International Journal of the Physical Sciences Vol. 6(7), pp. 1755-1765, 4 April, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.115
ISSN 1992 - 1950 ©2011 Academic Journals

*Full Length Research Paper*

# A semantic role-based methodology for knowledge acquisition from Spanish documents

**José Luis Ochoa, Maria Luisa Hernández-Alcaraz, Rafael Valencia-García* and Rodrigo Martínez-Béjar**

Facultad de Informática, Universidad de Murcia 30071 Espinardo (Murcia) Spain.

Accepted 14 March 2011

**The Semantic Web vision grounds on providing the current Web with structural knowledge that can be understandable by the machines without the intervention of human beings. Given that ontologies are a backbone technology for the Semantic Web, different mechanisms and methodologies for designing and building ontologies have been proposed. The need for overcoming the bottleneck provoked by the manual construction of ontologies has generated several studies and research on obtaining semiautomatic methods to build ontologies. Ontology learning from Web documents is considered to be an important activity to promote the Semantic Web. In this paper, an automatic method for acquiring knowledge from Spanish texts is described. The method is based on semantic roles, which have been employed in our research for extracting semantic relations between concepts. The method makes it possible to represent multiple semantic relations. A set of experiments have been performed with the approach implemented in the oncology domain that show promising results.**

**Key words:** Ontology learning, semantic role labeling, information extraction, semantic web, ontology.

## INTRODUCTION

Due to its size and the diversity of its textual information, the World Wide Web has become a precious resource for the acquisition of lexical information and for the compilation of corpora. Web sites contain information originally designed to be human-readable, so that a manual process is required for making that information machine-readable. This process can be tedious, difficult, and time-consuming (Han and Elmasri, 2004). In order to face this problem, several approaches have been proposed for different purposes such as the extraction of parallel corpora, lexical information (Santamaría et al., 2003).

In Berners-Lee et al. (2001) the Semantic Web was defined as an extension of the current Web in which information is provided with well-defined meaning, so allowing computers and people to work in a cooperative manner. In the Semantic Web, ontologies are used as a knowledge representation technology.

An ontology is viewed in this work as a formal specification of a domain knowledge conceptualization (van Heijst et al., 1997). In this sense, ontologies provide a formal, structured knowledge representation, and are reusable and shareable. In our methodology, ontologies are used to represent the knowledge extracted from texts.

Ontologies have been applied to a number of different domains, including biomedicine (García-Sánchez et al., 2008), finance (Valencia-Garcia et al., 2011), tourism (Ruiz-Martínez et al., 2009), education (Fernández-Breis et al., 2009; Hashim et al., 2010) and software engineering (Beydoun et al., 2009 a, b; Henderson-Sellers, 2011).

Due to the outstanding of the importance of ontologies, different methodologies for designing and building ontologies have been proposed. In this respect, it can be said that the manual ontology construction process constitutes a major problem, since it involves a time-resources consuming task (Fortuna et al., 2006). Hence, the generation and development of methods and software tools to support the construction of ontologies is a relevant research area, which is known as Ontology Learning. One of the most active subareas in Ontology Learning is the use of natural language texts to build

*Corresponding author. E-mail: valencia@um.es. Tel: +34 868888522. Fax: +34 868884151.

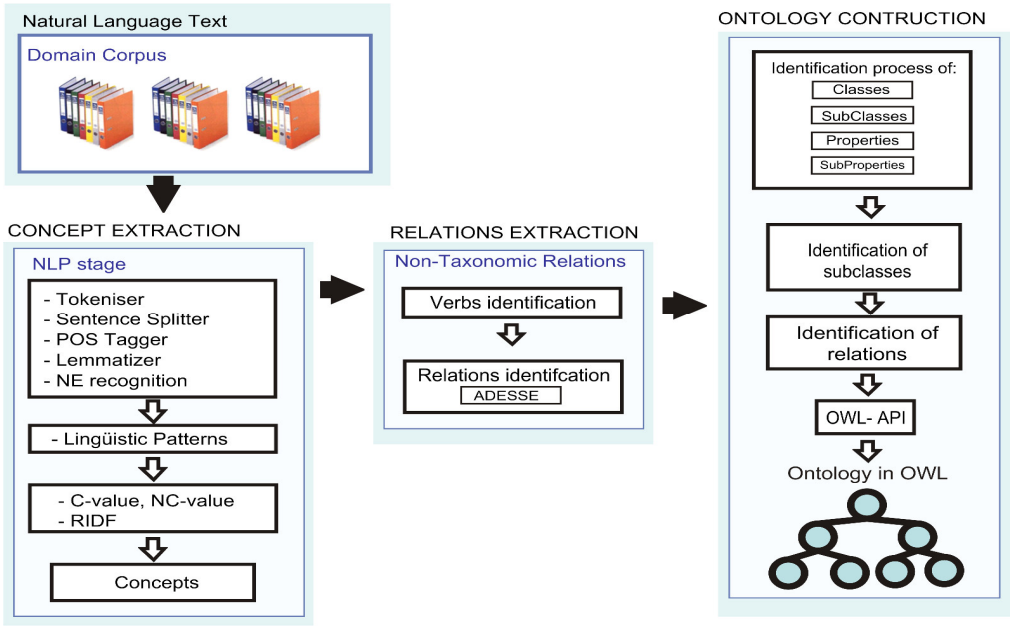


**Figure 1.** Ontology learning process.

ontologies.

