# Journal of Applied Sciences

ISSN 1812-5654

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 11 (4): 639-646, 2011
ISSN 1812-5654 / DOI: 10.3923/jas.2011.639.646

# Statistical and Lexical Analysis for Semi-automatic Extraction of Relevant Information from Legal Documents

F. Amato, R. Canonico, A. Mazzeo and A. Picariello
Dipartimento di Informatica e Sistemistica, Universita di Napoli Federico II,
via Claudio 21, 80125, Napoli, Italy

**Abstract:** The bureaucratic domain and the legal one, in particular, are characterized by a huge amount of information. In order to opportunely manage the knowledge embedded within documents for structuring, indexing and retrieval purposes, a suitable statistical-lexical approach is required for a quick identification of relevant and peculiar information. The main goal of this study is to describe two integrated strategies for semi-automatic extraction of significant and peculiar terms, starting from a corpus of documents belonging to legal domain. The extracted lexicon will provide a basis for the construction of a conceptual system to be used as knowledge base supporting the semantic processing of documents.

**Key words:** Relevant language, peculiar lexicon, statistical and lexical analysis, lexical information extraction, NLP, Legal information system

## INTRODUCTION

The exponential growth of digital documents has currently pointed out the need of more and more complex strategies for the management of the embedded unstructured knowledge in terms of analysis and extraction of relevant information. To gain access to this kind of information is nowadays more and more difficult: only the possibility to make searches based on textual data provided with explicit semantic content could ensure a more intelligent information retrieval and an effective management, as well as an intelligent sharing of knowledge.

An intelligent management of information requires some fundamental steps:

- Specification of the macro and micro structure of the text
- Indexing and extraction of the relevant and peculiar terminology
- Construction of a terminological and conceptual knowledge base

Thus, in a document management process the first step is to make explicit the data structure, consequently the attention will be primarily drawn on the analysis of the lexical items since vehicle of specific conceptual meanings.

The methodologies and the techniques performed for the lexical analysis follow an inductive approach, totally based on data extracted from real texts (notarial deeds) and gathered in a corpus. The analysis will be characterized by Lenci *et al*. (2005):

- Stages of linguistics elaboration, concerning the structure of the language and the words
- Stages of statistical elaboration, making clear the relevant but not immediately noticeable, linguistic phenomena

It is finally evident the need of an infrastructure composed by integrated linguistic and statistical resources able to:

- Transform the knowledge implicitly embedded in textual documents into explicitly structured knowledge
- Correlate, in the field of specific domains, the meanings of textual data in order to give a representation of their semantic potential
- Enable a semantic retrieval, allowing the access to the documental base not by queries, based on key words, but by contents

In this study we provide a description of two integrated strategies enabling a semi-automatic extraction

**Corresponding Author:** Flora Amato, Dipartimento di Informatica e Sistemistica, Universita di Napoli Federico II, via Claudio 21, 80125, Napoli, Italy

of statistically relevant domain terms, starting from a collection of notarial documents. In addition, we will show how the extracted peculiar lexicon will provide the basis for the construction of a conceptual system to be used as knowledge base supporting the semantic processing of document.

The analyses are performed on a statistically significant corpus of Italian notarial deeds, for this reason the lemmas in the paper are reported in the Italian language.

## PROCEDURE OVERVIEW

Here, we provide a description of a procedure for semi-automatic extraction of relevant domain terminology, starting from a significant collection of notarial deeds.

The first step, that is preparatory for the real analysis, consists in the acquisition of the set of notarial documents in a digital textual format (such as .txt, .doc, rtf).

Statistical and lexical analysis is composed by three different stages:

- Corpus Pre-processing
- Vocabulary Analysis
- Extraction of Relevant Terminology

These stages will enable the corpus exploration in order to analyze its contents and extract statistically relevant domain terms, that is corpus key-words that are the most semantically discriminating words because descriptors of the contents of the corpus with respect to the domain terminology.

