significant collection of notary documents.

Our approach is based on the idea that words are the basis of any textual document: consequently, the identification of words is a prerequisite for our textual analysis. Not all the words are useful for identifying the semantic of a documental corpus: this is the case, for example, of grammatical words that, even forming the connective tissue of a text, represent “noise” since they are not vehicle of meaningful contents.

Thus, let us consider as *peculiar lexicon* the set of relevant lexical items whereas the extraction of peculiar lexicon is to be described as identification and selection, among the different lexical items, of the most significant and representative key-words, in order to define the contents of the single textual fragments and in general, the whole domain whose corpus is a representative sample set.

Once extracted, the peculiar lexicon will provide the basis for the construction of the *domain conceptual system* able to perform a semantic processing of the documental contents by working with the meanings of the resources.

Identifying the most significant words in a text is a very difficult operation. Our system is founded on both linguistic and statistical approaches that are deeply integrated between them: the former goes into the linguistic structures of the text by analyzing the meaning of words; the latter, instead, provides quantitative representations of the identified phenomena.

2.1 System overview

Acquisition, loading and parsing of data. The first stage consists in the acquisition of the set of notary documents in a textual format (such as <.txt>, <.doc>, <.rtf>) processable by the system for further analysis. Once loaded, the collection is submitted to a parser for the definition and computation of the characters.

Pre-processing: tokenization and normalization. Objective of the tokenization process is the recognition and the segmentation of the single documents into minimal units of analysis (*tokens*). The process of normalization, instead, aims to standardize the orthographical variants of the same words in order to ensure a uniform processing of these items. As a result of this stage, a list of all different type words (or graphic forms) is extracted, the so-called *vocabulary*.

Lexical analysis – Statistical measures on vocabulary. Statistical measurements on vocabulary enable us to gather statistical and mathematical indexes on the vocabulary formerly extracted and on its classes of frequency: range of frequencies, normalized frequencies, indexes of lexical variety, ranks, as well as percentage of hapax, number of tokens and type words.

Lexical analysis - Identification of significant chunks and lexicalization. Objective of this stage is the selection (through the computation of the absorption coefficient) of semantically relevant sequences to be lexicalized, that is transformed in one token since its overall semantic content is greater than the one inside the single items. After this processing, the vocabulary results to be modified since the total number of occurrences tends to decrease whereas the size of vocabulary tends to increase.