As it could be expected, the vast majority of ontology learning methods have focused on the English language. In comparison with English language, Spanish has a much more complex syntax, and is nowadays the second most spoken language in the world[1]. These facts have led us to claim that the computerization of Internet domains in Spanish is of utmost importance.

In this paper, we propose a method for ontology learning from Spanish natural language texts based on the identification of semantic relations among concepts by using semantic roles.

## ONTOLOGY LEARNING PROCESS

The approach for ontology learning consists of three sequential subprocesses, namely: (1) Concept extraction, (2) relations extraction (3) and ontology construction (Figure 1). These subprocesses are applied to each text in the corpus, with the subsequent extraction of the knowledge entities (concepts and relationships) contained in them.

### Concept extraction process

Through this process, terms representing concepts are identified. It is assumed that there exist both multiword and single word terms. By taking into account this assumption, two different methods have been implemented: the NC-Value algorithm (Ochoa et al., 2010), which allows to obtain the multiword terms candidates to represent concept, and RIDF (Manning and Schütze, 1999), which has been employed to obtain terms formed by one word. This process can be decomposed into several phases as described next.

#### *NLP stage*

The main aim of this stage is the extraction of the morphosyntactic structure of each sentence. For this purpose, a set of NLP software tools including a sentence detection component, a tokenizer, a set of POS taggers, a set of lemmatizers and a set of syntactic parsers have been developed. For it, Freeling[2]2.2 has been employed. Spanish language has a very complex inflection system compared to the one presented in English language, specially in verb conjugation. In Spanish determiners, nouns and adjectives have gender and number. Moreover, both lexical categories determiners and adjectives have to agree in gender and number with their corresponding nouns. For example, the all the terms “desempleado”, “desempleada”, "desempleados” and “desempleadas” refer to the same concept, namely unemployed. For these reasons, it is important to note that the lemmatizer used must have a good accuracy, and the concept extraction process must use the lemmas

[1] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

[2] http://nlp.lsi.upc.edu/freeling/

**Table 1.** Linguistic patterns.

| Linguistic pattern | Term |
|---|---|
| NC + SP+ NC | *Tipo de interés* (Interest rate) |
| NC + AQ | *Cuadro macroeconómico* (Macroeconomic profile) |
| NC + SP + DA + NC | *Precio de el dinero* (Price of money) |
| NC + SP + AQ + NC | *Fondo de alto riesgo* (subprime funds) |
| NC+ SP+NC | *Beneficio antes de impuestos* (Pre-tax profits) |
| NC+ AQ+ AQ | *Crecimiento anual acumulativo* (Annual cumulative growths) |
| NC + AQ + SP +DA +NP | *Cuota empresarial a la Seguridad Social* (Employer's contribution to national insurance) |

and not the words of the text.

### *Linguistic patterns*

The candidate terms are identified by means of a hybrid method which uses a series of linguistic patterns in which the morphosyntactic structure of the terms is described. These patterns depend on the domain, and for their design it is possible to use a predetermined list of terms which can be obtained from several information sources such as previous studies, pre-defined ontologies, or terminological databases, for example WordNet (Miller, 1990) and EuroWordNet (Vossen, 1998). For the validation of this tool, the patterns have been defined manually from the corpus.

Table 1 shows some of the morphosyntactic patterns obtained for the financial domain as well as some terms matching the patterns. These patterns are language-dependent, so they must be defined for each language. For example, adjectives in English usually go before the noun while in Spanish they go after it. Besides, the saxon genitive does not exist in Spanish language, so it is usually represented in this language using the preposition "de" (of).

### *Multiword concept extraction stage*

Once a list of multiword candidate terms has been obtained, this list is filtered out by applying the NC-value algorithm. For that, the system arranges the terms list according to the amount of words contained in each term and calculates the values for several parameters, namely: the occurrence frequency of the candidate term within longer candidates, the occurrence frequency of the candidate term, the length of the candidate term and the total occurrence frequency of the candidate term in the corpus.

In order to obtain an acceptable precision level in the candidate term list, the NC-value method (Ochoa et al., 2010; Barrón-Cedeño et al., 2009) uses the morphological information from the context of the term under question. For this, we consider that verbs, adjectives and nouns are likely to be found in the neighbourhood of a term as it has been proposed elsewhere (Grefenstette, 1994).

The system processes context words and split them up according to their grammatical category (that is, Adjectives, Verbs or Nouns). With the method developed by Grefenstette (1994), a type of weight known as 'context weighting factor' is obtained. It calculates the probability of a context word appearing with a certain term. Next, both the C-Value and NC-Value algorithms are explained in detail.

### *C-value*

Its formula is provided in Equation 1. First, the system arranges the term list according to the amount of words contained in each term and calculates the frequency of occurrence of the candidate term within longer candidates, the frequency of occurrence of that longer candidate term, the length of the candidate term and the total frequency of occurrence of the candidate term in the corpus.

$$C-Value=\begin{cases}\log_2|a|*f(a) \\ \log_2|a|*\left(f(a)-\dfrac{1}{P(T_a)}\sum_{b\in T_a}f(b)\right)\end{cases} \quad (1)$$

Where:
*a:* is the candidate string,
$|a|$: is the length of the candidate string,
$f(a)$: is its frequency of occurrence in the corpus,
$T_a$: is the set of extracted candidate terms containing *a*,
$P(T_a)$: is the number of the longest candidate terms containing *a,*
$Sf(b)$: is the frequency of occurrence of *a* as a sub-term of any candidate term *b* as $|a| < |b|$.

