# Analysis and Evaluation of Techniques for the Extraction of Classes in the Ontology Learning Process

Rafael Pedraza-Jimenez, Mari Vallez, Lluís Codina, Cristòfol Rovira

Department of Communication, Pompeu Fabra University
Campus de la Comunicació, Roc Boronat 138
08018 Barcelona, Spain
{rafael.pedraza, mari.vallez, lluis.codina, cristofol.rovira}@upf.edu

**Abstract.** This paper analyzes and evaluates, in the context of Ontology learning, some techniques to identify and extract candidate terms to classes of a taxonomy. Besides, this work points out some inconsistencies that may be occurring in the preprocessing of text corpus, and proposes techniques to obtain good terms candidate to classes of a taxonomy.

**Keywords.** Semantic Web, Ontology engineering, Ontology learning, Text mining, Language processing.

## 1 Introduction

In 2001 Berners-Lee and his colleagues made known to the public at large the Semantic Web [1], a short, medium and long term project of the most important agency for the Web standardization: the World Wide Web Consortium (W3C). This proposal implied deep changes that would affect, and, in fact are already affecting, the fields of creation, edition and publication of web pages and sites.

The main goal of this project is to make understandable for machines the Web content [2]. However, three requirements would be necessary to make it possible: a) Web contents must be described: to this end different languages have been created, such as RDF [3], which allows the description of any resource on the Web with metadata. b) The different knowledge domains must be structured and formalized using ontologies [4]. c) Tools to interpret, compare, and merge data on a semantic base are needed: these tools work over ontologies, and they can be built using different languages. The most important of them is OWL [5].

Nevertheless, the formalization of Semantic Web [6], on the one hand describing their resources and on the other hand making ontologies, entails a high cost in time and money. As a result, in 2010 the Semantic Web is not yet a reality [7] and, although many of its technologies are already among us [8], the W3C has recently announced that the entire project can not be achieved in less than 10 years.

To solve the first of these problems several research groups, namely, the one that the authors of this paper belong to, DigiDoc (http://www.upf.edu/digidoc), are working in the development of editors and automatic extractors of metadata (such as DigiDocMeta: http://www.metaeditor.net). Regarding the second issue, in 2001 a new discipline developed, the Ontology Engineering [9], devoted to the study and the design of applications that help to develop, maintain and use these tools semi-automatically.

In this new discipline, the process called "Ontology learning" [10] is very important, which focuses on the generation of tools to import, extract, prune, refine and evaluate the taxonomy of an ontology semi-automatically.

This work is carried out in the Ontology learning field, and focuses on the analysis and evaluation of techniques commonly used to propose terms [11] that constitute the classes of the taxonomy resulting from this process.

This paper is structured as follows: the next section explains the ontology learning process; the following section sets out the main objectives of this research; the third section describes the methodology and tools used in experimentation. Then a discussion concerning the main results of this research is presented. Finally, some conclusions are stated.

## 2 Ontology learning

The Ontology learning [12] is a process carried out initially by a human expert, and consists basically of three stages. First, the expert gathers a corpus of documents from the specific domain for which we want to develop the ontology. Then he applies language processing techniques [13] to extract the candidate terms to "classes" or "categories" of the taxonomy. And finally, using classification algorithms he generates a tree or graph that represents the relationships between the most significant terms of the domain [14], and that constitute the taxonomy of the ontology (see Figure 1).

