Educational Research and Reviews Vol. 6(9), pp. 645-656, 5 September, 2011
Available online at http://www.academicjournals.org/ERR
ISSN 1990-3839 ©2011 Academic Journals

Full Length Research Paper

Identifying patterns for unsupervised learning of multiword terms

José Luis Ochoa, Ángela Almela and Rafael Valencia-García*

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

Accepted 20 July, 2011

The identification of valid terms in any domain is fundamental to its computerization. For this reason, in this paper we present a method for obtaining automated morphosyntactic patterns, which will help researchers to obtain valid terms from the proposed patterns, in order to build quality ontologies for the translation from one language to another, or to find important concepts in short sentences, which can be used as parameters in question-answer systems. For this purpose, we use some statistical methods which show candidates in a pattern vector. Then, a heuristic process unfolds to refine the pattern vector obtained, based on two main parameters: the statistical results previously obtained and the pattern length analyzed. As a result, we obtain the collection of the best patterns for the detection of real multiword terms.

Key words: Morphosyntactic patterns, multiword terms, incremental learning.

INTRODUCTION

Nowadays, the massive amount of information flowing through books, magazines, articles, and mainly through the Web, requires some systems and methods which facilitate its processing. In this line, automatic term recognition (ATR) approaches the task of automatically detecting and extracting the terminological units contained in those collections of texts (Fahmi et al., 2007; Korkontzelos et al., 2008; Zhang et al., 2008). After processing the corpus, the data obtained are stored in a structured language such as those described in (Lassila and Swick, 1999; Dean and Guus, 2004; Bray et al., 2008). Then, data are ready to be utilized for applications like ontology builders, as it is the case of (Gómez-Pérez et al., 2006), which is focused on the public administration domain, semantic search engines (Ding et al., 2004; Byungkyu and Kyungsook, 2010), and question-answering systems (Vargas-Vera and Lytras, 2010; Heinemann, 2010), to name but a few.

As regards the ontology learning process (Shamsfard and Barforoush, 2003; Buitelaar et al., 2005; Zhou, 2007), it entails a series of steps: (i) the extraction of valid terms from a corpus (either texts on the web, formatted texts, plain texts, databases, etc.); (ii) the establishment of taxonomic, non-taxonomic and other relationship types between concepts, along with restrictions and axioms; (iii) the building of the ontology depending on usage, purpose, content type, structure, and representation language; and (iv) the evaluation and maintenance of the created ontology. The present paper explores some methodologies for obtaining valid terms by virtue of morphosyntactic patterns. Our ultimate aim is to assist researchers in the field to perform this task with an unsupervised tool adapted to their specific needs, regardless of the domain and the language of the corpus as Stated by Ochoa et al. 2011.

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

RELATED WORK

In this line of research, Sánchez (2010) presents a domain-independent method for automatically learning terms from the Web for the building of ontologies. It has been manually evaluated in many domains. It uses a

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

basic set of patterns that includes verbal forms for taxonomic relationships, such as the following ones: NP's NP {is|are|was|were} → for example, camera's sensor is; NP of {the | a | an} NP {is | are | was | were} → for example, resolution of the camera is; NP in {the|a|an} NP {is|are|was|were} → for example, exposure in the camera is; NP {have|has|had} NP → for example, camera has ISO; NP {come|comes|came} with NP → for example camera comes with lens cap. In these examples, all the NPs before and after the verb are identified as their domain concepts.

Similarly, in (Imsombut and Kawtrakul, 2007) propose a method for extracting ontological concepts and taxonomic relationships by using explicit expressions of reference in Thai language, namely lexico-syntactic patterns and lists of items. An example of these patterns unfolds as follows: NP1 = (ncn | nct + ncn | npn) + NP, NP2 = NP1 + adj, NP3 = NP + VP where VP = vi | (vt + NP) and NP = NP4 + PP, where PP = prep + NP. The terms extracted from these NP* patterns are stored in a list of candidate terms by means of an estimation function which measures the lexical co-occurrence and eventually obtains the ontological concepts. In order to reduce the large number of candidate terms, co-occurrence scores are subsequently applied to the resulting list.

The effect of the use of different technologies for the establishment of taxonomic relationships has been studied by (Yang and Callan, 2009). He asserts that co-occurrence and lexico-syntactic relations are adequate parameters for obtaining kinship relations of type “is-a” and relations of type “part-of”. In addition, he states that the use of patterns with syntactic features is rather appropriate to obtain “specific terms”.

Finally, (Cimiano and Wenderoth, 2007) present a method for obtaining structures automatically from the web called “qualia”. When the tool was created in 1992, the user had to introduce the structures by hand, and, for this reason, it was not frequently used. Subsequently, the tool was updated with the automation of the process by means of the inclusion of lexico-syntactic patterns. Some of the patterns used were “NP_{QT} is made up of NP'_{C}”, “NP_{QT} comprises NP'_{C}”, and “NP_{QT} consists of NP'_{C}”. The abovementioned studies prove the fundamental need of lexical and morphosyntactic patterns in the automatic extraction of knowledge from text.

PATTERN LEARNING PROCESS

The pattern learning process comprises two sequential phases respectively known as *patterns* identification and debugging and patterns optimization (Figure 1). These stages are applied to each sentence in the text, with the subsequent extraction of the patterns contained in them.

Pattern learning background

It has been proved that a fundamental part in the computerization of a domain is indeed the automatic learning of valid terms; the mere detection of terms in a text is not sufficient. The ultimate goal is that the method is able to provide itself feedback and to learn over time, since there is an increasing amount of terms in each text to process. For this reason, we have developed a method for learning new language patterns from texts automatically and incrementally, which means in practice that morphosyntactic patterns are not necessary from the outset. Providing that the user includes initial patterns, the method will identify the best and the worst ones sorted by categories. Furthermore, the method suggests new patterns not included in the initial list, which implies that when the system processes a new text not only the original patterns are recalled, but also those learnt in previous texts. In this way, the system obtains new valid terms, which will progressively update and improve the term list.

Patterns identification and debugging

This phase involves a series of steps. Firstly, the setting of *guiding patterns*, which establishes the parameters of the most important patterns. Those parameters are *pattern length* and *level of accuracy of the pattern*. *Pattern length* is directly proportional to the length of the multiword term that the user wants to find. For instance, if the user is interested in extracting a valid term consisting of 4 words, a pattern with 4 morphosyntactic elements is required (Table 1).