Table 2 shows some of the C-value scores obtained for the financial domain.

**Table 2.** Example of C-Value scores.

| C-Value | Term |
|---|---|
| 9.51 | *Contrato de inserción* (Insertion contract) |
| 6.34 | *Cuarto de punto* (Quarter-point) |
| 1.5 | *Consejo de administración* (Board of directors) |

### *NC-value*

In order to obtain an acceptable degree of precision in the candidate term list, the NC-value method uses the morphological information from the context of the term under question. As Grefenstette (1994) has pointed out, all verbs, adjectives and nouns are likely to be found in the neighbourhood of a term.

The system processes context words and split them up according to their grammatical category of Adjectives, Verbs or Nouns. With the method developed by (Grefenstette, 1994), a type of weight known as 'context weighting factor' is obtained. It calculates the probability of a context word appearing with a term, as it is formalised in Equation 2.

$$weight(w) = \frac{t(w)}{n} \quad (2)$$

where:
w is the context word,
t(w) is the number of times that the context word appear with the term,
*n* is the total frequency of occurrence,
*weight (w)* is the context weighting factor.
Once the weighting factor has been obtained for each context word of the candidate term, the scores are standardized by means of Equation 3.

$$cw_1 \cdot w_1 + cw_2 \cdot w_2 + \ldots + cw_n \cdot w_n \quad (3)$$

where:
$cw_{x:}$ is the number of times that the context word *x* appears with the term,
$w_x$: is the weighting factor obtained for the context word *x*.
With the global score obtained, the C-value list may be rearranged. For this purpose, a further calculation is required:

$$NC-value(a) = 0.8C-value(a) + 0.2\sum_{b \in C_a} f_a(b) weight(b) \quad (4)$$

where:
*a:* is the candidate term,
$C_{a:}$ is the set of context words for *a*,
*b:* is a word from $C_a$,
$f_a$ (*b*): is the frequency of *b* as the context word for *a*,
weight (b): is the weight of *b* as the context word.

In Equation 4, the factors 0.8 in the first part and 0.2 in the second part have been assigned after several tests conducted by Grefenstette (1994). These values represent the best distribution in the precision of the extracted terms. If the C-value of the candidate term is multiplied by 0.8 and the result of the summation of the individual weight corresponding to the context words is multiplied by 0.2, we get the NC-value, which rearranges the list by placing the best candidate terms on the top of the list.

### *Single word concept extraction stage*

Residual IDF (RIDF) is defined as the difference between the logarithms of the actual document frequency and the document frequency predicted by a Poisson distribution (Manning and Schütze, 1999) (Equation 5).

$$RIDF\ (i) = Idf\ (i) + \log_2 \frac{1}{(1 - p(0; \lambda_i))}) \quad (5)$$

Where: *p* is the Poisson distribution with parameter $\lambda_i = \frac{cf_i}{N}$ (the average number of occurrences of each word per document). $1 - p(0; \lambda_i)$ is the Poisson probability of a document with at least one occurrence of *i*.

### Relation extraction process

Once the concepts have been identified in the corpus, the semantic relations of these concepts have to be obtained. In natural language, relations between concepts are usually associated with verbs (Valencia-García et al., 2008). A number of systems for learning relationships have been proposed that are based on the extraction and identification of verbs (Shamsfard and Barforoush, 2004; Sánchez and Moreno, 2008; Valencia-García et al., 2008). In this work semantic roles and semantic class membership for Spanish verbs are used in order to extract and identify these relationships (Table 3).

A semantic role is the relation between a syntactic constituent and a predicate. It defines the role of a verbal argument in the event expressed by the verb (Moreda et al., 2010). The semantic roles set developed in the Proposition Bank (PropBank) project (Palmer et al., 2005) and in the FrameNet project (Filmore, 2002) are the most widely used, although they are only useful for English. ADESSE (Vaamonde et al., 2010) collects nearly 4,300 semantic roles of Spanish verbs in a syntactic database of nearly 160,000 clauses retrieved from a Spanish corpus of 1.5 million words.

Spanish verb conjugation is very complex and highly irregular in some cases. For example, the Spanish verb

```
<predicate lemma="pertenecer">
<roleset id="posesión">
<roles>
  <role n="0" descr="posession"/>
  <role n="1" descr="holder"/>
</roles>
</roleset>
```

el [$_0$Banco Popular] pertenece al [$_1$IBEX 35]

[$_0$Banco Popular] belongs to [$_1$IBEX-35]

**Figure 2.** An example of the pertenecer frame in ADESSE.

**Table 3.** Examples of relations extracted.

| Sentence | Semantic relation |
|---|---|
| El parón del consumo provocará la subida de el IVA<br>(The halt in consumption will bring about the VAT increase) | parón_del_consumo provoca subida_del_IVA<br>(halt_in_consumption brings about VAT_increase) |
| El grupo Sacyr Vallehermoso registró en 2009 un beneficio neto de 505,9 millones de euros<br>(Sacyr Vallehermoso group reported a net profit of €505.9 million) | Sacyr_Vallehermoso registrar beneficio_neto<br>(Sacyr_Vallehermoso reports net_profit) |
| El objetivo de la política monetaria es garantizar la estabilidad de los precios, lo que significa…<br>(The monetary policy is aimed at ensuring price stability, which means...) | Política_monetaria garantiza estabilidad_de_los_precios<br>(Monetary_policy ensures price_stability) |
| El Banco Popular pertenece al IBEX-35<br>(Banco Popular belongs to IBEX-35) | Banco_Popular pertenece IBEX-35<br>(Banco_Popular belongs_to IBEX-35) |