Our analysis is founded on both linguistic and statistical approaches: the former goes into the linguistic structures of the corpus by analyzing the meaning of words; the latter instead, provides quantitative representations of the identified phenomena.

**Corpus pre-processing:** The pre-processing stage is essential for a correct parsing of the corpus. It identifies units of analysis discriminating sequence of tokens, which are related to the same concepts. In particular, two kinds of units are identified:

- **Simple:** Identified by a single word (as imposta, the italian word standing for duty)
- **Complex:** Identified by a sequence of related words (imposta di bollo standing for stamp duty) or phrases to be considered as single unit in the syntax of the sentence

Thus, the identification of these forms results in a fundamental pre-condition for a correct text analysis: for a computer a text is a sequence of alphanumeric strings, so a computer is not able to understand when a sequence of strings correspond to a lexical item. Thus, it is necessary to provide the computer with knowledge on the structures and conventions of the language used (Lebart *et al*., 1998).

The pre-processing stage is composed by tokenization and normalization.

Tokenization stage consists in the segmentation of the corpus into minimal units of analysis (tokens). The identification of tokens in a text is not a simple operation, since a token is a heterogeneous lexical item represented by simple words phrasal expressions, alphanumeric expressions, abbreviations or acronyms. The correct text tokenization requires some fundamental steps:

- Definition and computation of alphabetic and not alphabetic characters, to be considered as separators
- Disambiguation of punctuation marks: these marks must be considered as independent tokens, but there are cases where the same mark can have different uses producing different outcomes. To give an example, the point is to be considered as independent token at the end of the sentence but the same mark is to be considered together with the element with which it is used in the case of acronyms or abbreviations
- Unification in a single token of several words in order to identify complex units (for example, sequences as name and surname (Mario Rossi), dates (8 February 1980), monetary expressions (3 Euro), laws (dpr 28 dicembre 2000, n. 45)

A Tokenizer, provided with the necessary knowledge on the language conventions, is able to work by identifying the borders of the token and making the necessary transformations, by separating the sequences of strings to be considered as independent tokens, and by unifying the sequences to be considered as one token.

Actually, some words tend to co-occur with repetitiveness, producing relevant lexical chunks such as:

- Noun phrases (indicated as NP), often corresponding to Italian technical expressions such as capitale sociale, consiglio d'amministrazione, collegio notarile, distretto notarile, base imponibile, imposta di bollo, etc.
- Verb phrases (VP), such as avere ad oggetto, fare eccezione, fare riferimento, etc.

Table 1: List of the lexicalized phrases with higher IS

| Phrasal expression | IS | Phrasal expression | IS | Phrasal | IS | Phrasal expression | IS |
|---|---|---|---|---|---|---|---|
| Societa a responsabilita limitata | 7.12 | Collegio notarile | 3.30 | Parte acquirente | 2.28 | Rappresentanza della società | 1.12 |
| Repubblica italiana | 4.00 | Cittadini italiani | 3.27 | Norme urbanistiche | 2.26 | Socio semplice | 1.12 |
| Trascrizioni pregiudizievoli | 4.00 | Porzione immobiliare | 3.20 | Parte venditrice | 2.19 | Consiglio di amministrazione | 1.08 |
| Organo amministrativo | 3.82 | Valore nominale | 3.20 | Modalità di pagamento | 2.11 | Cessione di quota | 1.07 |
| Assistenza dei testimoni | 3.69 | Amministratore unico | 3.05 | Esercizio sociale | 2.04 | Società venditrice | 0.94 |
| Diritti ed obblighi condominiali | 3.64 | Registro delle imprese | 2.81 | Concessione edilizia | 1.71 | Quote sociali | 0.92 |
| Complesso edilizio | 3.56 | Operazioni commerciali | 2.75 | Quietanza di saldo | 1.69 | Piena proprietà | 0.92 |
| Gravami non apparenti | 3.50 | Strumenti urbanistici | 2.74 | Diritto di prelazione | 1.52 | Sede sociale | 0.88 |
| Identità personale | 3.45 | Ipoteca legale | 2.52 | Catasto fabbricati | 1.50 | Sede legale | 0.79 |
| Normativa vigente | 3.43 | Regolamento di condominio | 2.52 | Legale rappresentante | 1.24 | Assemblea dei soci | 0.61 |
| Scrittura privata | 3.38 | Servitù apparenti | 2.36 | Libro sociale | 1.18 | Atto di compravendita | 0.52 |

- Prepositional phrases (PrepP), such as a margine di, a carico di, da parte di, in regime di, etc.