The *level of accuracy* is the value attached to each morphosyntactic element. It is assigned by Freeling tagger, which includes the set of tags Eagles for the Spanish language. It provides a level of accuracy up to 8 degrees for some morphosyntactic categories, as in the case of pronouns, and a level of accuracy up to 7 degrees in the case of nouns and verbs (Table 2).Given these two parameters, the list of guiding patterns is shown in Table 3 with some examples of patterns which may be obtained after processing a text.It is worth noting that the two values or symbols (XX) presented in the examples for the morphosyntactic elements, do not imply a limitation in their number; in fact, as stated above, the user can define from 1 to 8 values depending on the morphosyntactic element. As a result of the processing of the text with the proposed method, a Candidate Pattern Vector (CPV) is obtained for each pattern length previously defined, including the patterns found with the statistical method, which provides the user with the frequency of occurrence of the patterns in the corpus (NTP).

The next step entails the filtering out of incorrect patterns. Patterns beginning or ending with functional words are not adequate, and thus they must be discarded in this phase. For this purpose, there is a stop-list with Candidate Pattern Vectors beginning and ending with prepositions, pronouns, numerals, determinants,

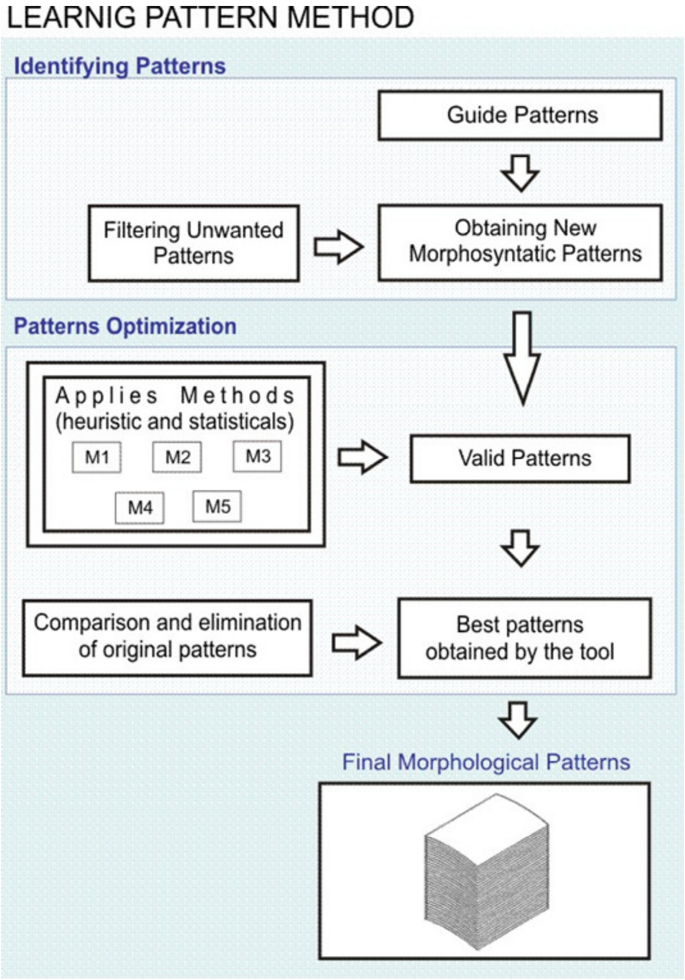

LEARNIG PATTERN METHOD
Identifying Patterns
Guide Patterns
Filtering Unwanted Patterns
Obtaining New Morphosyntactic Patterns
Patterns Optimization
Applies Methods (heuristic and statisticals)
M1
M2
M3
M4
M5
Valid Patterns
Comparison and elimination of original patterns
Best patterns obtained by the tool
Final Morphological Patterns

Figure 1. Pattern learning process.

Table 1. Term length measured by morphosyntatic patterns.

Term length	Morphosyntactic structure
2	xx·xx
3	xx·xx·xx
4	xx·xx·xx·xx

conjunctions, adverbs, verbs and interjections, as well as patterns containing verbs and dates.

Selection of the best linguistic patterns

Once the candidate patterns have been obtained, the best ones must be selected. For this purpose, a combination of heuristic and statistical methods has been applied. Specifically, a method has been implemented for finding patterns from scratch and another method has proved to be adequate for an incremental process with known patterns.

Methodology for pattern learning process from scratch

Firstly, the statistical results of the candidate pattern vector obtained in the previous phase are provided, that is, for each candidate vector, we obtain the first and last element, which will indicate the upper (LS) and the lower limit (LI) of the vector. It is also important to get the number of items contained in each vector; this value will be known as vector size (VS).

After obtaining these values, we apply one of the following equations:

$$if\ \ P(v) \geq 0.0\ and\ P(v) < 1.0 \Rightarrow P(v) * 100 \qquad (1)$$

$$if\ \ P(v) \geq 1.0\ and\ P(v) < 10.0 \Rightarrow P(v) * 10 \qquad (2)$$

$$if\ \ P(v) \geq 10.0 \Rightarrow P(v) * 1 \qquad (3)$$

Where probability P(v) is the result from the following equation:

$$P(v) = (LS * LI) \div VS \qquad (4)$$

Providing that probability exceeds the upper limit, the following equation applies:

$$if\ \ P(v) > LS \Rightarrow LI = 2 \qquad (5)$$

Since the lower limit is rather high in the application of the probability equation. In order to reduce it, the probability equation (4) is applied again.

Finally, if the minimum value is 1, it will be replaced by value 2, with the aim of reducing the candidate patterns and thus the number of candidate terms.

$$if\ \ LI = 1 \Rightarrow LI = 2 \qquad (6)$$

In order to benefit the most relevant patterns, Benefit Factor (BF) is used. It changes the value of each NTP, depending on pattern length (PL). The calculation unfolds as follows:

$$if\ \ LP(x) = 2 \Rightarrow NTP(x) = NTP(x) * 1.3 \qquad (7)$$

$$if\ \ LP(x) = 3 \Rightarrow NTP(x) = NTP(x) * 1.2 \qquad (8)$$

$$if\ \ LP(x) = 4 \Rightarrow NTP(x) = NTP(x) * 1.1 \qquad (9)$$

$$if\ \ LP(x) = 5 \Rightarrow NTP(x) = NTP(x) * 0.9 \qquad (10)$$

$$if\ \ LP(x) = 6 \Rightarrow NTP(x) = NTP(x) * 0.8 \qquad (11)$$

Table 2. Morphosyntatic structures words.

