Scientific Research and Essays Vol. 6(26), pp. 5563-5578, 9 November, 2011
DOI: 10.5897/SRE11.1299
Available online at http://www.academicjournals.org/SRE
ISSN 1992-2248 © 2011 Academic Journals

*Full Length Research Paper*

# Learning morphosyntactic patterns for multiword term extraction

**José Luis Ochoa[1], Ángela Almela[2], Maria Luisa Hernández-Alcaraz[3] and Rafael Valencia-García[3]***

[1]Departamento de Ingeniería Industrial, Universidad de Sonora. Blvd. Rosales y Transversal, Hermosillo, Sonora, México. C.P. 83000.
[2] English Department, Universidad de Murcia, España.
[3] Faculty of Computer Science, Universidad de Murcia 30071 Espinardo (Murcia). España.

Accepted 4 October, 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 obtain valid terms from the proposed patterns, in order to build quality ontologies for the translation from one language to another, or to find relevant terms 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, basing on two main parameters: the statistical results previously obtained and the length of the pattern 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 (Bray et al., 2008; Dean and Guus, 2004; Lassila and Swick, 1999). 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 (Byungkyu and Kyungsook, 2010; Ding et al., 2004), and question-answering systems (Heinemann, 2010; Vargas-Vera and Lytras, 2010), to name but a few.

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 (Ruiz-Martínez et al., 2009), education (Fernández-Breis et al., 2009; Hashim et al., 2010), natural language processing (Ercan, 2010; Sezer, 2011; Subramaniam et al., 2010; Yang et al., 2010a), information security (Vorobiev and Bekmamedova, 2010), web services (Cömert et al., 2010) and software engineering (Beydoun et al., 2009a,b; Lasheras et al, 2009).

Due to the outstanding importance of ontologies, different methodologies for their design and building have

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

been proposed. In this respect, it can be said that the process of manual ontology construction poses a major problem, since it involves several resources and time-consuming tasks (Fortuna et al., 2008). Thus, the generation and development of methods and software tools to support the automatic construction of ontologies from natural language texts is a relevant research area, which is known as ontology learning.

As regards the ontology learning process, it entails a series of steps (Buitelaar et al., 2005; Shamsfard and Barforoush, 2003; Zhou, 2007): (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.

This paper explores some methodologies for obtaining valid terms by virtue of morphosyntactic patterns. Lexical and morphosyntactic patterns are important in the automatic extraction of knowledge from text since they are needed in linguistic-based term extraction processes (Ochoa et al., 2011c). Other approaches use statistical methods for term extraction as well (Yang et al., 2010b). 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.

The remainder of this paper is structured as follows: 1) An overview of the related work and the pattern learning process; 2) Description of the results of the experiment; 3) Discussion about the results; 4) Conclusions and future work.

## MATERIALS AND METHODS

Firstly, related work underlining the importance of morphosyntactic patterns in term extraction for ontology learning is set out. Next, the pattern learning process is explained in detail.

### 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 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, (Imsombut and Kawtrakul, 2007) proposed 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). They 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 manually, 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 aforementioned studies prove the fundamental need of lexical and morphosyntactic patterns in the automatic extraction of knowledge from text.

### The pattern learning process

The pattern learning process comprises two sequential stages 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 both 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

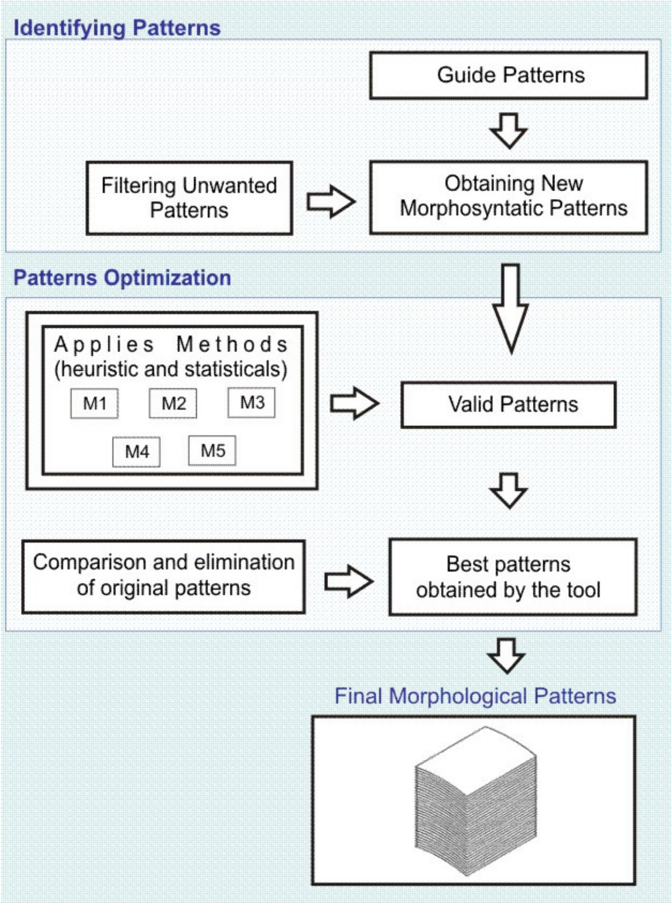

Identifying Patterns
Guide Patterns
Filtering Unwanted Patterns
Obtaining New Morphosyntactic Patterns
Patterns Optimization
Applies Methods (heuristic and statisticals)
M1
M2
M3
M4
M5
Valid Patterns
Comparison and elimination of original patterns
Best patterns obtained by the tool
Final Morphological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

**Figure 1.** Pattern learning process.

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 2).