**Table 4.** Subclasses assignment.

| Name of the class | Subclass |
|---|---|
| BANCO<br>(BANK) | BANCO_CENTRAL_EUROPEO BCE<br>(EUROPEAN_CENTRAL_BANK)<br>BANCO_SANTANDER<br>(BANCO_SANTANDER) |
| CUOTA<br>(SHARE) | CUOTA DE MERCADO<br>(MARKET_SHARE) |

system has 14 regular tenses which are subdivided into seven simple tenses and seven compound tenses. The compound tenses are formed by the auxiliary verb “haber” followed by the past participle. Spanish verbs are conjugated through three persons, each having a singular and a plural form. Finally, many of the most frequent verbs are irregular. The regular ones fall into one out of the three regular conjugations defined for Spanish. These regular conjugations are classified according to the two last symbols of their infinitive forms, namely: –“–ar”, “-er”, or “-ir”. All these (Spanish language) features have a dramatic influence on the functionality of the lemamatizer to be used in that it must take into account all the regular and irregular verbs.

The unfolding of this process is described next. The main verb of the current sentence is identified. Then, there is a search for the type of semantic relation associated with that verb in ADESSE. This search is

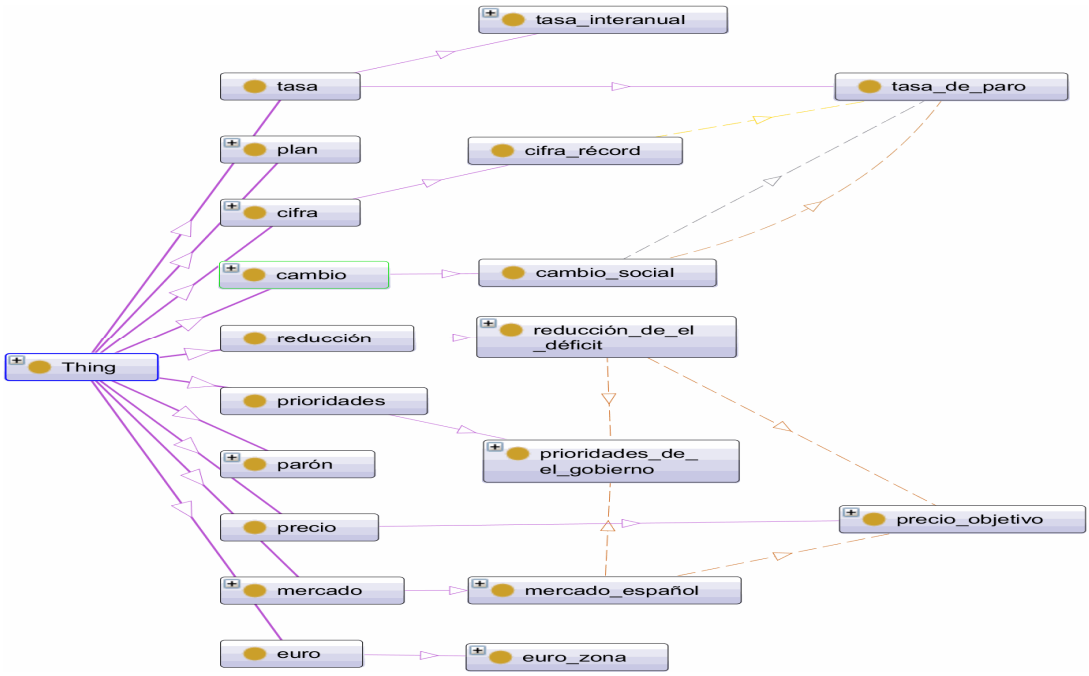


**Figure 3.** Final image of a part of an ontology created with this tool.

conducted on the ADESSE relational knowledge base by means of the lemmatized word of the verbal expression. Once the type of relation associated with the main verbal expression in the current sentence is found, the system selects those concepts which are related to that verb. For this purpose, the system looks for concepts on the right and left side of the verb.

In order to detect ontological semantic relations among entities, a mapping between semantic relations and semantic roles has to be done. For instance, Figure 2 shows the ADESSE *pertenecer (belong to)* frame. The example shows how the semantic role relates "Banco Popular" (a Spanish bank) and "Ibex 35" (the Spanish stock market index comprising the 35 most liquid Spanish stocks traded in the Madrid Stock Exchange General Index).

## Ontology construction process

At this stage, the ontology is built from the elements previously extracted. Specifically, the aim is the detection of the classes, subclasses and properties of the ontology. In an ontology, a property can be a datatype property or an object property. At this point, the system attempts to identify the subclasses of the concepts extracted at the first stage and then it inserts the detected relations at the second stage of the process.