Term	Lemma	Pattern	Description
		Pronouns	
those (aquellos)	that (aquel)	PD0MP000	P Pronoun D Demonstrative 0 Without person M Male P Plural 0 Without case 0 3rd person 0 Without politeness
your (vos)	your tú	PP3CN00P	P Pronoun P Personal 3 Third person C Gender common N Invariable 0 Without case 0 3rd person P Polite
		Noun	
kitten (gatito)	cat (gato)	NCMS00D	N Noun C Common M Male S Singular 00 Without semantic gender D Diminutive grade
		Verb	
we sing (cantamos)	sing (cantar)	VMIP1P0	V Verb P Main I Indicative P Present 1 First person P Plural 0 Without gender

Table 3. Examples of patterns obtained basing on guiding patterns.

Term length	Guiding patterns	Patterns obtained
2	xx·xx	AQ NPNP NP
3	xx·xx·xx	NC SP NCAQ RG AQ
4	xx·xx·xx·xx	AQ NC SP NCNC AO CC NC
5	xx·xx·xx·xx·xx	NC SP DA NC NPAQ SP DI DA NC
6	xx·xx·xx·xx·xx·xx	NC AQ SP NC SP NCAQ SP NC SP NC AQ

$$if\ LP(x)=7 \Rightarrow NTP(x)=NTP(x)*0.7 \tag{12}$$

$$if\ LP(x)=8 \Rightarrow NTP(x)=NTP(x)*0.6 \tag{13}$$

$$if\ LP(x)=9 \Rightarrow NTP(x)=NTP(x)*0.5 \tag{14}$$

$$if\ LP(x)\geq 10 \Rightarrow NTP(x)=NTP(x)*0.4 \tag{15}$$

The final probability values are rounded to integers. Subsequently, the user is provided with the patterns.The candidate patterns may be automatically selected with the parameters defined above. An illustrative example of

Table 4. Limits of each vector.

Variables	CPV (2)	CPV (4)
LS	3791	294
LI	1	1
VS	15	837

Table 5. Identification f patterns suggestedforCPV(2).

Pattern	NTP	NTP + BF	Clipping level	Suggested
NC AQ	3791	4928	506	X
NC NC	978	1271	506	X
AQ NC	885	1151	506	X
NC NP	359	467	506	
AQ AQ	347	451	506	
.				
AO NP	2	3	506	
AQ AO	1	1	506	
NC W	1	1	506	
NC Y	1	1	506	
NC AO	1	1	506	

the process is provided below:

1. The CPV(x) is to be automatically found for each level. For this example, we have used the following strings:

CPV(2):[NC AQ·3791, NC NC·978, AQ NC·885, NC NP·359, AQ AQ·347, NP NC·178, NP NP·121, AO NC·107, AQ NP·71, NP AQ·70, AO NP·2, AQ AO·1, NC W·1, NC Y·1, NC AO·1]
CPV(4):[NC SP DA NC·294, VM SP DA NC·127, NC SP DA NP·74, AQ SP DA NC·72, VM CS DA NC·60, NC SP NC AQ·49, VM DA NC AQ·46, NC AQ SP NC·35, AQ NC SP NC·33, NC SP DA AQ·32, VM VM DA NC·26, NC SP DI NC·24, VM SP DI NC·24, VM SP DA NP·23, NC SP NC VM·22, NC CC DA NC·22, VM DI NC AQ·22, AQ VM DA NC·21, …….., VM CS RN VA·1, AO NC NC VM·1, VM AQ P0 VM·1, NP SP DA AO·1, NC NP P0 VA·1, NC VM NC VM·1, VM NC VM NC·1, NC AQ PR NC·1, VM SP VM VS·1, VM VS RG AQ·1]

2. The upper and the lower limits and the vector size must be obtained for each level (Table 4).
3. Equation (1) is applied. Since the lower limit of the patterns of level 2 and 4 is equal to 1, equation (6) is also applied.

$$P(v2) = (3791 * 2) \div 15 = 505.4666$$
$$P(v4) = (294 * 2) \div 837 = 0.7025$$

4. The probability value is obtained for each level and the final result is rounded to integers; this value is known as *clipping level*:

Probability of CPV(2):

$$P(v2) = 505.47\, is > 10.0 \Rightarrow 505.47 * 1 = 505.47$$
$$P(v2) = 506$$

Probability of CPV(4):

$$P(v4) = 0.70\, is > 0.0 \,\&< 1.0 \Rightarrow 0.70 * 100 = 70.25$$
$$P(v4) = 70$$

5. Considering pattern length for CPV(2), equation (7) is applied. In this case, we have obtained a 30% of BF. Those NTP exceeding the Clipping level are the candidate patterns finally suggested.

$$NTP(x) = 3791 * 1.3 = 4928.3$$
$$NTP(x) = 978 * 1.3 = 1271.4$$
$$\ldots$$
$$NTP(x) = 1 * 1.3 = 1.3$$ (Table 5).

For CPV(4), the result of equation (9) is a 10% of BF:

$$NTP(x) = 294 * 1.1 = 323.4$$
$$NTP(x) = 127 * 1.1 = 139.7$$
$$\ldots$$
$$NTP(x) = 1 * 1.1 = 1.1$$ (Table 6).

Methodology for the incremental pattern learning process

This is based on a limit value (LV); it is shown in the

Table 6. Identification of patterns suggested forCPV(4).

Pattern	NTP	NTP + BF	Clipping level	Suggested
NC SP DA NC	294	323	70	X
VM SP DA NC	127	140	70	X
NC SP DA NP	74	81	70	X
AQ SP DA NC	72	79	70	X
VM CS DA NC	60	66	70	
NC SP NC AQ	49	54	70	
.				
NC VM NC VM	1	1	70	
VM NC VM NC	1	1	70	
NC AQ PR NC	1	1	70	
VM SP VM VS	1	1	70	
VM VS RG AQ	1	1	70	

Table 7. Limits of each vector.