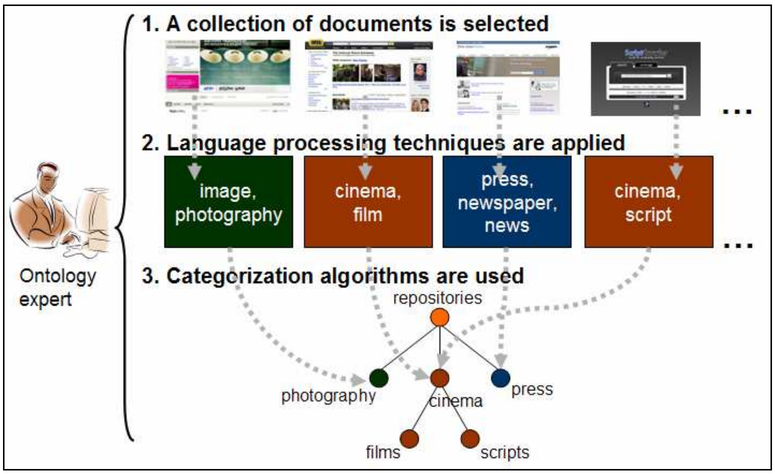

**Fig. 1**. Outline of the taxonomy extraction process

Unfortunately, the implementation of this process usually concludes with taxonomies composed by inappropriate "classes" or "categories", in many cases by their high degree of specificity, which usually also involves the generation of an excessive number of them.

## 3 Objectives

As already mentioned, Ontology learning is a process, and as such, its quality is determined by the quality of the worst of its stages. Thus, its success depends, among other things, on: a) having a corpus of documents from the domain to which we want to develop the ontology; b) preprocessing properly the documents to extract the most suitable terms to be used as classes or categories of the resulting taxonomy.

This paper emphasizes the latter aspect with two aims: first, to point out some inconsistencies that may be occurring in the preprocessing of text collections; second, demand a greater attention to this stage in the text mining field, and particularly in the Ontology learning process.

With these objectives, this research analyzes and evaluates some of the preprocessing techniques most commonly used for the extraction of the classes of a taxonomy. These techniques come mainly from the Information Retrieval field, eg. the statistical measure tf-idf, *Term Frequency - Inverse Document Frequency,* whose use is widespread in Textual Data Mining.

Finally, some new preprocessing techniques are presented, which could help to envision and propose new approaches to obtaining better terms candidate to classes or categories of the taxonomies.

## 4 Methodology

### 4.1 Tools

To analyze and assess the adequacy of the preprocessing techniques normally used in Ontology learning (as well as in Text mining) the following tools and resources have been used:

a. A corpus of documents from a specific domain.
b. Language processing programs developed ad hoc for this experimentation.
c. Software of linguistic analysis.
d. A lexical resource to control the semantic relationships existing between terms (more specifically, the relationship of hyperonymy).

#### 4.1.1 Corpus of documents

The experiments were performed on the Reuters-21578 collection (Distribution 1.0). This collection consists of 21,578 documents that appeared as economy news during 1987 in the Reuters database. In these experiments “The Modified Apte Split” has been used, which includes 12,902 documents, all indexed by human experts of Reuters Ltd. according to a set of 135 economics subjects.

The choice of this database for this research is motivated by its popularity, accessibility and labelling, as well as being one of the largest collections used in the clustering experiments.

### 4.1.2 Language processing programs

Tools made ad hoc to facilitate the statistical processing of documents (experiment 1). Their objectives are: a) Removing text labels (or stripping). b) Standardization of texts (control of lower and uppercase letters, stop words, punctuation marks, acronyms, dates and numeric quantities). c) The application of a stemming algorithm (specifically Morphy [15], the algorithm used by WordNet).

### 4.1.3 Software of linguistic analysis

A tool that enables the linguistic processing of documents has been also used (experiments 2, 3, 4, and 5). This tool is Freeling (version 3.0), a software developed by the “Center de Tecnologies i Aplicacions of Llenguatge i la Parla” (TALP) (http://www.talp.cat/) from the Polytechnic University of Catalunya (UPC).

Some of its main functionalities are: tokenization, morphological analysis, treatment of suffixes, identification of n-grams (Multiword); recognition of dates, numbers and currencies; annotations based on WordNet sense, etc.

### 4.1.4 Lexical resource

The lexical resource used is **WordNet** (WordNet 3.0, http://wordnet.princeton.edu/) [16]. It is a data base of lexical references. It groups words (nouns, verbs, adjectives and adverbs) into sets of synonyms called 'synsets', each representing a lexical concept. The senses are synsets associated with the different entries (words) of WordNet. Also, different semantic relations (meronymy, hyperonymy, hyponymy, etc.) relate the sets of synonyms.