### *Identification of subclasses*

The *subclass_of* relations are detected by means of the name of the class. In case a class's name is made up of other classes' names, then it would be a subclass of the first class. For instance, the BANCO_CENTRAL_EUROPEO (EUROPEAN_CENTRAL_BANK) is a subclass of the BANCO (BANK) concept, since this class's tokens are comprised by the first one. Some other subclasses instances can be observed in Table 4.

### *Identification of relations*

At this stage, concepts are related from the results obtained at the relation extraction stage. In order to identify the names of the properties, the lemmatized form of the verb is used.

OWL API[3] has been the library used for the development of the ontology in the OWL language. More concretely, OWL 2 (Grau et al., 2008), which is an extension and revision of OWL and has become the W3C recommendation for representing ontologies in the Semantic Web, has been used. OWL 2[4] addresses several problems and drawbacks that have been identified throughout the years of the extensive application of OWL in different contexts. Besides, OWL 2 adds several new features to OWL, including increased expressive power for properties, extended support for datatypes, simple metamodeling capabilities and extended annotation capabilities.

[3] http://owlapi.sourceforge.net/

[4] http://www.w3.org/TR/owl2-profiles/

**Table 5.** Evaluation results in the financial domain.

| | Concept extraction | | | Relation extraction (%) | Ontology construction | Total (%) |
|---|---|---|---|---|---|---|
| | Single word concepts (%) | Multi-word concepts (%) | Total (%) | | Subclass-of relations | |
| Precision | 60.87 | 83.69 | 69.54 | 79.42 | 55.90 | 69.81 |
| Recall | 73.68 | 87.41 | 79.38 | 91.47 | 83.33 | 84.01 |
| F1 | 66.67 | 85.51 | 74.14 | 85.02 | 66.91 | 76.25 |

An OWL ontology can be viewed from a logical point of view as a collection of axioms that must be satisfied. Figure 3 illustrates a part of the ontology obtained from a financial corpus.

## VALIDATION

The methodology described in this work has been validated across two application domains, namely: oncology and finance. The study of the results is based on two main scores: precision and recall. These measures are the most commonly used for the assessment of statistical extraction systems and trace their origins back to the information retrieval discipline (Subramaniam et al., 2010). The precision score (Equation 6) is the result of dividing the amount of knowledge entities suggested by the system and that are accepted by the expert user, into the total amount of knowledge entities suggested. The recall score (Equation 7) is the result of dividing the amount of the knowledge entities suggested by the system accepted by the expert into the total amount of knowledge entities existing in the fragment. That is:

$$precision = \frac{correct\ \ knowledge\ \ entities\ \ suggested}{total\ \ knowledge\ \ entities\ \ suggested} \quad (6)$$

$$recall = \frac{correct\ \ knowledge\ \ entities\ \ suggested}{total\ \ knowledge\ \ entities\ \ in\ \ the\ \ text} \quad (7)$$

In this work, the knowledge entities that have been studied are *concepts, relations and subclass_of*.

The F-measure score (Equation 8) has been also calculated. The F-measure score can be interpreted as a weighted average of the values corresponding to the two parameters precision and recall. F-measure scores range from 0 (that is, the worst case) to 1 (that is, the best case).

$$F-measure = 2 * precision * recall \Big/ precision + recall \quad (8)$$

All these measures have been calculated for the three processes which constitute the methodology proposed here, namely: concept extraction, relation extraction and ontology construction. In the concept extraction process, the results obtained by both the single word and multi word concept extraction sub-processes are presented. The results of the non-taxonomic relations identification are studied in the relation extraction process. Finally, in the last process, the results of the acquisition of taxonomic or subclass-of relations are shown.

### Validation in the financial domain

The amount of financial documents available on the Web such as news, reports, and papers is constantly increasing. In order to manipulate the mass of data contained in those documents, there is a great need for NLP tools which enable the automatic location, generation, organization and management of pieces of financial information. In the financial domain, as it is the case of any specialised area, words often acquire meanings which differ from those found in common language. Thus, a specialised vocabulary is needed for subsequent processing tasks.

Our experimental corpus has 37,396 words and comprises 82 documents. This corpus has been manually processed by domain experts, obtaining a total amount of 325 concepts or classes, 108 *subclass-of* relations, and 211 semantic relations. In Table 5, the results of the evaluation for each step of the process are shown.

#### *Concept extraction phase evaluation*

As shown in Table 5, the results obtained by the Multiword concept extraction sub-process are much better than the ones corresponding to the Single word extraction sub-process. More concretely, through single word extraction a low precision score (60.87%) is obtained, while precision achieved by using multi-word extraction is much higher (83.69%). That means that the system obtains a set of single word concepts that are not relevant in the domain, so decreasing the total precision measure (69.54%).

It is worth noting that the method obtains a total recall value of 79.38% in the detection of concepts, which means that the process does not identify 20% of the

**Table 6.** Evaluation results in the oncology domain.

| | Concept extraction | | | Relation extraction (%) | Ontology construction | |
|---|---|---|---|---|---|---|
| | Single word concepts (%) | Multi-word concepts (%) | Total (%) | | Subclass-of relations (%) | Total (%) |
| Precision | 62.50 | 86.67 | 74.19 | 84.75 | 76.11 | 78.20 |
| Recall | 83.33 | 92.07 | 88.06 | 74.63 | 78.06 | 80.59 |
| F1 | 71.43 | 89.29 | 80.53 | 79.37 | 77.07 | 79.38 |

the concepts extracted manually by the expert.