The level of accuracy is the value of each morphosyntactic element. This value is defined by the Eagles[1] tagset, which are used in the linguistic tool Freeling[2]. This application has been used for tagging the words in the corpus automatically, since it includes Spanish among other languages. As can be seen in Table 1, the level of accuracy is used for the definition of each morphological level in the words or in the morphosyntactical elements, given that there are words such as pronouns which can be defined with up to 8 levels of accuracy. For instance, “those” is defined as a masculine plural demonstrative pronoun; the morphological features scoring 0 are not relevant. As regards nouns, they have a maximum of 6 levels of accuracy; for example, "kitten" is defined as a masculine singular diminutive common noun. Thus, the degree of specialization varies according to the POS tag attached to each word. In the Eagles tagset, it ranges from a minimum of 1, such as numbers, punctuation marks or interjections, and a maximum of 8, as it is the case for pronouns. Other instances would be the abovementioned nouns, with a degree of specialization 6, or verbs, up to 7 degrees (Table 1).

Given these parameters, it is possible to define a list of guiding patterns which will comprise one or more elements grouped and identified with symbol “X” standing for the level of accuracy, and periods “·” delimiting the groups and stating pattern length, as shown in Table 2. For example, the guiding pattern “XX·XX” has a length of 2 components, with two groups of morphosyntactic elements and two degrees of specialization each. This means in practice that only the two first categories of each morphological tag will be obtained. Thus, given the text “aquellos gatitos” (those kittens), tagged by Freeling as “aquellos·PD0MP000 gatitos·NCMP00D”, the abbreviation “PD” will be kept from the first tag, and “NC” from the second one, giving as a result the pattern “PD·NC”. As can be seen in Table 2, this one and the other guiding patterns may offer a wide variety of pattern combinations extracted from the processing of the text.

It is worth noting that the symbols (XX) standing for the morphosyntactic elements in the examples do not imply a limitation in their number; in fact, as stated above, the user can define from 1 to 8 values. Therefore, the user is able to define terms from their most general element (level of accuracy 1), such as a noun represented as "N", to the highest level of complexity, like a pronoun represented as

[1] http://www.lsi.upc.es/~nlp/tools/parole-sp.html

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

**Table 1**. Morphosyntactic structures from some words.

| Term | Lemma | Morphological tag | Description |
|---|---|---|---|
| **Pronouns** | | | |
| Those (aquellos) | That (aquel) | PD0MP000 | P Pronoun<br>D Demonstrative<br>0 No person<br>M Male<br>P Plural<br>0 No case<br>0 3rd person<br>0 No politeness |
| You (vosotros) | You (tú) | PP3CN00P | P Pronoun<br>P Personal<br>3 3rd person<br>C Common for gender<br>N Invariable<br>0 No case<br>0 3rd person<br>P Polite |
| **Noun** | | | |
| Kitten (gatito) | Cat (gato) | NCMS00D | N Noun<br>C Common<br>M Male<br>S Singular<br>00 No semantic gender<br>D Diminutive grade |
| **Verb** | | | |
| we sing(cantamos) | Sing (cantar) | VMIP1P0 | V Verb<br>P Main<br>I Indicative<br>P Present<br>1 First person<br>P Plural<br>0 No gender |

"PP2CSN00", standing for pronoun, personal, 2nd person, common, nominative, and no number, respectively. The level of accuracy is of utmost importance, since the terms obtained from the text depend on it: the more specific the patterns, the more accurate the terms obtained. This would imply a lower amount of terms extracted, reducing the risk of noise in the extraction process. Among other reasons, the guiding patterns where so defined because of their capacity for adaptation to several languages, since, as would be discussed, the resulting patterns have to be filtered according to the particular features of the language, rejecting those unwanted morphosyntatic elements. It is worth noting that grammatical patterns may vary for each language. For instance, English adjectives usually precede nouns, for example, "red AQ car NC", whereas Spanish nouns usually precede adjectives, for example, "coche NC rojo AQ". Similarly, there are differences in the expression of possession: Saxon genitive is fairly frequent in English terms, for example. "Alzheimer's NP disease NC", whereas the equivalent in Spanish would require a preposition, "enfermedad NC de SPS Alzheimer NP".

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 by means of 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, ending or involving functional words are not deemed adequate in this case, and thus they must be discarded in this phase. For this purpose, there are three stoplists containing a candidate pattern vector with prepositions, pronouns, numerals, determinants, conjunctions, adverbs, verbs and interjections, that is, morphological elements which are unwanted. Particularly in this stage, the features of the language are taken into account, since the stoplist will comprise the unwanted morphological elements in each language in the adequate

**Table 2.** Examples of patterns obtained basing on guiding patterns.

| Term length | Guiding patterns | Patterns obtained |
|---|---|---|
| 2 | xx·xx | AQ NC<br>NC NC<br>CS PD<br>DA NC |
| 3 | xx·xx·xx | NC SP NC<br>AQ RG AQ |
| 4 | xx·xx·xx·xx | AQ NC SP NC<br>NC AO CC NC |
| 5 | xx·xx·xx·xx·xx | NC SP DA NC NC<br>AQ SP DI DA NC |
| 6 | xx·xx·xx·xx·xx·xx | NC AQ SP NC SP NC<br>AQ SP NC SP NC AQ |

**Table 3.** Limits of each vector.

| | |
|---|---|
| LS | 2946 |
| LI | 62 |
| VS | 8 |

position.

### *Selection of the best linguistic patterns*

Once 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 been adopted for an incremental process with known patterns. Specifically, five methods (M1-M5) have been applied for the selection of the best linguistic patterns; these methods will be explained in detail below. M1-M2 have been developed for pattern learning from scratch, whereas M3, M4 and M5 are adopted for incremental learning.

### **Methodology for pattern learning process from scratch (methodology M1*)***

Firstly, the statistical results of the candidate pattern vector obtained in the previous stage are provided, that is, for each candidate vector, we obtain the first and the 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 parameter will be referred to 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:

$$if\ P(v) > LS \Rightarrow LI = 1 \qquad \textit{(5)}$$

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. 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 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 (1): [NC·2946, VM·1898, AQ·1043, NP·933, VS·179, VA·154, AO·62]

2) The upper and the lower limits and the vector size must be obtained for each level (see Table 3).