## 4.2 Experiments

In this section five experiments have been defined, each one corresponding to a different preprocessing technique, and all of them defined to contrast, compare and evaluate the terms obtained as candidates for classes with them.

The characteristics of these five experiments can be seen in table 1, which specifies:

a. **Terminology Extraction Method:** i.e., the type of preprocessing that has been applied to documents to extract the most important terms.

b. **Vocabulary:** it is the number of terms extracted to represent the content of the collection.

c. **Term Weighting:** it is the statistical measure to determine the importance of terms [17] that describe the contents of each document in the collection. The assignment of weights is made using one of the following techniques:

   i. **Term frecuency**: this technique assigns to each term a value equal to its frequency of repetition within the document.

   ii. **Tf-idf**: it is the measure most commonly used in the Information retrieval and Text mining fields. This measure assigns to terms the value obtained after dividing the frequency of a term in the document between the frequency of the same term throughout the entire collection.

   iii. **Tf-Mod1**: This measure has been proposed by the authors. Here, each term obtains the value from the sum of its frequency in the document and the frequency of the same term throughout the collection.

   iv. **Tf-Mod2**: this measure has been also proposed by the authors, and supposes a slight variation from the previous one. In this case, the weight of each term is obtained by dividing the value obtained in Tf-Mod1 between the quantity of terms contained in the document.

d. **Relevance of terms in the collection**: it is the method used to propose the most representative terms of the collection as a whole. Basically, this measure helps to generate, for each experiment, a ranking that sorts the terms of the vocabulary according to their importance (or significance) for the collection. To do that, two measures have been used:

   i. **Mutual information (MI)**: is a measure widely used in the Information Theory and the Probability Theory fields, which estimates, from the weights of the terms, which are the most significant terms to represent the contents of the collection [18].

   ii. **Overall frequency of the terms (OF)**: we have applied this measure in the experiments that use the term weighting techniques proposed by the authors. Here the terms are arranged in a ranking based on their overall frequency in the collection.

**Table 1**. Characteristics of the preprocessing techniques.

| | ***Exp. 1*** | ***Exp. 2*** | ***Exp. 3*** | ***Exp. 4*** | ***Exp. 5*** |
|---|---|---|---|---|---|
| ***Terms Extraction Method*** | Statistical processing of documents: stripping, control of lower and uppercase letters, stop words, punctuation marks, acronyms, dates and numeric quantities, application of stemming algorithms | Software of linguistic analysis “Freeling” | Software of linguistic analysis “Freeling” | Software of linguistic analysis “Freeling” | Software of linguistic analysis “Freeling” |
| ***Vocabulary*** | 10.877 | 3.787 | 3.787 | 3.787 | 3.787 |
| ***Term Weighting*** | Term frecuency | Term frecuency | Tf-idf | Tf-Mod1 | Tf-Mod2 |
| ***Relevance of terms*** | MI | MI | MI | OF | OF |

## 5 Results

Once the terms candidate to classes have been extracted, it is necessary to begin the analysis of results based on the following criteria:

1. Similarity among terms proposed by each experiment.
2. Coverage of keywords assigned by human expert indexers.
3. Semantics of the terms proposed.

### 5.1 Similarity of the terms proposed by the experiments.

The first analysis involves the comparison of the terms proposed by each experiment, to determine their degree of similarity. To this end, the extracted terms are sorted according to their degree of relevance and are grouped into sets, so as to obtain the 10, 50, 100, 200, 300, 400 and 500 most significant terms of each experiment.

Then, these sets are compared with each other, obtaining the percentages of similarity between the different experiments (Table 2).

When analysing the data of this table we must bear in mind that when we increase the number of terms compared, the similarities also increase, especially among the experiments 2, 3, 4 and 5, since all work on the same set of terms, but each of them sort these terms differently. Therefore, if all terms are compared, the level of coincidence of these four experiments would necessarily be 100%.