***Relation extraction phase evaluation***

As it has been explained before, through this phase the non-taxonomic relationships from the semantic roles of ADESSE are obtained. As it can be seen in Table 5, the best results have been obtained at this stage. Here, the system gets a precision and recall values of 79 and 91%, respectively.

***Ontology construction evaluation***

Through this phase, taxonomic relations are extracted and the ontology is built. Since the subclasses identification process is the only one that acquires knowledge in our methodology, the performance of that process has been evaluated. In particular, the worst results are obtained with a precision value of 56%.

***Overall evaluation***

The global results of the evaluation seem promising, although the precision of 68.91% is not very high due to the features of the financial domain, which is a wide domain where it is difficult to extract specialized concepts. On the other hand, the recall score is quite good and the method obtains a 84.01% of the corpus ontological elements.

**Validation in the oncology domain**

Vast amounts of medical knowledge reside within text documents, so that the automatic extraction of such knowledge would certainly be beneficial for clinical activities (Valencia-Garcia et al., 2004; Miranda-Mena et al., 2006).

The experimental corpus in the oncology domain has 115,257 words distributed into 83 documents. This corpus has been manually processed by domain experts, so obtaining a total amount of 653 concepts or classes *subclass-of* relations, and 670 semantic relations. As it is shown in Table 6, the results of the validation are promising and better than those obtained by the previous described experiment.

***Concept extraction phase evaluation***

At first sight, it can be observed that the Multiword concept extraction is better than the Single word concept extraction in this domain. The precision values obtained suggest that multiword medical terms are more specific than single word medical terms. On the other hand, the recall score states that almost all the multiword terms in the text have been identified by this process (92.07%). The global results of the concept extraction in this domain seem promising, as a precision value of 76.67% and a recall value of 88.06% have been obtained. These values allow us to affirm that the oncology domain is more specific that the financial one, since the performance in extracting specific concepts is quite good in this application domain.

***Relation extraction phase evaluation***

The precision improves in the detection of non-taxonomic relations in this domain (84.75%). On the other hand, the values of the recall are lower in this domain because ADESSE contains lots of general verbs and the domain under question is very specific in comparison with the finance domain.

***Ontology construction evaluation***

Although the subclass of relation extraction method obtains a better precision value in this domain (76.11%) in relation to the finance domain, the recall value is lower than its counterpart in the financial domain (78.06%). As it has been stated previously, in the oncology domain it is difficult to identify all the taxonomic relations with this approach.

***Overall evaluation***

The global results of the evaluation are solid. The total precision and recall scores, namely, 78.20 and 80.59%, suggest that the approach presented in this paper have a quite acceptable performance in oncology, that is, a