Many phrase-structures are often outcomes of standardized linguistic uses: they are very used in jargons and are often specializations of generic terms (for instance, in the legal domain we can find the words imposta and its specialization imposta di bollo).

Normalization stage serves to level orthographical variants in order to ensure a uniform processing of lexical items. As a matter of fact, a same word in the same text can appear in different ways:

- Compounds or prefixed words can be, for example, separated by a hyphen or a blank: data-base, data base
- The same date can be written differently: 8 February 1980, 08/02/1980, etc.
- The same word can be abbreviated in different ways: the word page can appear as pag or pg
- The same acronym can be written differently: USA or U.S.A.
- The same word can appear in small or in capital letter: in this specific case to simply cancel this distinction would make it difficult for a computer the identification of the beginning of a sentence or the distinction between a name of person (such as Rosa) and the name of a flower (e.g., rosa) or even to recognize the distinction between an acronym (e.g., USA) and a verb (e.g., USA, 3rd sing. pers.)

Pre-processing operations will be performed with the help of software for lexical and textual processing named Taltac (http://www.taltac.it/it/index.shtml), enabling the identification of simple units and relevant lexical chunks as well as the vocabulary standardization. The presence of specific structures, such as proper nouns, compounds, abbreviations, acronyms or common phrasal expressions will be performed through comparisons with glossaries and thesauri, enumerating well-known structures to be identified as single token.

In particular, the selection of new semantically relevant chunks, not immediately recognized during the tokenization process, will be performed by computing the Index of Significance (IS) which filters the relevant chunks in accordance with their capacity of absorption of the occurrences of the compositional words. Obviously, the more a lexical segment will be relevant the higher the segment power of absorption of the single words composing it will be. By applying the IS index together with an empirical analysis, we have produced a list of 44 semantically relevant chunks from the corpus under examination. Then these chunks, reported in Table 1 have been lexicalized, in order to be transformed in one single occurrence (or token).

Most of the identified chunks are noun phrases, in particular sequences Noun-Adjective and Noun-Prepositional Phrase.

Outcome of the pre-processing module is a list containing all the different words of the document collection, the so-called vocabulary, expressed through graphic forms and occurrences, that is the number indicating how often a given lexical form appears in the corpus.

**Vocabulary analysis:** Our approach is based on the idea that words are vehicle of the conceptual contents of a domain, consequently, the analysis of the corpus vocabulary, together with the computation of the words, is a fundamental prerequisite: the lexical analysis integrated with statistical techniques will enable a systematic exploration of the corpus lexical structure (dimensions, occurrences, prevailing lexical categories, subtext specific lexical forms), ensuring the depth and the relevance of the interpretations.

In this paragraph we provide a description of some fundamental strategies for vocabulary analysis, resulting in a real study of the language used in the corpus and in single subtexts (Bolasco and Pavone, 2007).

The lexicometric analysis of vocabulary enables to gather quantitative and qualitative information on the vocabulary formerly extracted, in terms of:

- Corpus dimension N (total number of occurrences or word-tokens) and vocabulary dimension V (number of graphic forms or word-types)
- Indexes of lexical variety, such as type/token ratio V/N and percentage of hapax legomena (i.e., words occurring only one time in the corpus) $V_1/V$
- Indexes of word frequency, such as relative, normalized and cumulative frequency, sub-occurrences, etc.