3) Equation (1) is applied.

$$P(v1) = (2946 * 62) \div 8 = 22831.5$$

**Table 4.** Identification of patterns suggested for CPV(1).

| Pattern | NTP | Clipping Level | Suggested |
|---|---|---|---|
| NC | 2946 | 368 | X |
| VM | 1898 | 368 | X |
| AQ | 1043 | 368 | X |
| NP | 933 | 368 | X |
| VS | 179 | 368 | |
| VA | 154 | 368 | |
| AO | 62 | 368 | |

4) Since the probability value is greater than the maximum value, Equation (5) is also applied.

$$P(v1) = (2946 * 1) \div 8 = 368.25$$

5) 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(1):*

$$P(v1) = 368.25\, is > 10.0 \Rightarrow 368.25 * 1 = 368.25$$
$$P(v1) = 368$$

6) Comparing each pattern with the Clipping Level, those NTP exceeding it will be the candidate patterns finally suggested (Table 4).

**Methodology for pattern learning process from scratch (methodology M2)**

M1 serves as a basis for M2, with two main differences: the modification of Equation (5) and the addition of a Benefit Factor (BF) to the NTP values obtained. In this way, Equations (1) to (4) remain the same in this method, and providing that probability exceeds the upper limit, the following equation applies:

$$if\ P(v) > LS \Rightarrow LI = 2 \quad (6)$$

In order to reduce it, Equation (4) is applied again. Finally, if the minimum value is 1, it will be replaced with value 2, with the aim of reducing the candidate patterns and thus the number of candidate terms.

$$if\ LI = 1 \Rightarrow LI = 2 \quad (7)$$

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:

$$if\ LP(x) = 2 \Rightarrow NTP(x) = NTP(x) * 1.3 \quad (8)$$

$$if\ LP(x) = 3 \Rightarrow NTP(x) = NTP(x) * 1.2 \quad (9)$$

$$if\ LP(x) = 4 \Rightarrow NTP(x) = NTP(x) * 1.1 \quad (10)$$

$$if\ LP(x) = 5 \Rightarrow NTP(x) = NTP(x) * 0.9 \quad (11)$$

$$if\ LP(x) = 6 \Rightarrow NTP(x) = NTP(x) * 0.8 \quad (12)$$

$$if\ LP(x) = 7 \Rightarrow NTP(x) = NTP(x) * 0.7 \quad (13)$$

$$if\ LP(x) = 8 \Rightarrow NTP(x) = NTP(x) * 0.6 \quad (14)$$

$$if\ LP(x) = 9 \Rightarrow NTP(x) = NTP(x) * 0.5 \quad (15)$$

$$if\ LP(x) \geq 10 \Rightarrow NTP(x) = NTP(x) * 0.4 \quad (16)$$

The final probability values are rounded to integers. Subsequently, the user is provided with the patterns. Thus, the candidate patterns may be automatically selected with the parameters defined above. An illustrative example 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, $NTP(x) = 1 * 1.1 = 1.1$ (Table 7).

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 (see Table 5).

3) Equation (1) is applied. Since the lower limit of the patterns of level 2 and 4 is equal to 1, Equation (7) is also applied.

**Table 5.** Limits of each vector.

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

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

4) The Clipping Level is computed:

*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 (8) is applied. In this case, we have obtained a 30% 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 6).

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

$$NTP(x) = 294 * 1.1 = 323.4$$
$$NTP(x) = 127 * 1.1 = 139.7$$

### Methodology for the incremental pattern learning process (methodology M3 to M5)

It is based on a limit value (LV); it is shown in the following equation:

$$Limit\,Value\,(LV) = (LS + LI) \div 3 \qquad (17)$$

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 (7)).

The limit value is used to obtain 4 ranks, which are different from each other. The following equations have been computed:

Rank 1 (methodology M3)

$$LI = LV(x) * 0 \qquad (18)$$

$$LS = LV(x) * 1 \qquad (19)$$

Rank 2 (methodology M4)

$$LI = LV(x) * 1 \qquad (20)$$

$$LS = LV(x) * 2 \qquad (21)$$

Rank 3 (methodology M5)

$$LI = LV(x) * 2 \qquad (22)$$

$$LS = LV(x) * 3 \qquad (23)$$

Subsequently, the benefit factor has been calculated (Equations 8 to 16). 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, in this case we focus on Rank 4 (methodology M5), as shown by the instance offered here; for the other ranks, we apply the same process, only changing the equations for each methodology (M3, M4):

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 the lower limits and the vector size must be obtained for each level (Table 8).

3. Equation (17) 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 (21) and (22), respectively (Table 9):

**Table 6.** Identification of patterns suggested for CPV(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 | |

**Table 7.** Identification of patterns suggested for CPV(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 8.** Limits of each vector.