Variables	CPV (2)	CPV (3)
LS	3791	2767
LI	1	1
VS	15	110

following equation:

$$Limit\,Value\,(LV) = (LS + LI) \div 3 \qquad (16)$$

Where:
LS is the largest number of terms contained in each pattern level;
LI is the smallest number of terms contained in each pattern level, and
3 is a constant which divides this range into 3 sections.
When the minimum value is 1, this value is replaced by value 2 (Equation (6)).
The limit value is used to obtain 4 ranks, which are different from each other. Here, we focus on rank 4, with the following equations:

$$LI = LV(x) * 3 \qquad (17)$$

$$LS = LV(x) * 4 \qquad (18)$$

Subsequently, the benefit factor has been calculated (Equations 7 to 15):

The final probability values are rounded to integers, and the user is provided with the patterns.The incremental candidate patterns may also be automatically selected with the parameters defined, as shown by the instance offered here:

1. Pattern vectors are to be automatically found for each level. In this case, we have used the following ones:

CPV(2):[NC AQ·3791, NC NC·978, AQ NC·885, NC NP·359, AQ AQ·347, NP NC·178, NP NP·121, AO NC·107, AQ NP·71, NP AQ·70, AO NP·2, AQ AO·1, NC W·1, NC Y·1, NC AO·1]
CPV(3): [NC SP NC·2767, NC CC NC·687, NC DA NC·371, AQ SP NC·339, NC AQ AQ·239, NC RG AQ·205, NC AQ NC·164, NC SP NP·157, NC NC NC·153, AQ CC AQ·153, NC NC AQ·147, AQ DA NC·134, NC SP AQ·132, AQ CC NC·104, ……, NP DA NP·1, AO NC AQ·1, NC P0 NP·1, NP RG AQ·1, AO CC NC·1, NC P0 NC·1, AQ RG NP·1, NP NC Y·1, NC Y NP·1, NP RG NC·1, NC AO NC·1]

2. The upper and lower limits and the vector size must be obtained for each level (Table 7).

3. Equation (16) is applied to the CPV(2) and CPV(3):

Limit value for CPV(2):

$$LV = (3791 + 2) \div 3 = 1264.33$$

Limit value for CPV(3):

$$LV = (2767 + 2) \div 3 = 923$$

4. The upper and the lower limits are obtained for each level, depending on the range in which they are; by means of the application of Equations (17) and (18), respectively (Table 8):

For CPV (2):

Table 8. Limits of each vector.

Variables	Range 4	
	LI	LS
CPV (2)	3792	5056
CPV (3)	2769	3692

Table 9. Patterns suggested for CPV(2).

Pattern	NTP	NTP + BF	LI	LS	Suggested
NC AQ	3791	4928	2528	3792	X
NC NC	978	1271	2528	3792	
AQ NC	885	1151	2528	3792	
NC NP	359	467	2528	3792	
AQ AQ	347	451	2528	3792	
.					
AO NP	2	3	2528	3792	
AQ AO	1	1	2528	3792	
NC W	1	1	2528	3792	
NC Y	1	1	2528	3792	
NC AO	1	1	2528	3792	

Table 10. Patterns suggested for CPV(3).

Pattern	NTP	NTP + BF	LI	LS	Suggested
NC SP NC	2767	3320	1846	2769	X
NC CC NC	687	824	1846	2769	
NC DA NC	371	445	1846	2769	
AQ SP NC	339	407	1846	2769	
.					
AQ DD NC	17	20	1846	2769	
NP AQ NC	11	13	1846	2769	
NP CC AQ	8	10	1846	2769	
NP SP AQ	5	6	1846	2769	

$$LI = (1264 * 3) = 3792$$
$$LS = (1264 * 4) = 5056$$

For CPV (3):

$$LI = (923 * 3) = 2769$$
$$LS = (923 * 4) = 3692$$

5. Considering pattern length for CPV(2), equation (7) is applied. In this case, we have obtained a 30% of BF. Those NTP who are inside the Range, are the candidate patterns finally suggested (Table 9):

$$NTP(x) = 3791 * 1.3 = 4928.3$$

$$NTP(x) = 978 * 1.3 = 1271.4$$
$$\ldots$$
$$NTP(x) = 1 * 1.3 = 1.3$$

For CPV(3), equation (8) is applied, with a 20% of BF as a result (Table 10):

$$NTP(x) = 2767 * 1.2 = 3320.4$$
$$NTP(x) = 687 * 1.2 = 824.4$$
$$\ldots$$
$$NTP(x) = 1 * 1.2 = 1.2$$

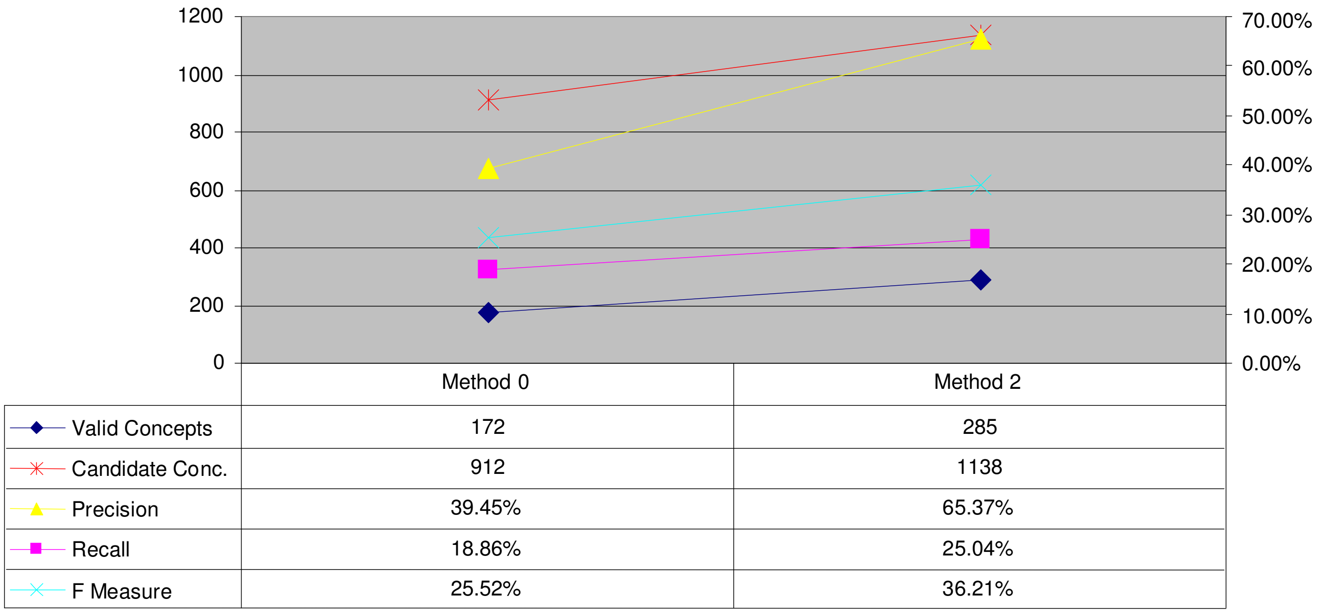
VALIDATION IN THE FINANCIAL AND CANCER DOMAIN

The selected domain

We have conducted our research on the financial

Table 11. Guiding table of patterns obtained by the tool and by the expert.