Our corpus is composed by N = 162.628 word-tokens corresponding to V = 13.466 word-types. The type/token ratio (V/N = 0,0828) and the percentage of the hapax ($V_1/V$ = 42,46%) point out the presence of a poor language. As a matter of fact, the values of the type/token ratio vary from 0 to 1, consequently, values near 0 indicate a poor vocabulary; the highest value 1 is obtained when the vocabulary dimension is equivalent to the corpus dimension, that is when the vocabulary is entirely formed by hapax.

Arranging the graphic forms according to decreasing values of frequency, it is possible to identify ranks and distribution of words in ranges of frequency:

- High-frequency range, containing a low number of lexical items, most of which are grammatical words (articles, prepositions and conjunctions) occurring with high and different frequencies
- Middle-frequency range, containing a number of words higher than the previous range, that are different for typology and occurrences
- Low-frequency range, containing the most part of the vocabulary and including all the classes of frequency descending until 1

A distributive analysis of the corpus lexical items permits to observe that the most semantically relevant words tends to be in the middle and low ranges of frequency, differently from grammar words that usually are in the high range.

Grammatical words are generally qualified as functional words, having little lexical meaning since they serve to express grammatical relations with other words within the sentence: they are frequently used and in a very predictable way, independently from the text typology or the topic handled. Words like nouns, verbs, adverbs and adjectives, instead, are usually classified as content words: generally nouns indicates entities (people, things or places); verbs are used to denote actions, states or processes; adjectives are indicators of qualities or properties of nouns; adverbs represent modifiers (directional/locative, modal, time) of other classes. All these words serve to semantically discriminate the texts; in addition, the rarest words are the most informative ones.

In our vocabulary we can find at the first ranks grammatical words like prepositions and articles: only one word (numero) out of 13 doesn't belong to the class of the grammatical words.

The graphic forms numero, atto, articolo and societa occupy respectively the ranks 9, 16, 18 and 23 and they are the first content or lexical words. We have empirically observed that the most part of notarial domain terms, (such as statuto, quota, vendita, socio, proprieta, etc.) belong to the low-range of frequency.

The lexicometric analysis points out that frequency values cannot determine if a word belongs to a certain domain: domain terms can have both high and low values of frequency, consequently even hapax can be interesting.

With Part-of-speech tagging the main goal is assigning a lexical category to each identified lexeme by solving all the ambiguousness of the language. Actually, different lexical categories can be assigned to a same word dependently from the syntactic and semantic context. Thus, the automated tagging implies two kinds of problems:

- Finding the tag or all the possible tags for each lexical item
- Choosing, among all the possible tags, the correct one

The first problem can be solved by using a glossary as reference; the second problem, that is a problem of word-category disambiguation (or syntactic disambiguation), can be solved by:

- Using contextual evidences, that is examining the context where the word is used
- Using probabilistic evidences starting from a tagged corpus to be used as training set

In our case, the Key-words in Context (KWIC) analysis has been performed for the disambiguation. This analysis performs a systematical study of the local context (or co-text) of the word to be disambiguate: for each ambiguous word in the corpus, it is possible to observe all its occurrences in the text, that is all the words preceding (left surrounding) and following it (right surrounding), in order to identify its correct lexical category and therefore its meaning.

As can be in Table 2, the word pubblico can be a verb (1st occurrence), an adjective (2nd occurrence) or

Table 2: Example of word pubblico in context

| | | | |
|---|---|---|---|
| c.1 | Richiesto io notaio ho ricevuto il presente atto che | pubblico | Mediante lettura da me datane ai comparenti che lo |
| s.4 | Ente locale ed ha personalità giuridica di diritto | pubblico | . L' Unione è costituita allo scopo di esercitare |
| s.5 | Qualunque attività di raccolta del risparmio presso il | pubblico | . Articolo 5 = capitale sociale il capitale sociale |
| s.5 | Qualunque attività di raccolta del risparmio presso il | pubblico | . Art . 4 - capitale sociale , diritti dei soci |