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

**Table 9.** Limits of each vector.

| | Range 4 | |
|---|---|---|
| | ***LI*** | ***LS*** |
| CPV (2) | 2528 | 3792 |
| CPV (3) | 1846 | 2769 |

*For CPV (2):*

$$LI = (1264 * 2) = 2528$$

$$LS = (1264 * 3) = 3792$$

*For CPV (3):*

$$LI = (923 * 2) = 1846$$
$$LS = (923 * 3) = 2769$$

5) Considering pattern length for CPV(2), Equation (8) is applied. In this case, we have obtained a 30% BF. Those NTP exceeding the range are the candidate patterns finally proposed (Table 10):

$$NTP(x) = 3791 * 1.3 = 4928.3$$
$$NTP(x) = 978 * 1.3 = 1271.4$$
$$\cdots$$
$$NTP(x) = 1 * 1.3 = 1.3$$

For CPV(3), Equation (9) is applied, with a 20% BF as a result (Table 11):

$$NTP(x) = 2767 * 1.2 = 3320.4$$

**Table 10.** Patterns suggested for CPV(2).

| 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 | |

**Table 11** Patterns suggested for CPV(3).

| 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 | |

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

## RESULTS

### The selected domain

We have conducted our research on the financial domain, since it is not an area properly explored for computerization. In addition, it is perfectly valid for the construction of ontologies, due to the fact that it is a stable specialized domain plenty of valid multiword terms and acronyms. Apart from this, some tests have been undergone within the cancer subdomain, which has been thoroughly explored by ontology engineers and has a large number of well-defined concepts in the SNOMED[3] terminology database.

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

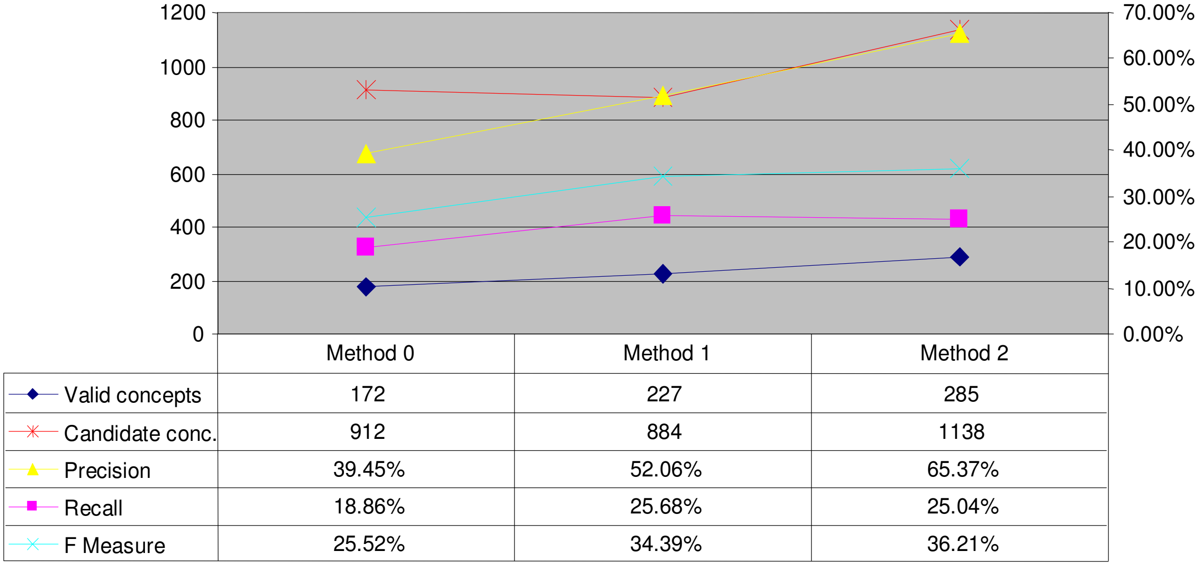
In order to achieve reliable results, we have conducted different procedures modifying certain parameters of the applied heuristics. Specifically, we have performed the test with five different methods, M1 and M2 working from scratch, and M3, M4 and M5 being employed for incremental learning.

### Detection of patterns from scratch in the financial domain

To identify the best method, we have compared the set of patterns proposed by the expert for extracting terms from the corpus with the total amount of valid terms detected through these patterns. As can be seen in Table 12, the method which has obtained the highest number of recommended patterns from scratch is M2. A total amount of 12 was originally proposed by the expert.

From a total of 36 patterns devised by the expert, the

[3] http://www.ihtsdo.org/publications/introducing-snomed-ct/



| | Method 0 | Method 1 | Method 2 |
|---|---|---|---|
| ◆ Valid concepts | 172 | 227 | 285 |
| ✳ Candidate conc. | 912 | 884 | 1138 |
| ▲ Precision | 39.45% | 52.06% | 65.37% |
| ■ Recall | 18.86% | 25.68% | 25.04% |
| × F Measure | 25.52% | 34.39% | 36.21% |

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

tool puts forward 39, from which 12 were originally proposed and 4 were not found in the corpus. Those 16 patterns account for 41.03% of the original patterns proposed by the expert. The amount of patterns obtained by means of M1 was lower than those suggested by the expert. Then, by applying the patterns from M1 and M2 to the financial corpus with a sample of 70%, we obtain the results observed in Figure 2.

The results from method M0 have been obtained through the patterns proposed by the expert, and the results from methods M1 and M2 have been produced with the patterns devised 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 a modest improvement.

## Detection of patterns by means of the incremental method

Like in the previous assessment, in this stage we have compared the set of patterns suggested by the expert for extracting terms from the corpus with the total amount of valid terms detected by means of these patterns. As for incremental learning, it is advisable to obtain a relatively low amount of proposed and original patterns; otherwise, the patterns previously identified by the expert would be obtained again. Furthermore, with every new corpus processed, more patterns will be added; hence the importance of the suitability of the proposed patterns so as not to produce noisy terms. For this reason, we decided that M5 was the best method, since from a total amount of 4 patterns correctly identified by the tool, 2 have been proposed by the expert (Table 12). This is similar to M4, but a higher number of patterns are suggested.

By applying the patterns obtained in methods M3, M4 and M5 to the financial corpus with a sample of 70%, the results shown in Figure 3 have been achieved.

The results from method M0 have been obtained through the patterns proposed by the expert, and the results from method M3, M4 and M5 have been produced with those patterns and with the ones devised by the tool. In this case, the number of valid terms has grown from 270 to 279, which entails an improvement in precision from 61.93 to 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 a strength of the method. Specifically, method M3 produces a higher amount of valid terms, but it also increases the number of candidate terms.