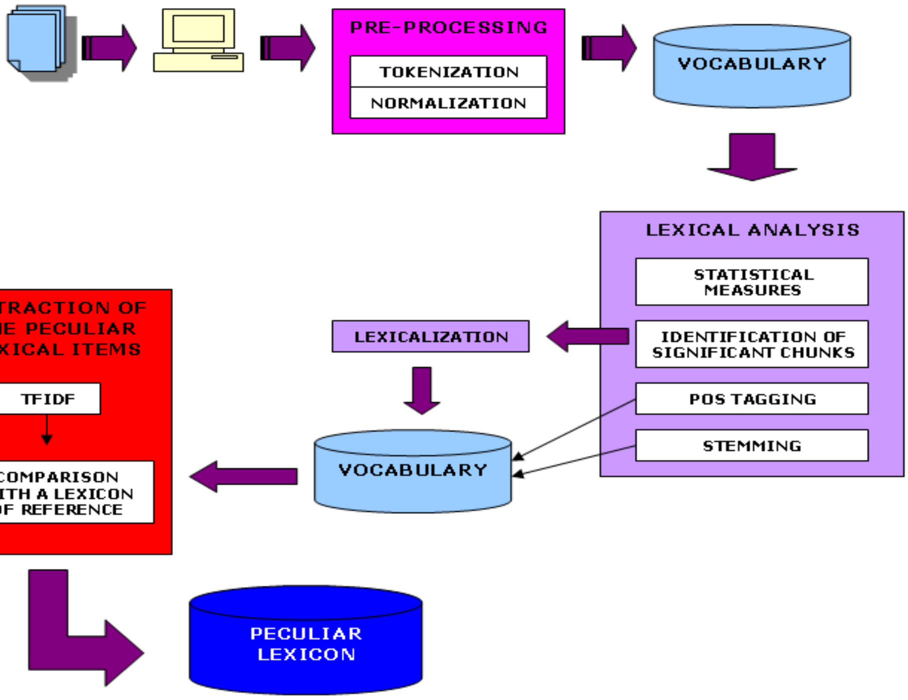

Figure 1: Semi-automatic system for extraction of peculiar lexicon

Fragment of List of relevant segments	
trascrizioni pregiudizievoli	parte promittente
base imponibile	assistenza dei testimoni
collegio notarile	raccolta n.
ipoteca legale	repertorio n.
parte acquirente	piena ed esclusiva proprietà
notaio autenticante	quietanza di saldo
parte venditrice	sottoscrizione delle parti
a carico di	comunione legale dei beni
ai sensi di	trasferimento della proprietà
immobile compravenduto	uso e abitazione
servitù apparenti	imposta di registro
parte cedente	diritto di prelazione
gravami non apparenti	atto di compravendita
…	

Table 1: List of relevant segments

Lexical analysis - Part of Speech tagging and stemming. Objective of POS tagging is the identification and the assignment of a grammatical category to each token. This allows to increase the precision in the information retrieval, sorting out the ambiguity of certain lexical forms which can belong, according to the co-text, to different grammatical categories: the *analysis of the concordance* permits to discriminate the different uses and the different meanings of a same word. Second objective is the stemming of the vocabulary: each lexical item is reduced to its stem in order to obtain from the vocabulary of the corpus a list of stemmed words.

Extraction of the peculiar lexical items. The strategies for the extraction of the peculiar lexicon are:

1. Extraction of the TF-IDF index (Term Frequency Inverse Document Frequency) through which it is possible to extract the most relevant lexical forms because concentrated only on few documents;

2. Comparison with other lexicons for the recognition of shared and not-shared words, and for the identification of the lexical items which are over or under- used with respect to the lexicon of reference.

Objective of this stage is, 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 secondly the lexical database of JurWordNet in order to extract a more and more specialized lexicon.

3. Extraction of RDF Triple based on the Peculiar Lexicon

The list of segments resulting by the vocabulary creation processing are codified in RDF triples containing all the relevant concepts retrieved in the buying-selling document.

We have implemented a prototype system in JAVA on top of the Oracle 10g back ends that is able to manage RDF technology.

The system implements functions to support most of the operations described in this paper. Note that some operations still require a manual contribute of domain experts, as the morpho-syntactic annotation and the identification of the relevant segments based on the evaluated absorption coefficient.

We have tested our prototype over a collection of about 100 notary documents belonging to two main categories:

- Buying-Selling Notary document and
- Notary Data Base (documents belongings to the italian “Banca Dati Notarile”)

Qualitative evaluation performed by domain experts has showed that the list of extracted words and segments are really relevant for the domain in examination.

4. Conclusion and Future Works

In this paper we have presented a procedure to extract relevant words and segment starting from documents belonging to Notary Domain.

The extracted information, that are codified in RDF triples, can be used in future works in order to design an ontology able to codify the concepts of interest belonging to the notary documents.

Bibliography

Visser, P., (1996), The formal specification of a legal ontology.

Tiscornia, D., (1996), Some ontological tools to support legal regulatory compliance, with a case study. Workshop on Regulatory Ontologies and the Modeling of Complaint Regulations (WORM CoRe 2003) Springer LNCS.

Cruse, D.A. (1982), Lexical Semantics. Cambridge University Press.

Bolasco S., Pavone P. (2008). Multi-class categorization based on cluster analysis and TFIDF, in S. Heiden & B. Pincemin (eds.), JADT2008, Presses Universitaires de Lyon, vol. 1, pp. 209-218.

Bolasco S., Pavone P. (2007). Automatic dictionary and rule-based systems for extracting information from text, in Classification and Data Analysis 2007. Book of short papers CLADAG2007. EUM - Edizioni Università di Macerata, pp. 255-258.