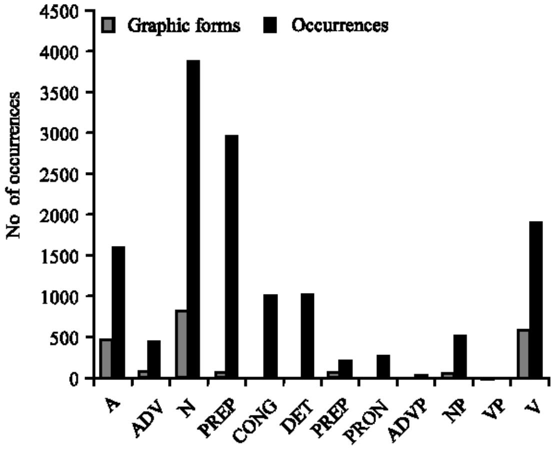

Fig. 1: Incidence of POS tags between graphic forms and occurrences

Table 3: Morphemic Information

| TAG | Graphic forms | %Graphic | Occurrence | %Occurrences |
|---|---|---|---|---|
| Conjv-imperf-s-3 | 2 | 0.06 | 2 | 0.01 |
| Conjv-pres-pl-3 | 12 | 0.35 | 18 | 0.10 |
| Conjv-pres-s-1/2 | 7 | 0.20 | 12 | 0.07 |
| Ger-pres-indf | 19 | 0.56 | 35 | 0.20 |
| Indic-fut-pl-3 | 9 | 0.26 | 23 | 0.13 |
| Indic-fut-s-3 | 20 | 0.59 | 52 | 0.29 |
| Indic-imperf-pl-3 | 1 | 0.03 | 1 | 0.01 |
| Indic-imperf-s-3 | 1 | 0.03 | 1 | 0.01 |
| Indic-pres-pl-3 | 44 | 1.29 | 128 | 0.72 |
| Indic-pres-s-1 | 5 | 0.15 | 11 | 0.06 |
| Indic-pres-s-3 | 43 | 1.26 | 301 | 1.69 |
| Inf-pres-indf | 105 | 3.07 | 237 | 1.33 |
| Past-part-pl-f | 15 | 0.44 | 32 | 0.18 |
| Past-part-pl-m | 46 | 1.35 | 92 | 0.52 |
| Past-part-s-f | 80 | 2.34 | 220 | 1.23 |
| Past-part-s-m | 90 | 2.63 | 352 | 1.97 |
| Pres-part-pl-indf | 14 | 0.41 | 29 | 0.16 |
| Pres-part-s-indf | 14 | 0.41 | 44 | 0.25 |
| Past | 266 | 44.19 | 754 | 39.77 |
| Present | 330 | 54.82 | 1110 | 58.54 |
| Future | 29 | 4.82 | 75 | 3.96 |

a noun (3rd and 4th occurrences). We have chosen among these three possible word categories, having more representations in the corpus.

After the tagging process, the ambiguous lexical items have been manually disambiguated: mostly of them were present and past participles. Generally, present and past participles can also have function of adjective (parte acquirente, documenti annessi) and noun (l'acquirente, il convenuto): we have decided to tag participle forms as verbs only when supporting a direct or indirect object.

POS tagging process has been performed with the software Taltac that has permitted the identification of nouns, verbs, adjective, adverbs, articles, prepositions, conjunctions, exclamations, pronouns as well as abbreviations, numerals and dates. In addition, we have manually classified the previously lexicalized forms and other automatically identified chunks as noun phrases, verb phrases, preposition phrases and adverb phrases.

Figure 1 presents some statistics about word lexical categories in terms of graphic forms and occurrences: nouns, verbs and adjectives represent the most relevant lexical categories in terms of graphic forms; outcomes are different with regard to the total occurrences (The results do not take into account abbreviations, exclamations, numerals and proper nouns representing the 32% of the graphic forms). The used x-axis variables, in particular, correspond to parts of speech classes, in which words are traditionally grouped into: Adjectives (A), adverbs (ADV), Nouns (N), prepositions (PREP), conjunctions (CONG), articles or determiners (DET), prepositional phrases (PREPP), pronouns (PRON), adverbial phrases (ADVP), noun phrases (NP), verb phrases (VP), Verbs (V). In the y-axis the number of corpus occurrences and graphic forms for these classes is reported.