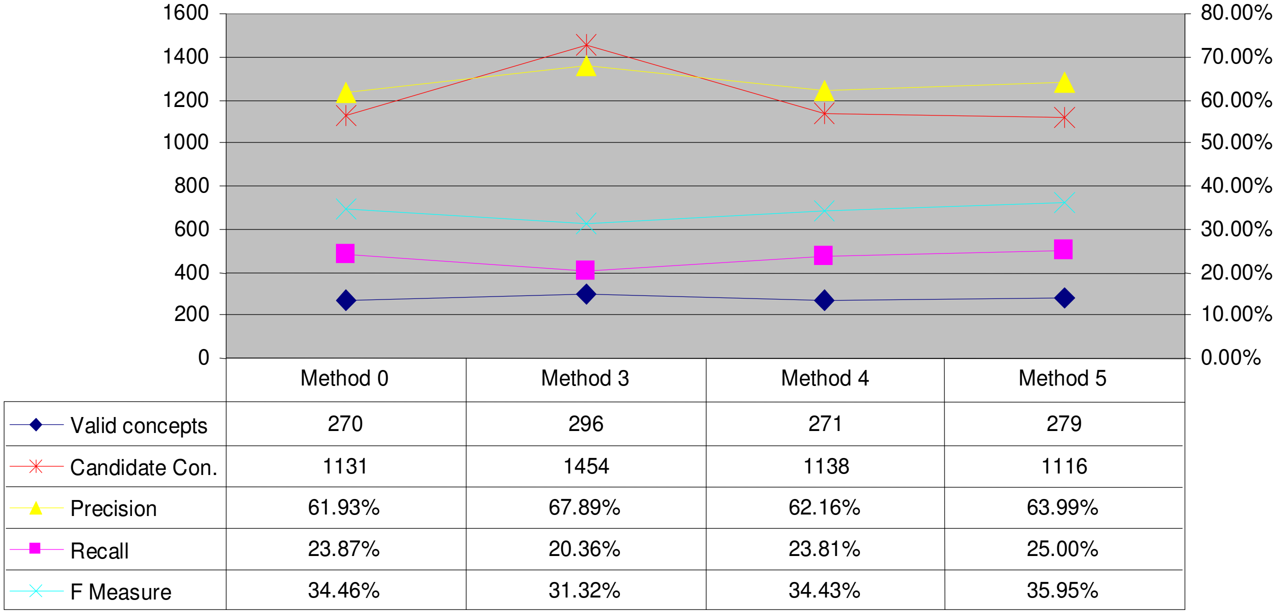
## Detection of patterns from scratch in the cancer domain

As can be seen in Table 13, the method which has obtained the highest number of recommended patterns from scratch is M2. A total amount of 17 was originally proposed by the expert (Table 13).

Accordingly, from a total of 36 patterns proposed by the expert, the tool recommends 185 patterns for method M2, from which 17 were originally suggested by the expert,

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

| | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| Recommended patterns | 27 | 39 | 12 | 6 | 4 |
| Originally 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 3 | Method 4 | Method 5 |
|---|---|---|---|---|
| ◆ Valid concepts | 270 | 296 | 271 | 279 |
| ✳ Candidate Con. | 1131 | 1454 | 1138 | 1116 |
| ▲ Precision | 61.93% | 67.89% | 62.16% | 63.99% |
| ■ Recall | 23.87% | 20.36% | 23.81% | 25.00% |
| × F Measure | 34.46% | 31.32% | 34.43% | 35.95% |

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

having 3 of them not been found in the corpus; this yields a result of 20 patterns, accounting for 55.55% of the original patterns proposed by the expert. As regards method M1, it also offers an improvement, although it is not as significant as method M2. Subsequently, by applying the patterns with M1 and M2 to the cancer corpus with a sample of 70%, we have obtained the results shown in Figure 4.

The results from method M0 have been obtained here through the patterns proposed by the expert, and the results from method M1 and M2 have been produced with the patterns devised by the tool. In this test, there has been an increase in the amount of valid terms from 1521 to 1840, resulting in improvements in precision value –from 34.14 to 41.30%– and in recall value –from 21.18 to 25.70%.

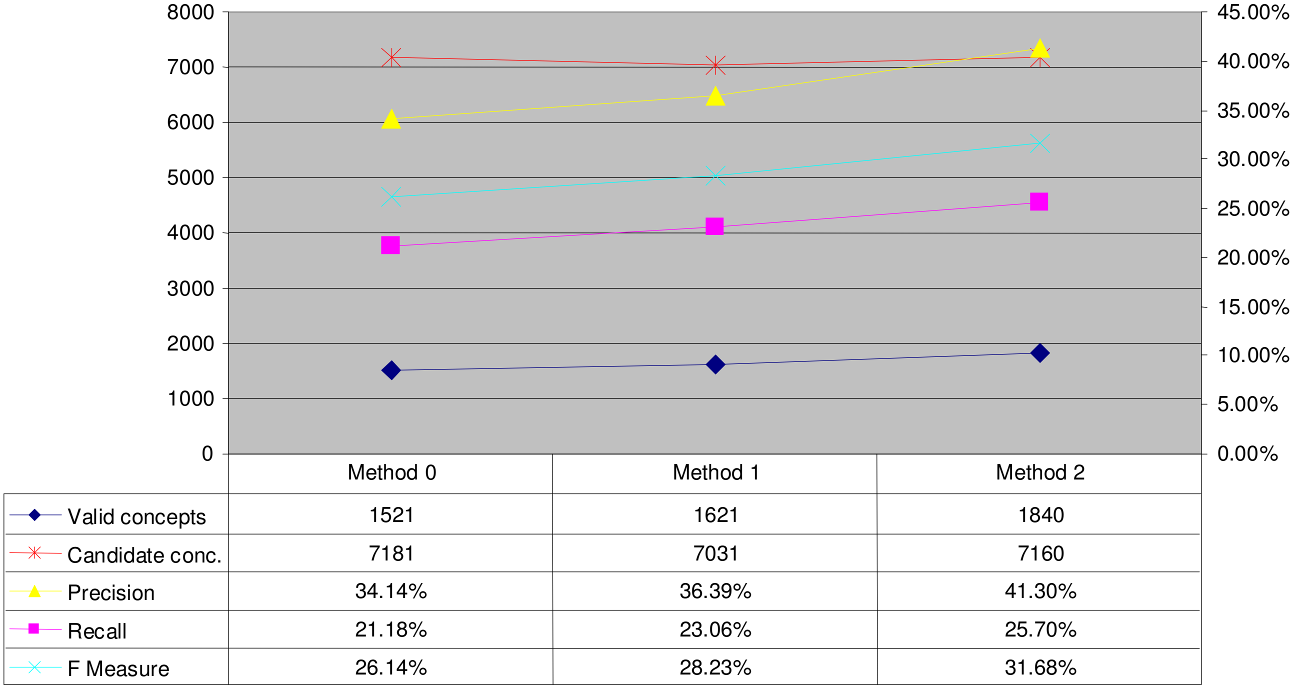
## Detection of patterns by means of the incremental method in the cancer domain