Note also that the data from those experiments that share a high similarity with the analyzed experiment have been marked in bold, and in italics those data (specially Exp. 3) which are characterized by low similarity.

**Table 2**. Similarity of the terms proposed by the experiments.

| | | ***T10*** | ***T50*** | ***T100*** | ***T200*** | ***T300*** | ***T400*** | ***T500*** |
|---|---|---|---|---|---|---|---|---|
| **Exp. 1** vs. | **Exp. 2** | 30% | 32% | 34% | 34,50% | 42% | 43% | 42% |
| | ***Exp. 3*** | *0* | *6%* | *10%* | *15,50%* | *19,30%* | *26,25%* | *28,20%* |
| | **Exp. 4** | 40% | 44% | 47% | 51,50% | 50,30% | 50% | 48,40% |
| | **Exp. 5** | 30% | 48% | 47% | 49% | 49,60% | 49,50% | 47,80% |
| **Exp. 2** vs. | **Exp. 3** | 30% | **60%** | **61%** | 65,50% | 66,60% | 75,50% | 77% |
| | **Exp. 4** | 40% | 42% | 48% | 58,50% | 65,30% | 67,50% | 70,20% |
| | **Exp. 5** | 20% | 40% | 46% | 57,50% | 62,60% | 67,25% | 69,60% |
| **Exp 3** vs. | ***Exp. 4*** | *0* | *8%* | *15%* | *27%* | *33,60%* | *42,75%* | *48%* |
| | ***Exp. 5*** | *0* | *8%* | *15%* | *26,50%* | *32,60%* | *44,25%* | *48%* |
| **Exp. 4** vs. | **Exp. 5** | **60%** | **72%** | **84%** | **87,50%** | **87%** | **87,75%** | **89,80%** |

In table 2 we can see that the Exp. 1 has a low percentage of matching up with the Exp. 3, but presents a relatively high percentage of similarity with the rest of experiments.

Exp. 2 also has a high degree of coincidence with other experiments, being quite significant its similarity with Exp. 3, especially from fifties terms (T50).

Exp. 3 shows a high coincidence with Exp. 2 (at T50 they already share the 60% of their terms) and low matching with other experiments.

Exp. 4 has a high percentage of similarity with experiments 1, 2 and 5, and particularly with the latter, with which shares 60% of its terms from tenth terms (T10). However, it is also significant its low rate of coincidence with Exp. 3.

Exp. 5 has the same behaviour than Exp. 4, with a remarkable percentage of similarity with Exp. 4.

The analysis of these results reveals that there are experiments with a high similarity, such as the Exp. 2 and Exp. 3, which from now on will be called "Group 1", and Exp. 4 and Exp. 5, and that from here on in will be called "Group 2".

### 5.2 Coverage of keywords assigned by human experts.

The corpus of documents used in these analyses was described by human experts from Reuters Ltd., using 135 keywords from the economic field. In this work is interesting to evaluate the degree of coverage that these five experiments make of these keywords. The fewer words an experiment needs to cover keywords used by human experts, the better its coverage. The table 3 shows the percentages of coverage of these experiments.

**Table 3**. Coverage of the keywords assigned by human experts.

| | | ***Group 1*** | | ***Group 2*** | |
|---|---|---|---|---|---|
| ***Coverage*** | ***Exp. 1*** | ***Exp. 2*** | ***Exp. 3*** | ***Exp. 4*** | ***Exp. 5*** |
| ***10*** | 0 | 0 | 0 | 0 | 0 |
| ***50*** | 0,74% | 3,70% | 3,70% | 0,74% | 1,48% |
| ***100*** | 1,48% | 8,14% | 8,14% | 2,22% | 2,96% |
| ***200*** | 8,88% | 22,96% | 25,92% | 12,59% | 14,07% |
| ***300*** | 14,07% | 26,66% | 31,85% | 16,29% | 17,77% |
| ***400*** | 16,29% | 31,11% | 34,81% | 19,25% | 22,96% |
| ***500*** | 17,77% | 34,07% | 37,03% | 24,44% | 27,40% |
| ***600*** | 18,51% | 36,29% | 37,03% | 26,66% | 28,88% |
| ***700*** | 20% | 37,03% | 37,77% | 30,37% | 32,59% |
| ***800*** | 21,48% | 39,25% | 38,51% | 32,59% | 34,07% |
| ***900*** | 24,44% | 41,48% | 42,96% | 32,59% | 34,07% |
| ***1000*** | 26,66% | 42,22% | 44,44% | 34,07% | 35,55% |
| ***2000*** | 38,51% | 52,59% | 52,59% | 51,11% | 51,11% |
| ***3787*** | 50,37% | 55,55% | 55,55% | 55,55% | 55,55% |