POS tagging information have been integrated by further morphemic specifications, such as inflectional information (Inflection is the way language handles grammatical relations and relational categories such as gender (masculine/feminine) and number (singular/plural) for nouns; tense, mood, person, voice and aspect for verbs).`

As shown in Table 3, the clearly prevalent use of third person (singular and plural) is evidence of a predominant referential function in the text: the notarial deed attests declarations, facts and events where the will of the contracting parties is identifiable. Further evident are the prevalent use of the indicative mood and the present tense: they are mood and tense of objectivity and certainty, of the purely noticed verbal action, where sentences often have the form of statements.

POS tagging is performed together with word-stemming: each lexical item is reduced to its lemma (For the Italian language, the stem corresponds to the singular masculine/feminine noun, the singular masculine adjective and the infinitive verb) in order to obtain a list of canonical forms (or citation forms) as well as graphic words (that are nothing but inflected lexical forms). The stemming process is performed by a stemmer provided with some reference lists, for example:

Table 4: List of subtext typical lexical items

| Purchase deed | | | Partnership formation | | |
|---|---|---|---|---|---|
| Noun phrases | Nouns | Verbs | Noun phrases | Nouns | Verbs |
| Base imponibile | Appartamento | Accettare | Amministratore unico | Bilancio | Convocare |
| Complesso edilizio | Catasto | Acquistare | Assemblea dei soci | Deliberazioni | Nominare |
| Ipoteca legale | Compravendita | Confinare | Attività sociale | Durata | |
| Parte acquirente | Gravami | Dichiarare | Capitale sociale | Partecipazioni | |
| Parte venditrice | Immobile | Garantire | Collegio sindacale | Sede | |
| Porzioni | Planimetria | Rilasciare | Consiglio | Società | |
| Quietanza di saldo | Proprietà | Trasferire | Esercizio sociale | Socio | |
| | Vendita | Vendere | Organo amministrativo | Statuto | |

- A list of derivation and suffixation affixes
- A list of stems
- A mini-grammar containing the combination rules between stems and affixes

Word-stemming enables a more general approach to the corpus vocabulary analysis.

POS tagging and word-stemming will be useful later for the extraction of relevant lexical categories, such as nouns and noun phrases, verbs and verb phrases, with the exclusion of other categories, that are not useful for a semantic characterization of the corpus.

Studying a corpus vocabulary also means to identify subtext typical lexical items, that is identifying possible significant differences in terminology within the same document collection by comparing the different text typologies.

The attribute Type has been assigned to each single document fragment in the corpus, therefore, notarial deeds have been divided into two subtexts: Purchase deed and Partnership Formation. The analysis of lexical specificities will enable to compute if and how much a word is typical and specific of one of these two subtexts.

At the base of this computation there is a probabilistic hypothesis of equi-distribution of the lexical items in the corpus: a word is identified as specific of a subtext if it shows (with respect to a threshold probability) a significant over-use in this subtext in opposition to the other subtext. As a matter of fact, if the over-used word shows all its occurrences, significantly concentrated within one subtext, it is to be considered typical of that subtext.

Table 4 represents some over-used lexical items describing the peculiarity of the subtexts they belong to, having considered as threshold the value $p<0.025$.

Lexical items characterizing Purchase deeds, for instance, identify the formal agreement through which the selling party transfers to the buying party the ownership of a real estate on payment of the equivalent sum of money. There are many words describing the transferred asset as well as words indicating its location, boundaries and cadastral data. Obviously, there are words regarding the selling price and the ways of payments as well as possible guarantees or charges.