The elements for comparison in this section are the same ones as in the previous test. M5 column in Table 13 shows a total of 5 patterns identified by the tool, 2 of them having been proposed by the expert. M3 column in the same table shows a higher number of patterns, whereas M4 column shows 4 patterns, from which 3 have been already identified. Through the application of the patterns obtained in methods M3, M4 and M5 to the cancer corpus with a sample of 70%, we have obtained the results presented in Figure 5.

The results from method M0 have been obtained through the patterns suggested by the expert, and the results from methods M3, M4 and M5 have been produced

**Table 13.** Guiding table of patterns obtained by the tool and by the expert.

| | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| Recommended patterns | 119 | 185 | 12 | 4 | 5 |
| Originally 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 |



| | Method 0 | Method 1 | Method 2 |
|---|---|---|---|
| ◆ Valid concepts | 1521 | 1621 | 1840 |
| ✳ Candidate conc. | 7181 | 7031 | 7160 |
| ▲ Precision | 34.14% | 36.39% | 41.30% |
| ■ Recall | 21.18% | 23.06% | 25.70% |
| ✕ F Measure | 26.14% | 28.23% | 31.68% |

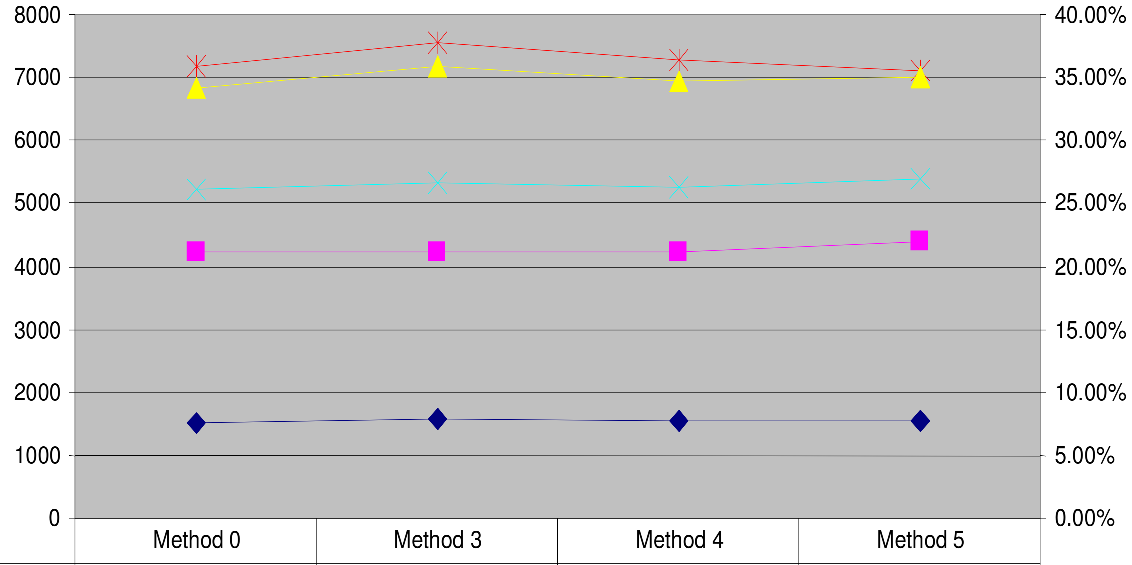
**Figure 4.** Improvements achieved by the method from scratch in the cancer domain.

with those patterns and with the ones identified by the tool. There is an increase in the amount of valid terms from 1521 to 1560, which constitutes an improvement in precision –from 34.14 to 35.02%– as well as in recall –from 21.18 to 21.91%. Method M3 produces a higher amount of valid terms, but it also increases the number of candidate terms.

## DISCUSSION

In this study, two methods for the automatic learning of patterns have been defined, one of them for finding patterns from scratch and the other one for identifying patterns incrementally. Each of them offers certain advantages. The first method allows the user to obtain patterns automatically, without the need of knowing the different pattern combinations, being the only requirement the definition of the unwanted morphological elements at the beginning, at the end and in the middle of the patterns by the user. The second method offers the user the possibility to add new patterns to a pre-existent list from processed corpora, as long as it shares domain with the source corpus for the initial patterns. Furthermore, this method has the same requirement as the first one regarding the unwanted morphological elements.