Patterns	M1	M2	M3	M4	M5
Recommended patterns	27	39	12	6	4
Original recommended patterns	14	12	9	4	2
Not recommended patterns:	1213	1202	1228	1234	1236
Original patterns not recommended	18	20	22	28	30
Patterns not covered	0	0	0	0	0
Original patterns not found	4	4	4	4	4
Original patterns off limits	0	0	0	0	0



	Method 0	Method 2
◆ Valid Concepts	172	285
✳ Candidate Conc.	912	1138
▲ Precision	39.45%	65.37%
■ Recall	18.86%	25.04%
× F Measure	25.52%	36.21%

Figure 2. The improvements achieved by the scratch method in the financial domain.

For our study, we have collected a corpus of 31 financial articles comprising 15.868 words, and a second corpus of 19 articles on cancer containing 94.829 words.

Assessment procedures

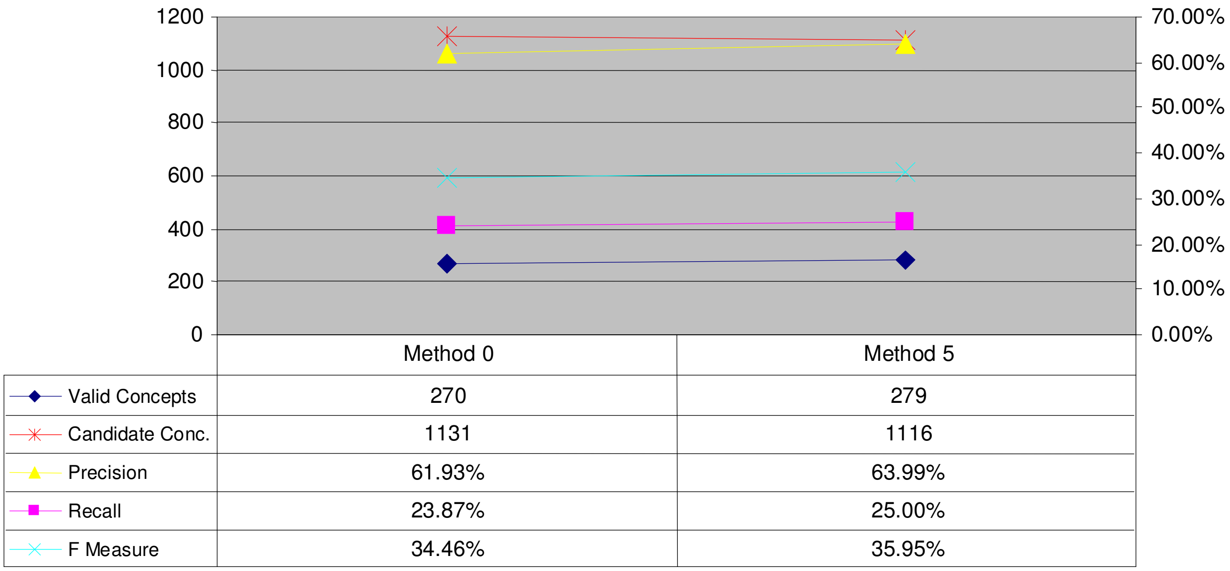
In order to achieve reliable results, we have conducted different procedures modifying certain parameters of the applied heuristics. Specifically, we have performed the test with five different methods. In this section, we are going to focus on Method 2 (M2) and Method 5 (M5), which are the most adequate ones for morphological patterns from scratch and incremental learning.

Detection of patterns from scratch in the financial domain

To identify the best method, we have compared the set of patterns created by the expert for extracting terms from the corpus with the total amount of valid terms detected by means of these patterns.

As can be seen in Table 11, the method which has obtained the highest number of recommended patterns is M2. A total amount of 12 of them, was originally recommended by the expert.

From a total of 39 patterns created by the expert, 12 of them, were originally recommended and 4 of them, were not found in the corpus, giving a total of 16 patterns, accounting for 41.03% of the original patterns created by the expert. By applying these patterns to the financial corpus with a sample of 70%, we obtain the results observed in Figure 2, where Method 0, are the results obtained with the patterns created by the expert and Method 2 are the results obtained with the patterns created by the tool. As can be seen, Precision has improved from 39.45 to 65.37%. On the other hand, Recall level has increased from 18.86 to 25.04% (from 172 to 285 valid terms), which can be deemed as a modest improvement.



	Method 0	Method 5
◆ Valid Concepts	270	279
✳ Candidate Conc.	1131	1116
▲ Precision	61.93%	63.99%
■ Recall	23.87%	25.00%
× F Measure	34.46%	35.95%

Figure 3. The improvements achieved by the incremental method in the financial domain.

Table 12. Guiding table of patterns obtained by the tool and by the expert.

	M1	M2	M3	M4	M5
Recommended patterns	119	185	12	4	5
Original recommended patterns	19	17	8	3	2
Not recommended patterns	3885	3819	3992	4000	3999
Original patterns not recommended	14	16	25	30	31
Patterns not covered	1	1	1	1	1
Original patterns not found	3	3	3	3	3
Original patterns off limits	0	0	0	0	0

Detection of patterns by means of the incremental method

Like in the previous assessment, here we have compared the set of patterns created by the expert for extracting terms from the corpus with the total amount of valid terms detected by means of these patterns.

In M5, from a total of 4 patterns correctly identified by the tool, 2 of them have been created by the expert (Table 11).