The analysis of subtext typical lexical items has produced a list of 74 lemmas for Purchase deeds (corresponding to the following graphic forms: 52 nouns, 9 noun phrases, 14 verbs) and 64 lemmas for Partnership Formation documents (corresponding to the following graphic forms: 51 nouns, 6 noun phrases and 13 verbs).

**Extraction of domain peculiar terminology:** In this paragraph, we provide a description of two integrated strategies enabling the extraction of a set of terms serving to semantically characterize the analyzed corpus and, more in general, to have a minimum terminology peculiar of the domain at issue.

The extraction of the peculiar domain terminology involves two different integrated strategies:

- Use of endogenous resources (statistical approach) for the creation of a sub-set containing the most significant and corpus representative key-words
- Use of exogenous resources (lexical approach) for the selection, from the previously obtained word-subset, of terms that are typical and peculiar of the domain at issue since designating specific concepts of the domain itself

Firstly, the TFIDF index (Term Frequency Inverse Document Frequency) is computed on the corpus vocabulary (Bolasco, 2005), enabling the assignment of a weight to each word: this index is computed on the base of the term frequency and its distribution within the corpus (Balbi and Di Meglio, 2004a). TFIDF index is, in fact, based on:

- Term frequency (tf), corresponding to the number of times a given term occurs in the collection: the more a term occurs in the same document, the more it is representative of its content
- Inverse document frequency (idf), concerning the term distribution within the collection: it relies on the principle that term importance is inversely proportional to the number of documents from the corpus where the given term occurs. Thus, the more documents contain that given term, the less discriminating it is

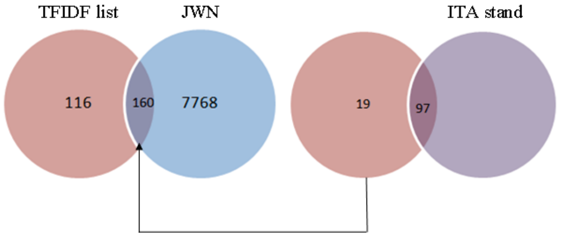

Fig. 2: Lexical comparisons

Therefore, TFIDF enables the extraction of the most discriminating lexical items because frequent and concentrated on few documents (Balbi and Di Meglio, 2004b). The computation of the index has enabled the extraction of the following graphic forms from the analyzed corpus: 203 nouns (out of 837), 36 noun phrases (out of 79), 90 verbs (out of 606) and 1 verb phrase (out of 3) producing a list of 276 lemmas.

In order to specialize the list obtained with respect to the analyzed domain, we need to resort to the comparison with exogenous resources that is external to specialized lexicons.

The list of key-words has been firstly compared to JureWordNet (Gangemi *et al*., 2003) lexical database (7768 lemmas) in order to obtain an inventory of 160 words in common, that are the corpus key-words pertaining to the legal domain (Lame, 2005).

Then, this list of 160 lexical items has been further specialized by integration of words belonging to the notarial domain and identified from the remaining 116 words with no correspondence in JureWordNet (JWN). This identification has been performed by eliminating general words in common with a standard lexicon of the Italian language. This has produced a list of 19 corpus specific lemmas that have been integrated to the initial list of 160 lemmas. A schematic representation of the sets of words under comparison is depicted in Fig. 2, in which the cardinality of the involved sets is depicted too. The reported sets are: the discovered relevant words belonging to notarial domain (TFIDF LIST), the list of terms defined in JureWordNet (JWN) lexical database, the standard lexicon of the Italian language (ITA STAND) and the set at the beginning of the arrow, containing the discovered 116 words, belonging to the notarial domain, but not included in JureWordNet (JWN), whose 19 words, not used in ITA STAND, are exploited for the integration of the list of 160 peculiar terms.

This stage has, therefore, enabled the extraction of 179 terms from the corpus, that are the most semantically pregnant and technical words because descriptors of the contents dealt within the corpus. Furthermore, these terms represent a sub-set of the terminology of the respective specialist language, which designate the concepts of the domain of interest.