The difference between the results from methods M1 and M2 is mainly due to the benefit factor added, as can be



| | Method 0 | Method 3 | Method 4 | Method 5 |
|---|---|---|---|---|
| ◆ Valid concepts | 1521 | 1595 | 1543 | 1560 |
| ✳ Candidate conc. | 7181 | 7540 | 7287 | 7120 |
| ▲ Precision | 34.14% | 35.80% | 34.64% | 35.02% |
| ■ Recall | 21.18% | 21.15% | 21.17% | 21.91% |
| ✕ F Measure | 26.14% | 26.59% | 26.28% | 26.95% |

**Figure 5.** Improvements achieved by the incremental method in the cancer domain.

seen in the corresponding columns in Tables 12 and 13. Furthermore, Figures 2 and 4 show how method M2 is substantially better than the method considered as a baseline, M0. As regards the difference between the financial and the cancer domains, the first one provided better results for all the parameters, such as Precision, which is nearly doubled, the correctly identified terms, which experienced a large increase, and candidate terms, which demonstrated just a slight rise in number. This is mainly due to the difficulty inherent in manually obtaining every possible pattern in this complex domain.

As far as methods M3, M4 and M5 are concerned, the major difference lies in the basis for comparison, which was established in an attempt to rule out all those patterns with a low frequency on the corpus, since they generated mainly non-valid terms. For this reason, the amount of recommended patterns obtained in method M2 is higher than in method M1 (Tables 12 and 13). Similarly, in methods M3, M4 and M5 there is a decrease in the proportion of recommended patterns. Nevertheless, the primary aim of obtaining the best recommended patterns is near to be fulfilled, since method M5 has achieved the best results, with the highest number of terms and the best Precision, Recall and F-Measure values.

## 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 the processing of the text, an ordered list with the best patterns found will be obtained. As mentioned above, the method may be used in any domain and it is language independent; 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 necessary. 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 in this line of research are (Abascal-Mena, 2009; Blaschke and Valencia, 2002; Pulido et al., 2007; Sánchez and Moreno, 2004) 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-García et al., 2006) for Spanish; and (Passant, 2007) for French. Other related pieces of research are (Carpuat et al., 2002), in which the building of ontologies across Dutch, Italian, Spanish, German, French, Czech and Estonian are studied; Cimiano's (2006) exploration of domain concepts acquisition; López et al.'s (2010) study of German-Spanish machine translation; 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 and non taxonomic relations (Ochoa et al., 2011a, b). 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 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 (2004) procedure for improving precision in extraction systems. 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; this may produce an excessive amount of candidates to be rejected. This module will be included when the system had processed several corpora from different domains, since many patterns will have been stored for the same domain. 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 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.

## ACKNOWLEDGEMENT

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

## REFERENCES

Abascal-Mena R (2009). Towards a semantic web: Ontology development based on the extraction of semantic concepts from digital documents. In: Mastorakis NE, Mladenov V, Bojkovic Z, Kartalopoulos S, Varonides A (eds) Proceedings of the WSEAES 13th International Conference on Computers, held at Rhodes, Greece. Wisconsin: World Sci. Eng. Acad. Soc., pp. 519-525.

Almuhareb A, Poesio M (2004). Attribute-based and value-based clustering: An evaluation. In: Lin D, Wu D (eds) Proceedings of the 2004 Conference on Empirical Methods and Natural Language, held at Barcelona, Spain. Barcelona: ACL Press, pp. 158-165.

Almuhareb A, Poesio M (2006). MSDA: Word-sense discrimination using context vectors and attributes. In: Brewka G, Coradeschi S, Perini A, Traverso P (eds) Proceedings of the 17th European Conference on Artificial Intelligence, held at Riva del Garda, Italy. Amsterdam: IOS Press, pp. 543-547.

Berland M, Charniak E (1999). Finding parts in very large corpora. In: Dale R, Church K (eds) Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, held at Maryland, USA. Massachusetts: Morgan Kaufmann, pp. 57-64.

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. Inf. Softw. Technol., 51(5): 832-845.

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

Bontas EP, Schlangen D, Schrader T (2005). Creating ontologies for content representation — the OntoSeed suite. LNCS, 3761: 1296-1313.

Bray T, Paoli J, Sperberg-McQueen CM (2008). Extensible markup language (XML) 1.0. W3C recommendation, available at <http://www.w3.org/TR/REC-xml/>

Buitelaar P, Cimiano P, Magnini B (2005). Ontology learning from text: An overview. In Buitelaar et al. (eds) Ontology learning from text: Methods, evaluation and applications, IOS Press, Amsterdam, pp. 1-10.

Byungkyu P, Kyungsook H (2010). An ontology-based search engine for protein-protein interactions. In: Parida L, Myers G (eds) Proceedings of the Eighth Asia-Pacific Bioinformatics Conference, held at Bangalore, India. BMC Bioinformatics, 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: Fellbaum C, Voseen P (eds) Proceedings of the 1st Global WordNet Conference, held at 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: Ananiadou S (ed) Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, held at Prague, Czech Republic, pp. 888-895.

Cömert Ç, Ulutaş D, Akinci H, Kara G (2010). Semantic web services for implementing national spatial data infrastructures. Sci. Res. Essays., 5(7): 685-692.

Dean M, Guus S (2004). OWL Web Ontology Language Reference. W3C recommendation, available at <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: Gravano L, Zhai CX, Herzog O, Evans DA (eds) Proceedings of the 13th ACM Conference on Information and Knowledge Management, held at Washington DC, USA. New York: ACM Press, pp. 652-659.

Ercan T (2010). Hash-based document extraction in corporate mobile devices using ontological architectures. Sci. Res. Essays., 6(2): 440-446.

Fahmi I, Bouma G, van der Plas L (2007). Improving statistical method using known terms for automatic term extraction. Comput. Linguist.. in the Netherlands, 17: 1-8.

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.