Table 3 shows that none of these experiments extract all keywords proposed by human indexers, being the maximum coverage of them of 55%. This percentage is obtained only by experiments that extracted their terms using the software of linguistic analysis (Exp. 2, 3, 4, and 5). Furthermore, experiments in Group 1 achieved greater coverage of these keywords, being the Exp. 3 the one that provides the best coverage with a lower number of terms.

## 5.3 Semantics of the terms proposed.

Finally, it is analyzed the semantic value of the terms proposed by each experiment. Table 4 shows the top ten terms each experiment proposed as more representative of the collection:

**Table 4**. The ten most relevant terms proposed by experiments.

| | | *Group 1* | | *Group 2* | |
|---|---|---|---|---|---|
| ***Terms*** | ***Exp. 1*** | ***Exp. 2*** | ***Exp. 3*** | ***Exp. 4*** | ***Exp. 5*** |
| ***1*** | say | nil | cattle | pct | pct |
| ***2*** | dollar | pct | cooperative | year | year |
| ***3*** | pct | rate | nil | share | share |
| ***4*** | year | bank | buffer | company | company |
| ***5*** | bank | bond | cocoa | loss | rate |
| ***6*** | ct | stock | beef | bank | profit |
| ***7*** | billion | cattle | soybean | price | loss |
| ***8*** | share | trade | cotton | market | sale |
| ***9*** | company | buffer | acre | rate | country |
| ***10*** | US | dollar | farm | stock | month |

A human expert analysis of these data conclude that Group 1, and especially Exp. 3, is characterized by its high level of specificity. This is because Exp. 3 extracts its terminology using the technique tf.idf, which is designed to identify those terms that best discriminate a document from others. This technique, combined with the calculation of mutual information, makes its terms to be very specific.

In contrast, Group 2 recommends sets of terms that are apparently more general, and, a priori, closer to classes of a taxonomy. However, these observations are made by human experts (authors, in this case), and could be subjective. It would be interesting to be able to evaluate automatically the degree of specificity of these sets of terms.

WordNet has been used with this aim, and we have searched the number of hypernyms that this linguistic resource associates to the first 20 terms proposed by Exp. 3 and Exp. 5. As a result, 92 hypernyms has been obtained for Exp. 3, and 71 hypernyms for Exp. 5. This finding confirms that the terms proposed by Exp. 5 are more general and, in consequence, probably closer to classes or categories, that was the original goal of this experiment.

## 6 Conclusions

From the results observed in the analyses performed we conclude that:

a. The experiments that used an application of linguistic analysis, and specifically through the selection of nouns that appear in the texts, obtained a greater coverage of keywords proposed by human experts. Also, they extracted a lower number of terms and, in consequence, entailed a lower computational cost. Besides, terms proposed for these experiments showed semantics closer to that required by the classes of a taxonomy.
b. As a consequence of what is mentioned in the previous section, we remove from this analysis the Exp. 1. The rest of experiments can be divided into two groups according to their similarity. In the case of Group 1, this similarity may be due to the use of techniques from the Information retrieval field, which let us to extract a terminology characterized by its level of specificity. In contrast, in Group 2 the similarity is motivated by the importance given to the cumulative frequency of the terms in the collection, which the authors propose to identify terms closer to classes or categories.
c. In each one of these two groups there is an experiment that stands out for its high coverage of the keywords proposed by human indexers. In the case of Group 1 the Exp. 3 has the greatest coverage, whereas in the case of Group 2 the experiment with more coverage is Exp. 5. Furthermore, since the level of coincidence between the terms of these experiments is very low, it could be interesting to join them in a new approach, that would increase the percentage of keywords covered with a smaller number of terms.
d. Finally, the results obtained in this work show that if we want to identify terms closer to classes or categories, it is better to use preprocessing techniques such as those proposed in Exp. 4 and Exp. 5, better than approaches from the Information Retrieval field.