## TERMS, CONCEPTS AND EVENTS: EXTRACTION OF RDF TRIPLE BASED ON RELEVANT TERMINOLOGY

An ontological domain is composed by concepts and semantic relations among them, besides it contains specifications about processes, actions and events involving different entities performing different roles and functions (Staab and Studer, 2004).

The characterization of events involves the identification of predicative structures (statements) enabling the recognition of: i) parties in the event (Who? What?); ii) characterizing properties (Where? When? How?), iii) the event itself, generally denoted by the verb.

It is possible to obtain a conceptual organization of the extracted terminology by looking for regular syntactic patterns relating terms to identify semantic relations. For example, the use of prepositions can be indicative of a meronymic relation: in the following noun phrase prezzo della vendita, the prepositional item is evidence of a meronymic relation between the head of the phrasal expression (prezzo) and the word to which it is related (vendita), being one of its conceptual components. In other cases, the prepositional item can be index of a relation of semantic inclusion (hyponymy) as in imposta di bollo vs. imposta.

The list of the extracted relevant lexical items is codified in RDF (Resource Description Framework) triples containing all the relevant concepts retrieved in the notarial documents. The RDF standard permits both the structuring of the acquired knowledge and the representation of the context the resources belong to.

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

## CONCLUSION

In this study we have presented a semi-automatic procedure to extract meaningful terms starting from documents belonging to the notarial domain. Qualitative evaluation performed by domain experts has showed that the list of extracted lexical items is really relevant for the domain in examination.

The extracted information, that are codified in RDF triples can be used in future works in order to design lexical resources and specific knowledge bases as well as systems for the advanced management of queries in natural language, that is an ontology able to codify the concepts of interest belonging to the notarial documents.

Further works will be devoted to extend our analysis on different corpora, and compare results with outcomes of different methodologies, considering, for example approach based on discourse analysis like (Nitti *et al*., 2010).

## REFERENCES

Balbi, S. and E. Di Meglio, 2004a. A text mining strategy based on local contexts of words. JADT 2004: 7es Journees Internationales d'Analyse Statistique des Donnees Textuelles.

Balbi, S. and E.D. Meglio, 2004b. Contributions of Textual Data Analysis to Text Retrieval. In: Classification, Clustering and Data Mining Applications, Banks, D., L. House, F.R. McMorris, P. Arabie and W. Gaul (Eds.). Springer-Verlag, Berlin, pp: 511-520.

Bolasco, S., 2005. Statistica testuale e text-mining: Alcuni paradigmi applicativi. Quaderni Statistica, 7: 17-53.

Bolasco, S. and P. Pavone, 2007. Automatic dictionary and rule-based systems for extracting information from text, in classification and data analysis 2007. Book of Short Papers CLADAG 2007. EUM-Edizioni Universita di Macerata, pp: 255-258.

Gangemi, A., A. Prisco, M.T. Sagri, G. Steve and D. Tiscornia, 2003. Some ontological tools to support legal regulatory compliance, with a case study. Lecture Notes Comput. Sci., 2889: 607-620.

Lame, G., 2005. Using NLP techniques to identify legal ontology components: Concepts and relations. Law Semantic Web, 3369: 169-184.

Lebart, L., A. Salem and L. Berry, 1998. Exploring Textual Data. Kluwer Academic Publisher, Dordrecht.

Lenci, A., S. Montemagni and P. Vito, 2005. Text and Computer: Elements of Computational Linguistics. Carocci, Roma, ISBN-13: 9788843034253, (Original Article in Italian).

Nitti, M., E. Ciavolino, S. Salvatore and A. Gennaro, 2010. Analyzing psychotherapy process as intersubjective sensemaking: An approach based on discourse analysis and neural networks. Psychother. Res., 20: 546-563.

Staab, S. and R. Studer, 2004. Handbook on Ontologies (International Handbooks on Information Systems). 1st Edn., Springer, New York.