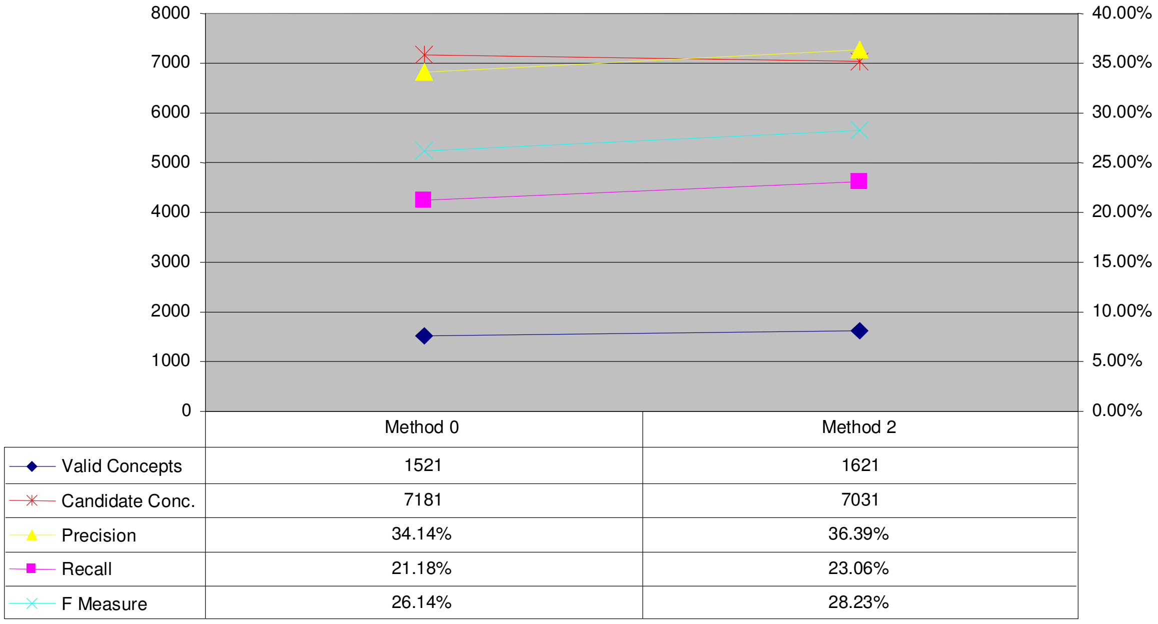
By applying these patterns to the financial corpus with a sample of 70%, we obtain the results observed in Figure 3: Where Method 0, are the results obtained with the patterns created by the expert and method 5 are the results obtained with the patterns created by the expert, including the patterns created by the tool. In this case, the number of valid terms has grown from 270 to 279, which entails an improvement in Precision from 61.93to 63.99% and in Recall from 23.87 to 25%. Despite the fact that this is not a remarkable improvement, it is worth noting that the amount of candidate terms has been reduced from 1131 to 1116, which is to be considered as strength of the method.

Detection of patterns from scratch in the cancer domain

As can be seen in Table 12, the method which has obtained the highest number of recommended patterns is M2. A total amount of 17 was originally recommended by the expert.

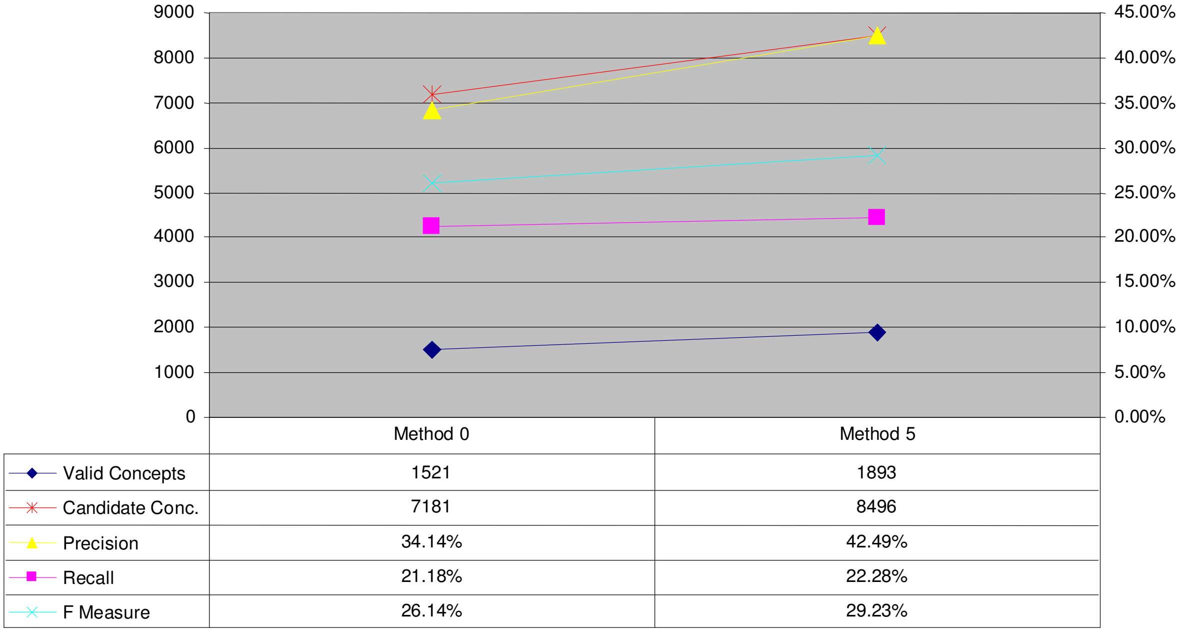
This means that from a total of 185 patterns devised by the expert, 17 of them were originally recommended by the expert and 3 of them, were not found in the corpus, giving a total of 20 patterns, accounting for 55.55% of the original patterns created by the expert. By applying these patterns to the cancer corpus with a sample of 70%, we have obtained the results observed in Figure 4.

Where method 0, are the results obtained with the patterns created by the expert and Method 2 are the results obtained with the patterns created by the tool. In this test, there has been an increase in the amount of valid terms from 1521 to 1621, resulting in improvements



	Method 0	Method 2
◆ Valid Concepts	1521	1621
✳ Candidate Conc.	7181	7031
▲ Precision	34.14%	36.39%
■ Recall	21.18%	23.06%
✕ F Measure	26.14%	28.23%

Figure 4. The improvements achieved by the scratch method in the cancer domain.



	Method 0	Method 5
◆ Valid Concepts	1521	1893
✳ Candidate Conc.	7181	8496
▲ Precision	34.14%	42.49%
■ Recall	21.18%	22.28%
✕ F Measure	26.14%	29.23%

Figure 5. The improvements achieved by the incremental method in the cancer domain.

in Precision value (from 34.14 to 36.39%) and in Recall value (from 21.18 to 23.06%).

