Automatising the Learning of Lexical Patterns: an Application to the Enrichment of WordNet by Extracting Semantic Relationships from Wikipedia ⋆

Maria Ruiz-Casado Enrique Alfonseca Pablo Castells

Computer Science Dept., Universidad Autonoma de Madrid, 28049 Madrid, Spain

Abstract

This paper describes an automatic approach to identify lexical patterns that represent semantic relationships between concepts in an on-line encyclopedia. Next, these patterns can be applied to extend existing ontologies or semantic networks with new relations. The experiments have been performed with the Simple English Wikipedia and WordNet 1.7. A new algorithm has been devised for automatically generalising the lexical patterns found in the encyclopedia entries. We have found general patterns for the hyperonymy, hyponymy, holonymy and meronymy relations and, using them, we have extracted more than 2600 new relationships that did not appear in WordNet originally. The precision of these relationships depends on the degree of generality chosen for the patterns and the type of relation, being around 60-70% for the best combinations proposed.

Key words:
Lexical Patterns, Information Extraction, Relation Extraction, Ontology and Thesaurus Acquisition

1 Introduction

Since the first World Wide Web (WWW) project was presented by Tim Berners Lee in 1989, the advances in web technologies have been large. From the

⋆ This work has been sponsored by MEC, project number TIN-2005-06885.
Email addresses: Maria.Ruiz@uam.es (Maria Ruiz-Casado), Enrique.Alfonseca@uam.es (Enrique Alfonseca), Pablo.Castells@uam.es (Pablo Castells).

development of HTML code and HTTP protocol to the most recent advances, the progress in facilities for publishing, retrieving and interacting with web content have strongly stimulated the success of the WWW. Nowadays, the web has at least 4×10^9 static pages (1), and that is taking into account only the so-called *surface web*, which includes only static web pages. Some authors consider that the so-called *deep web*, the web pages that are dynamically generated from information stored in underlying knowledge bases, is even greater than the total volume of printed information existing in the world (2).

Though the exponential growth of the web contents has transformed it into a universal information resource, the huge availability of data hinders sometimes the tasks of searching, retrieving and maintaining it (3; 4), specially when these tasks have to be totally or partially carried out in a manual way.

One of the difficulties that prevents the complete automation of those processes (5) is the fact that the contents in the WWW are presented mainly in natural language. Therefore, web pages are ambiguous and hard to process by a machine.

The Semantic Web (SW) constitutes an initiative to extend the web with machine readable contents and automated services far beyond current capabilities (3). A common practise to make explicit the meaning of web content, and therefore easily processable by a machine, is the annotation of domain-specific terms in the web pages using an ontology. One of the most accepted definitions of an ontology is an agreed explicit specification of a conceptualisation (6). In most of the cases, ontologies are structured as hierarchies of concepts, by means of the relation called hyponymy (*is-a*, class inclusion or subsumption) and its inverse hyperonymy, which arranges the concepts from the most general to the most specific one. Additionally, there may be other relationships, such as meronymy (the part-whole relation) and its inverse holonymy; telicity (*purpose*), or any other which may be of interest, such as *is-author-of*, *is-the-capital-of*, *is-employee-of*, etc. In many cases, ontologies distinguish nodes that represent concepts (classes of things, e.g. *person*) from nodes that represent instances (examples of concepts, e.g. *John*) (7).

Like the web itself, sometimes, these ontologies have to include a high amount of information, or they undergo a rapid evolution. Populating an ontology can result in a very high cost when working in a manual way. In the case of general-purpose ontologies, they usually contain a large quantity of concepts and relationships. On the other hand, many specialised domains also require a very fine-grained ontology, with large vocabularies of specific concepts and many kinds of relationships. Furthermore, the ontologies usually undergo an evolution in which new concepts and relationships are added, and existing ones are removed. Finally, once the ontologies are built, texts have to be annotated using the concepts included in them, adding an extra effort. All these tasks,

required to get semantically annotated contents, can become unfeasible when the goal is to adapt high amounts of preexisting, unannotated contents.