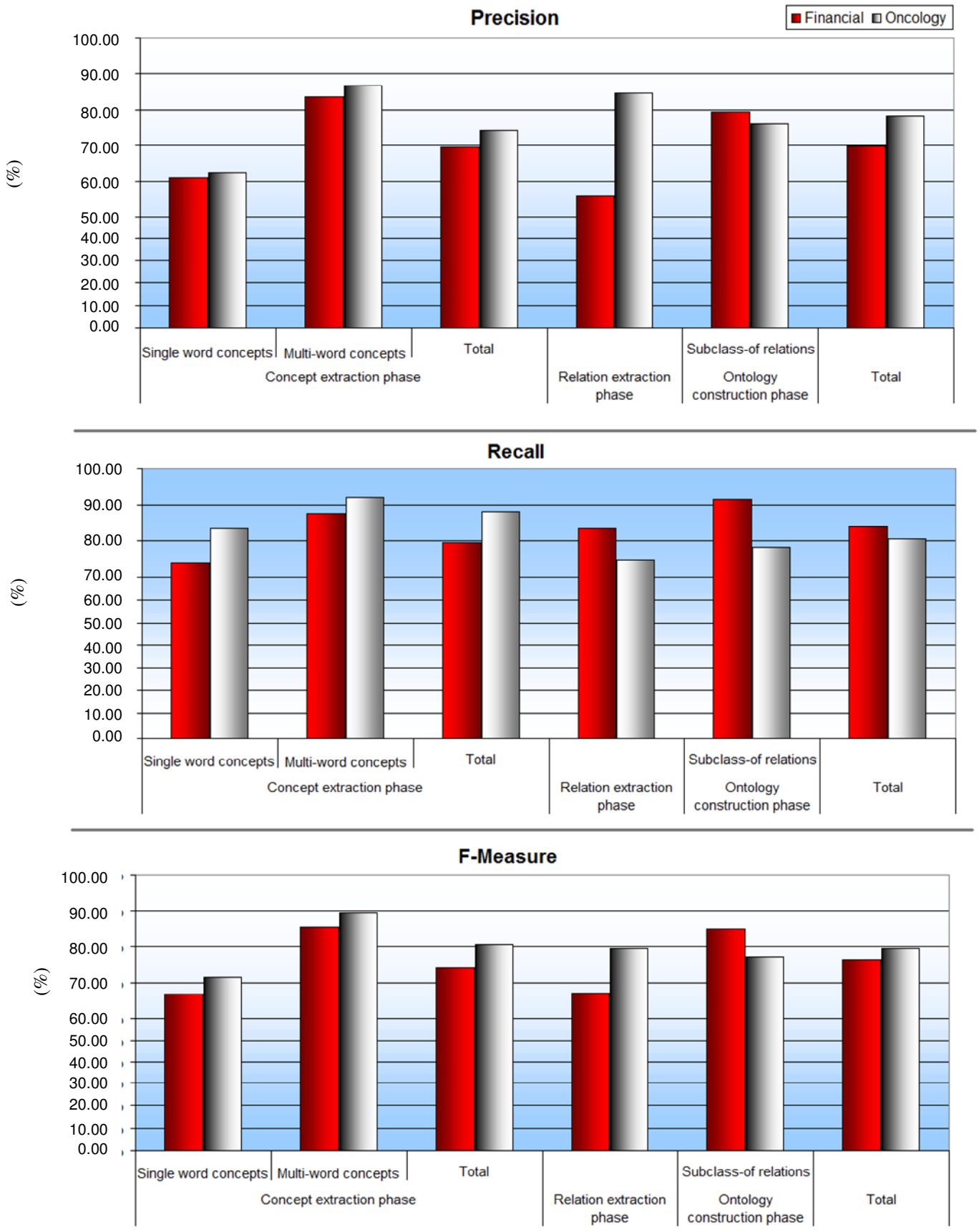


**Figure 4.** Evaluation results from all domains.

specific domain. As it has been argued before in this paper, the recall score is lower in this domain compared to what happens in the finance domain because ADESSE is mainly based on general verbs, while the oncology domain is very specific. Besides, there are a number of verbs that represent relations in this domain and which are not included in this Spanish syntactic database.

Graphically, the results of applying our approach to the two domains under question can be visualized in Figure 4.

## CONCLUSIONS AND FUTURE WORK

In this paper, an automatic method for acquiring knowledge from texts has been presented. This approach is based on the use of semantic roles from ADESSE in order to extract semantic relations between concepts.

Ontology building from free text documents is an important activity for the knowledge engineering community. One of the hottest research trends in this area is ontology learning from Web documents (Moreda et al., 2010; Sánchez and Moreno, 2008), which is considered to be an important activity to promote the Semantic Web (Shamsfard and Barforoush, 2004). The approach presented in this work is totally automatic. Another key feature of this approach is that it works not only with taxonomies, but also with multiple semantic relations.

Similarly, (Sánchez and Moreno, 2008) presents a methodology for the detection of non-taxonomic relations from Web texts. It is based on the identification of the relevant verbs in the text chunks. These verbs are then used as a knowledge basis for learning and tagging non-taxonomic relations automatically and without supervision. Both studies use linguistic patterns for obtaining taxonomic relations.

The authors in (Jiang and Tan, 2010) introduce a methodology which has a similar common ground to ours. The aim of these authors is the development of a high-quality ontology, and for this purpose they use a combination of statistical and lexical-semantic methods.
This method has been validated in two different application domains, namely finance and oncology. The results obtained in both domains have been analyzed and discussed. In this sense, the obtained results of the concept extraction phase show that such a phase is more effective in the oncology domain than in the finance domain. The reason for this must be found in the very nature of the oncology domain compared to that of the finance domain. To be more precise, oncology terms are more specific than finance-related ones, which often refer to more general concepts. It is also interesting to note that the number of concepts that the experts have identified in both domains differ. In particular, the oncology domain has more concepts to discover than the finance domain and the method allows to obtain better results if the number of concepts to be found is high.

On the other hand, in the relation extraction the results obtained by the finance domain are better because the verbs contained in ADESSE are mainly general verbs and the verbs used in specialized domains, including medical ones, are very specific and they are not included in ADESSE. Finally, in the identification of subclass relations the results obtained in the oncology domain are more solid that the results of the finance domain.

Oncology domain is a more specific domain than the finance domain and the approach presented in this paper obtains better results in specialized domains, where more concepts have to be extracted from the text with respect to the finance domain. The major drawback of this method is that ADESSE does not contain specialized verbs and although the method obtains a good accuracy in the detection of concepts, it cannot identify the most important relations in such a specific domain. As future work, we are planning to include more specific semantic roles in the medical domain by using ADESSE in order to improve the relation extraction performance.

Further validations of the system are planned by means of its application to texts from different medical domains and by using statistical methods for the analysis of the results obtained. Moreover, we intend to extend the system to cover axioms such as the work presented in (Terrientes at al., 2010). The main forecast problem concerning axioms is, however, that the number of participants is a priori unknown. Notwithstanding this fact, the amount of axioms present in a text is irrelevant in comparison to the amount of other knowledge entities.

## ACKNOWLEDGMENTS

This work has been supported by the Spanish Government through project SeCloud (TIN2010-18650).

### REFERENCES

Barrón-Cedeño A, Sierra G, Drouin P, Ananiadou S (2009). An Improved Automatic Term Recognition Method for Spanish. Lecture Notes Comput. Sci., 5449: 126-136.

Berners-Lee T, Hendler J, Lassila O (2001). The semantic web. Sci. Am., 284: 34-43.

Beydoun G, Low G, Henderson-Sellers B, Mouraditis H, Sanz JJG, Pavon J, Gonzales-Perez C (2009). FAML: A Generic Metamodel for MAS Development. IEEE Trans. Softw. Eng., 35(6): 841-863.

Beydoun G, Low G, Mouraditis H, Henderson-Sellers B (2009). A Security-Aware Metamodel For Multi-Agent Systems. J. Inf. Softw. Technol., 51(5): 832-845

Filmore CJ (2002). Framenet and the linking between semantic and syntactic relations. In: Shu-Cuan Tseng (eds): Proceedings of 19th International Conference on Computational Linguistics, pp. 27-36.