Detection of patterns by means of the incremental method in the cancer domain

The elements for comparison in this section are the same ones as in the previous test. M5 section in Table 12 shows a total of 5 patterns found by the tool, 2 of them having been created by the expert. By means of the application of these patterns to the cancer corpus with a sample of 70%, we have obtained the results represented in Figure 5.

Where Method 0, are the results obtained with the patterns created by the expert and Method 5 are the results obtained with the patterns created by the expert, including the patterns created by the tool. There is an increase in the amount of valid terms from 1521 to 1893, which entails a refining of Precision (from 34.14 to 42.49%), as well as of Recall (from 21.18 to 22.28%). Nonetheless, unlike the test on the financial domain, in this case the number of candidate terms was not reduced.

CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a method which provides the user with morphological patterns automatically, without the need to know any specific pattern. The only requirement is a list indicating pattern length and specialization level. For instance, in order to obtain patterns of length 3 with expertise levels 2 · 3 · 2 in their morphological elements respectively, the correct combination would be "XX · XXX · XX". As a result of theprocessing of the text, we will obtain an ordered list with the best patterns found. As previously mentioned, the method may be used in any domain. The input data must be in plain text format.

Apart from the list with the best patterns, the discarded patterns may be retrieved too if needed. In the event of having a list of patterns extracted by an expert, the tool provides the option to compare these patterns with those automatically obtained in order to assess their quality.The method presented in this paper has been developed to assist researchers in the field to detect valid terms in a corpus. This system may well complement previous work on the building of ontologies in different languages (Cimiano, 2006). Some representative instances of this line of research are (Abascal-Mena, 2009; Sánchez and Moreno, 2004; Blaschke and Valencia, 2002; Pulido et al., 2007) for the English language; (Lee et al., 2007) for the Chinese language; (Kawtrakul et al., 2004) for Thai; (Khosravi and Vazifedoost, 2007) for Persian; (Bontas et al., 2005; Kietz et al., 2000) for the German language; (Valencia-Garcia et al., 2006) for Spanish; and (Passant, 2007) for French. Other related pieces of research are (Carpuat et al., 2002), in which we study the building of ontologies across Dutch, Italian, Spanish, German, French, Czech and Estonian; Cimiano's exploration of domain concepts acquisition (Cimiano, 2006); study of German-Spanish machine translation (López et al., 2010); as well as research conducted on ontologies and geographical information (Kauppinen et al., 2006), on information retrieval (Cimiano and Wenderoth, 2007; Ruiz-Casado et al., 2007), on search engines technology (Ding et al., 2004), and even research on word senses disambiguation (Almuhareb and Poesio, 2006). All these works may benefit from the implementation of our method, since manual terminology management is unable to process the massive amount of information published daily (Sánchez, 2010).

One of the major advantages gained from the application of linguistic patterns is the finding of taxonomic relationships and non taxonomic relations (Ochoa et al., 2011a, b, c). In this line, (Hearst, 1992) has studied and defined a set of independent patterns for hyponymy, which has served as a basis for new learning approaches (Pasca, 2004). Similarly, hypernym can also define linguistic patterns expressing functions (Cimiano and Wenderoth, 2007), as well as metaphors and similes (Veale and Hao, 2007), and other semantic relationships such as meronymy, holonymy, telicity, etc. (Ruiz-Casado et al., 2007). A representative example of this appears in (Berland and Charniak, 1999), where these authors define a set of general guidelines for finding meronymic relationships in the text.

A further strength of our method is that it includes Almuhareb and Poesio's procedure for improving precision in extraction systems (Almuhareb and Poesio, 2004). They proved that the presence of verbal forms and the avoidance of modifiers guarantees that the attributes stand for real terms. An example of this may be found in (Ochoa et al., 2010).

The main weakness of the present method is that it still lacks a proper module for filtering the stored patterns over time. The inclusion of this module will become important when the system had processed several corpora from different domains, since many patterns will have been stored for the same domain. This may produce an excessive amount of candidates to be discarded. We are currently working to develop this module, along with new heuristics which will make a more precise selection of candidate patterns. For this purpose, we will intend to take the candidate terms obtained through each pattern as a baseline, after a comparison against a complete list of valid terms from each specific domain. In this way, this tool can indeed facilitate the progress of NLP in Spanish.

REFERENCES

Abascal-Mena R (2009). Towards a semantic web: ontology development based on the extraction of semantic concepts from digital documents. Proceedings of the WSEAES 13th international conference on Computers, Rodos, Greece, pp. 519-525.

Almuhareb A, Poesio M (2004). Attribute-based and value-based clustering: an evaluation. In: Proceedings of the Conference on Empirical Methods and Natural Language Proceedings, Barcelona, Spain, pp. 158–165.

Almuhareb A, Poesio M (2006). MSDA: Word-sense discrimination using context vectors and attributes. In: Proceedings of European Conference on Artificial Intelligence, pp. 543–547.

Berland M, Charniak E (1999). Finding parts in very large corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Maryland, USA, pp. 57–64.

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

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

Blaschke C, Valencia A (2002). Automatic ontology construction from the literature. Genome Inform., 13: 201–213.

Bontas EP, Schlangen D, Schrader T (2005). Creating ontologies for content representation — The OntoSeed suite. Lect. Notes Comput.

Sci., 3761: 1296-1313.

Bray T, Paoli J, Sperberg-Mc Queen CM (2008). Extensible markup language (XML) 1.0 W3C recommendation. Technical report, Available: http://www.w3.org/TR/REC-xml/.

Buitelaar P, Cimiano P, Magnini B (2005). Ontology learning from text: An overview. In P. Buitelaar, P. Cimiano, B. Magnini (Eds.), Ontology learning from text: Methods, Eval. Appl., 123: 3-12.

Byungkyu P, Kyungsook H (2010). An ontology-based search engine for protein-protein interactions. In the Eighth Asia-Pacific Bioinformatics Conference (APBC 2010), Bangalore, India, 11: S23.

Carpuat M, Ngai G, Fung P, Church K (2002). Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet. In Proceedings of the 1st Global WordNet Conference, Mysore, India, pp. 284-292.

Cimiano P (2006). Ontology Learning and Population from Text. Algorithms, Evaluation and Applications. Springer-Verlag, New York.