## 7 Acknowledgments

The development of this research is partially supported by projects CSO 2008-02627 and CSO2009-13713-C05-04 of the Spanish Ministry of Science and Innovation.

## 8 References

1. Berners-Lee, T., Hendler, J., and Lassila, O. *The Semantic Web*. Scientific American, vol. 284, nº 5, pp. 34-43 (2001)
2. Codina, L., Marcos, M.C., and Pedraza-Jimenez, R. (coords.) *Web semántica y sistemas de información documental*. Trea (2009)
3. RDF Working Group. Resource *Description Framework (RDF)*, (2004) http://www.w3.org/rdf

4. Pedraza-Jimenez, R., Codina, L., and Rovira, C. *Web semántica y ontologías en el procesamiento de la información documental*. El Profesional de la Información, vol.16, nº 6, pp. 569-578, (2007)
5. OWL Working Group. *OWL Web Ontology Language*, (2004) http://www.w3.org/2004/owl
6. Codina, L., and Rovira, C. *La Web semántica*. Tramullas, Jesús, Eds. Tendencias en documentación digital, chapter 1. Trea, (2006)
7. Pedraza-Jimenez, R., Codina, L., and Rovira, C. *Semantic web adoption: online tools for web evaluation and metadata extraction*. The 8th International FLINS Conference on Computational Intelligence in Decision and Control, Madrid (2008)
8. Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E., and Stephens, S. *The Semantic Web in Action*. Scientific American, vol. 297, nº 6, pp. 90-97 (2007)
9. Maedche, A. and Staab, S. Ontology learning for the Semantic Web. Intelligent Systems, IEEE, volume 16, issue 2, pp. 72-79 (2001)
10. Maedche, A., and Staab, S. *Ontology learning*. S. Staab and R. Studer, editors, Handbook on Ontologies, pp. 173-189. Springer (2003)
11. Buitelaar, P., Cimiando, P., Magnini, B. Ontology Learning from Text: Methods, Evaluation and Applications, Frontiers in Artificial Intelligence and Applications Series, vol. 123, IOS Press, (2005)
12. Gómez-Pérez, A., Manzano-Macho, D. *An overview of methods and tools for ontology learning from texts*, The Knowledge Engineering Review, vol. 9, nº. 3, pp. 187-212 (2005)
13. Vallez, M. and Pedraza-Jimenez, R. *Natural Language Processing in Textual Information Retrieval and Related Topics*. "Hipertext.net", num. 5 (2007) http://www.hipertext.net/english/pag1025.htm
14. Hotho, A., Staab, S., and Stumme, G. *Explaining text clustering results using semantic structures*. In Principles of Data Mining and Knowledge Discovery, 7th European Conference, PKDD (2003)
15. Beckwith, R., Miller, G. A. and Tengi, R. *Design and Implementation of the WordNet Lexical Database and Searching Software. Description of WordNet.* Technical report (1993)
16. Miller, G. *WordNet: A lexical database for english*. Communications of the ACM, vol. 38, nº 11 (1995)
17. Chisholm E. and Kolda, T. *New term weighting formulas for the vector space method in information retrieval*. Technical report, Oak Ridge National Laboratory (1999)
18. Church, K. W. and Hanks, P. *Word association norms, mutual information, and lexicography*. Computational Linguistics, vol. 16, nº 1, pp. 22-29 (1990)