Fortuna B, Lavrac N, Velardi P (2008). Advancing topic ontology learning through term extraction. In: Ho TB, Zhou ZH (eds) Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence, held at Hanoi, Vietnam. Berlin: Springer Verlag, pp. 626-635.

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.

Gómez-Pérez A, Ortiz-Rodríguez F, Villazón-Terrazas B (2006). Legal ontologies for the Spanish e-Government. LNCS, 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: Kay M, Peccoud F, Zampolli A, Boitet C (eds) Proceedings of the 14th International Conference on Computational Linguistics, held at Nantes, France. Stroudsburg: ACL Press, pp. 539-545.

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

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

Kauppinen T, Henriksson R, Väätäinen J, Deichstetter C, Hyvönen E (2006). Ontology-based modelling and visualization of cultural spatio-temporal knowledge. In: Hyvönen E, Kauppinen T, Kortela J, Laukkanen M, Raiko T, Viljanen K (eds) New Developments in Artificial Intelligence and the Semantic Web: Proceedings of the 12th Finnish AI Conference, held at Espoo, Finland. Helsinki: Helsinki Univ. Technol., pp. 37-45.

Kawtrakul A, Suktarachan M, Imsombut A (2004). Automatic Thai Ontology Construction and Maintenance System. In: Oltramari A, Paggio P (eds) Ontologies and Lexical Resources in distributed environments: Proceedings of the Workshop OntoLex 2004, held at Lisbon, Portugal.

Khosravi F, Vazifedoost A (2007). Creating a Persian ontology through thesaurus reengineering for organizing the digital library of the National Library of Iran. In: Abdullah A (ed) Proceedings of the International Conference on Libraries, Information and Society, held at Petaling Jaya, Malaysia. 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: Dieng R, Corby O (eds) Proceedings of the EKAW-2000 Workshop Ontologies and Text, held at Juan-Les-Pins, France, pp. 2-6.

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

Lasheras J, Valencia-García R, Fernández-Breis JT, Toval A (2009). Modeling reusable security requirements based on an ontology framework. J. Res. Prac. Inf. Tech., 41(2): 119-133.

Lassila O, Swick RR (1999). Resource Description Framework (RDF) model and syntax specification, available at <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. In: Zafar NA, Mahadevan V (eds) Proceedings of the 2010 International Conference on Intelligence and Information Technology, held at Lahore, Pakistan. Chengdu: Instit. Elect. Elect. Eng. Inc., pp. 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. In: Govil J, Giri D, Park SC, Purohit S, Uddin A (eds) Proceedings of the 3rd International Conference on Computer Modeling and Simulation, held at Mumbai, India. Chengdu: Institute of Electrical and Electronics Engineers, Inc., pp. 104-108.

Ochoa JL, Hernández-Alcaraz ML, Valencia-García, R, Martínez-Béjar R (2011b). A semantic role based ontology learning approach for Spanish texts. In Abraham A, Corchado JM, Rodríguez-González S, de Paz-Santana JF (eds) Proceedings of the 2011 International Symposium on Distributed Computing and Artificial, held at Salamanca, Spain. Berlin: Springer Verlag, pp. 273-280.

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: Gravano L, Zhai CX, Herzog O, Evans DA (eds) Proceedings of the 13th ACM Conference on Information and Knowledge Management, held at Washington DC, USA. New York: ACM Press, pp. 137-145.

Passant A (2007). Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. In: Nicolov N, Glance N, Adar E, Hurst M, Liberman M, Martin J, Salvetti F (eds) Proceedings of the 1st International Conference on Weblogs and Social Media, held at Boulder, Colorado.

Pulido JRG, Flores SBF, Reyes, PD, Díaz RA, Castillo JJC (2007). In the quest of specific-domain ontology components for the semantic web. In: Ritter H, Haschke R (eds) Proceedings of the 6th International Workshop on Self-Organizing Maps, held at Bielefeld, Germany. Berlin: Springer Verlag.

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-García R, Fernández-Breis JT, García-Sanchez F, Vivancos-Vicente PJ, Castejón-Garrido JS, Camón JB, Martínez-Béjar R (2009). Accessing touristic knowledge bases through a natural language interface.
LNCS. 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 Adv. in AI Res. Devel. 113: 11-18.

Sezer EA (2011). Performance assessment of a semantic information retrieval system using stagnant and active images. Sci. Res. Essays., 5(16): 2099-2106.

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-García R, Castellanos-Nieves D, Fernández-Breis J, Vivancos-Vicente P (2006). A methodology for extracting ontological knowledge from Spanish documents. LNCS, 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 Cybern. Paart A-Syst. Hum., 41(1): 121-136.

Vargas-Vera M, Lytras MD (2010). AQUA: A closed-domain question answering system. Inform. Syst. Manag., 27: 217-225.

Vorobiev A, Bekmamedova N (2010). An ontology-driven approach applied to information security. J. Res. Prac. Inf. Tech., 42(1): 61-76.

Yang H, Callan J (2009). Feature Selection for Automatic Taxonomy Induction. In: Allan J, Aslam JA, Sanderson M, Zhai CX, Zobel J (eds) Proceedings of the 32nd Annual ACM SIGIR Conference (SIGIR2009), held at Boston, USA. New York: ACM, pp. 19-23.

Yang C, Liu C, Li J, Yu JX, Wang J (2010a). A query system for XML data stream and its semantics-based buffer reduction. J. Res. Prac. Inf.

Tech., 42(2): 111-128.

Yang YH, Lu Q, Zhao TJ (2010b). A delimiter-based general approach for Chinese term extraction. J. Am. Soc. Inf. Sci. Technol., 61(1): 111-125.

Zhang Z, Iria J, Brewster C, Ciravegna F (2008). A comparative evaluation of term recognition algorithms. In: Calzolari N, Bartolini R (eds) Proceedings of the Sixth International Language Resources and Evaluation, held at Marrakech, Morocco, pp. 2108-2113.

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