Fortuna B, Mladenić D, Grobelnik M (2006). Semi-automatic construction of topic ontologies. In: Semantics, Web Min. Lect. Notes Comput. Sci., 4289: 121-131.

Fernández-Breis JT, Castellanos-Nieves D, Valencia-García R (2009). Measuring individual learning performance in group work from a knowledge integration perspective. Inf. Sci., 179(4): 339–354.

García-Sánchez F, Fernández-Breis JT, Valencia-García R, Gómez JM, Martínez-Béjar R (2008). Combining Semantic Web Technologies with Multi-Agent Systems for Integrated Access to Biological Resources. J. Biomed. Inf., 41(5): 848-859.

Grefenstette G (1994). Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers Norwell, MA, USA.
Grau BC, Horrocks I, Motik B, Parsia B, Patel-Schneider P, Sattler U (2008). OWL 2: The next step for OWL. J. Web Semantics: Sci. Services Agents World Wide Web, 6(4):309–322.
Han H, Elmasri R (2004). Learning rules for conceptual structure on the web. J. Intell. Inf. Syst., 22: 237–256.
Hashim F, Alam GM, Siraj S (2010). Information and communication technology for participatory based decision-making-E-management for administrative efficiency in Higher Education. Int. J. Phys. Sci., 5(4): 383-392.
Jiang X, Tan AH (2010). CRCTOL: a semantic-based domain ontology learning system. J. Am. Soc. Inf. Sci. Technol., 61: 150–168.
Henderson-Sellers B (2011) Bridging metamodels and ontologies in software engineering. J. Syst. Softw., 84(2): 301-313.
Manning CD, Schütze H (1999). Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge.
Miller GA, Beckwith R, Fellbaum C, Gross D, Miller K (1990). WordNet: An on-line lexical database. Int. J. Lexicogr., 3: 235-244.
Miranda-Mena TG, Benítez USL, Ochoa JL, Martnez-Bejar R, Fernandez-Breis JT, Salinas J (2006). A knowledge-based approach to assign breast cancer treatments in oncology units. Expert Syst. Appl., 31: 451-457.
Moreda P, Llorens H, Saquete E, Palomar M (2010). Combining semantic information in question answering. Information Processing and Management (In press corrected proof). doi:10.1016/j.ipm.2010.03.008.
Ochoa JL, Almela A, Ruiz-Martínez JM, Valencia-García R (2010). Efficient Mul-tiword Term Extraction in Spanish. Application to the Financial Domain. In Proceedings of International Conference on Intelligence and Information Technology. Lahore, Pakistan, pp. 426–430.
Palmer M, Gildea D, Kingsbury P (2005). The proposition bank: An annotated corpus of semantic roles. Comp. Lingustics, 1(31): 71-106.
Ruiz-Martínez JM, Castellanos-Nieves D, Valencia-Garcia R, Fernández-Breis, JT, Garcia-Sanchez F, Vivancos-Vicente PJ, Castejón-Garrido JS, Bosco Camón J, Martínez-Béjar R (2009). Accessing Touristic Knowledge Bases through a Natural Language Interface. Lect. Notes Comput. Sci., 5465: 147-160.
Sánchez D, Moreno A (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. Data Knowl. Eng., 64: 600–623.
Santamaría C, Gonzalo J, Verdejo F (2003). Automatic association of web directories with word senses. Comp. Linguistics, 29(3): 485–502.
Shamsfard M, Barforoush A (2004). Learning ontologies from natural language texts. Int. J. Human-Comp. Stud., 60(1): 17-63.
Subramaniam T, Jalab HA, Taga AY (2010) Overview of textual anti-spam filtering techniques. Int. J. Phys. Sci., 5(12): 1869-1882.
Terrientes L, Moreno A, Sánchez D (2010). Discovery of Relation Axioms from the Web. In Knowledge Science, Engineering and Management, Lect. Notes Computer Sci., 6291: 222-233.
Vaamonde G, González-Domínguez F, García-Miguel JM, (2010). ADESSE. A Database with Syntactic and Semantic Annotation of a Corpus of Spanish. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), Valletta (Malta), pp. 1903-1910.
Valencia-García R, Fernández-Breis JT, Ruiz-Martínez JM, García-Sánchez F, Martínez-Béjar R (2008). A knowledge acquisitionmethodology to ontology construction for information retrieval from medical documents. Expert Systems: Knowledge Eng. J., 25(3): 314-334.
Valencia-García R, García-Sánchez F, Castellanos-Nieves D, Fernández-Breis JT (2011). OWLPath: An OWL ontology-guided query editor. IEEE Trans. Syst. Man Cybernet. Part A: Syst. Hum., 41(1): 121-136.
Valencia-García R, Ruiz-Sánchez JM, Vivancos–Vicente PJ, Fernández-Breis JT, Martínez-Béjar R (2004). An incremental approach for discovering medical knowledge from texts. Expert Syst. Appl., 26(3): 291-299.
Van Heijst G, Schreiber AT, Wielinga BJ (1997). Using explicit ontologies in KBS development. Int. J. Hum. Comput. Stud., 46(2–3): 183–292.
Vossen P (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer, Dordrecht, The Netherlands.