Cimiano P, Wenderoth J (2007). Automatic acquisition of ranked qualia structures from the web. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Prague, pp. 888–895.

Dean M, Guus S (2004). OWL Web Ontology Language Reference, W3C Recommendation, Available: http://www.w3.org/TR/2004/REC-owl-ref-20040210.

Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi VC, Sachs J (2004). Swoogle: A search and metadata engine for the semantic web. In: Proceedings of the 13th ACM Conference on Information and Knowledge Management, ACM Press, pp. 652–659.

Fahmi I, Bouma G, van der Plas L (2007). Improving statistical method using known terms for automatic term extraction. In Computational Linguistics in the Netherlands. CLIN 17, pp. 1-8.

Fernández-Breis JT, Castellanos-Nieves D, Valencia-García R (2009). Measuring individual learning performance in group work from a knowledge integration perspective. Inform. 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. Inform., 41(5): 848-859.

Gómez-Pérez A, Ortiz-Rodríguez F, Villazón-Terrazas B (2006). Legal Ontologies for the Spanish e-Government. Lect. Notes Comp. Sci., 4177: 301-310.

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.

Hearst MA (1992). Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, pp. 539–545.

Heinemann T (2010). The question-response system of Danish. J. Pragmat., 42: 2703-2725.

Henderson-Sellers B (2011) Bridging metamodels and ontologies in software engineering. J. Syst. Softw., 84(2): 301-313.

Imsombut A, Kawtrakul A (2007). Automatic building of an ontology on the basis of text corpora in Thai. Lang. Resour. Eval., 42: 137-149.

Kauppinen T, Henriksson R, Väätäinen J, Deichstetter C, Hyvönen E (2006). Ontology-based modeling and visualization of cultural spatio-temporal knowledge. In Developments in Artificial Intelligence and the Semantic Web, Proceedings of the 12th Finnish AI Conference STeP, pp. 37-45.

Kawtrakul A, Suktarachan M, Imsombut A (2004). Automatic Thai Ontology Construction and Maintenance System. Proc. of OntoLex Workshop on LREC, Portugal.

Khosravi F, Vazifedoost A (2007). Creating a Persian Ontology through Thesaurus Reengineering for Organizing the Digital Library of the National Library of Iran, ICOLIS 2007, Kuala Lumpur: LISU, FCSIT, pp. 41-53.

Kietz JU, Volz R, Maedche A (2000). A method for semi-automatic ontology acquisition from a corporate intranet. In EKAW-2000 Workshop “Ontologies and Text”, Juan-Les-Pins, France, pp. 2-6.

Korkontzelos I, Klapaftis IP, Manandhar S (2008). Reviewing and evaluating automatic term recognition techniques. Lect. Notes Comp. Sci., 5221: 248-259.

Lassila O, Swick RR (1999). Resource Description Framework (RDF) Model and Syntax Specification, Available: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

Lee CS, Kao YF, Kuo YH, Wang MH (2007). Automated ontology construction for unstructured text documents. Data Knowl. Eng., 60: 547-566.

López VF, Alonso L, Moreno MN (2010). A SOMAgent for machine translation. Expert Syst. Appl., 12: 7993-7996.

Ochoa JL, Almela A, Ruiz-Martínez JM, Valencia-García R (2010). Efficient multiword term extraction in Spanish. Application to the Financial Domain, Proc. ICIIT 2010, Lahore, Pakistan, 1: 426-430.

Ochoa JL, Hernandez-Alcaraz ML, Almela A, Valencia-Garcia R (2011a). Learning Semantic Relations from Spanish Natural Language Documents in the Financial Domain, ICCMS, Mumbai, India.

Ochoa JL, Hernández-Alcaraz ML, Valencia-García, R, Martínez-Béjar R (2011b). A semantic role based Ontology Learning approach for Spanish texts. In DCAI 2011, Salamanca, Spain, pp. 273-280. Doi: 10.1007/978-3-642-19934-9_35.

Ochoa JL, Hernández-Alcaraz ML, Valencia-García R, Martínez-Béjar R (2011c). A semantic role-based methodology for knowledge acquisition from Spanish documents. Int. J. Phys. Sci., 6(7): 1755-1765.

Pasca M (2004). Acquisition of categorized named entities for web search. In: Proceedings of the 13th ACM International Conference on Information and Knowledge Management, USA, pp. 137–145.

Passant A (2007). Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs. In Proceedings of International Conference on Weblogs and Social Media, Boulder, Colorado.

Pulido JRG, Flores SBF, Reyes, PD, Diaz RA, Castillo JJC (2007). In the quest of specific-domain ontology components for the semantic web. The 6th International Workshop on Self-Organizing Maps (WSOM), Bielefeld University, Germany, 2007.

Ruiz-Casado M, Alfonseca E, Castells P (2007). Automatising the learning of lexical patterns: an application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. Data Knowl. Eng., 61: 484–499.

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 Comp. Sci., 5465: 147-160.

Sánchez D (2010). A methodology to learn ontological attributes from the Web. Data Knowl. Eng., 69: 573-597.

Sánchez D, Moreno A (2004). Creating ontologies from Web documents. Recent Advances in Artificial Intelligence Research and Development. IOS Press, 113: 11-18.

Shamsfard M,Barforoush AA (2003). The State of the Art in Ontology Learning: A Framework for Comparison. Knowl. Eng. Rev., 18: 293-316.

Subramaniam T, Jalab HA, Taqa AY (2010) Overview of textual anti-spam filtering techniques. Int. J. Phys. Sci., 5(12): 1869-1882.

Valencia-Garcia R, Castellanos-Nieves D, Fernández-Breis J, Vivancos-Vicente P (2006). A Methodology for Extracting Ontological Knowledge from Spanish Documents. Lect. Notes Comp. Sci., 3878: 71-80.

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.

Vargas-Vera M, Lytras MD (2010). AQUA: A Closed-Domain Question Answering System. Inf. Syst. Manage., 27: 217-225.

Yang H, Callan J (2009). Feature Selection for Automatic Taxonomy Induction. In Proceedings of the 32nd Annual ACM SIGIR Conference (SIGIR2009), Boston, MA, USA, pp. 19-23.

Zhang Z, Iria J, Brewster C, Ciravegna F (2008). A comparative evaluation of term recognition algorithms. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), pp. 2108-2113.

Zhou L (2007). Ontology learning: State-of-the-art and open issues. Inf. Technol. Manage., 8: 241–252.