Therefore, it is also highly desirable to automatise or semi-automatise the acquisition of the ontologies. This problem has been object of recent increasing interest, and new approaches for automatic ontology enrichment and population are being developed, which combine resources and techniques from Natural Language Processing (NLP), Information Extraction, Machine Learning and Text Mining (8; 9).

In this paper, we present a procedure for automatically enriching an existing lexical semantic network with new relationships extracted from on-line encyclopedic information. The approach followed is mainly based in the use of lexical patterns that model each type of relationship and natural language processing resources. The semantic network chosen is WordNet (10), given that it is currently used in many applications, although the procedure is general enough to be used with other ontologies. The encyclopedia used is the Wikipedia, a collaborative web-based resource which is being constantly updated by its users. The experiments have been performed with the Simple English version [1], because the vocabulary and syntactic structures found in Simple English are easier to handle by a parser than those found in fully unrestricted text. In addition, the fact that it is written with less supervision than an academic encyclopedia means that the language used is freer, sometimes colloquial, and the techniques that work well here are expected to be easier to port to the web than if we worked with a more structured reference text.

One of the main goals when developing this procedure to extract relationships was to automatise not only the extraction of particular relations (e.g. Jupiter *is part of* the Solar System), but also the extraction of the textual patterns that model them. For instance, *X is part of Y*, and *X is one of the PLURAL-NOUN in Y*, are patterns used to communicate a holonymy relationship. In this way, given an interest in a particular type of relation (e.g. hyponymy, holonymy, etc), patterns can be automatically collected from a textual source and generalised so as represent several different ways in which a human language conveys that relation in a text.

Some existing systems already use similar patterns that have been designed manually, mostly for hyponymy. They are usually modelled through the study of sample sets of diverse sizes and generalised by a human expert. The automation in the creation of the patterns has the advantage of being able to work with corpora of arbitrary size, and being able to extract quickly relationships for rare, domain-dependent relations. The automatic procedure to generalise lexical patterns presented in this work is domain-independent, and may be

[1] http://simple.wikipedia.org/wiki/Main_Page

applied to other type of relationships apart from those included in this study.

The need for automation in the extraction of relations is not limited to ontologies for the Semantic Web. The procedures here described can be also used in other fields that may require ontologies, semantic networks, or other knowledge representation models that include relations between words. Amongst these disciplines we can find Natural Language Processing and Generation, Knowledge Management, e-Commerce, Machine Translation or Information Retrieval (8). The algorithm presented in this work has also been applied to other NLP applications such as Named Entities Recognition (11).

This paper is structured in the following way: Section 2 describes related work; Sections 3 and 4 detail the approach followed, and the evaluation performed; and, finally, Section 5 concludes and points out open lines for future work.

2 Related work

Automatic extraction of information from textual corpora is now a well-known field with many different applications. It is possible to automatise the acquisition of many kinds of information, such as selectional restrictions (12; 13), proper nouns (14), collocations (15), syntactic rules (16; 17; 18), multilingual links (19), or new word senses (20).

In particular, concerning automatic ontology enrichment, we may classify current approaches in the following groups:

- Systems based on distributional properties of words: it consists in studying co-occurrence distributions of terms in order to calculate a semantic distance between the concepts represented by those terms. This distance metric can next be used for conceptual clustering (21; 13), Formal Concept Analysis (22) or for classifying words inside existing ontologies (23; 24; 25; 26). The previous are usually applied to enrich the ontologies with new concepts. On the other hand, (27) learn association rules from dependency relations between terms which, combined with heuristics, are used to extract non-taxonomic relations.
- Systems based on pattern extraction and matching: these rely on lexical or lexicosemantic patterns to discover ontological and non-taxonomic relationships between concepts in unrestricted text. (28; 29; 30) manually define regular expressions to extract hyponymy and part-of relationships. (31) learns such patterns for company merge relationships. (32) quantifies the error rate of a similar approach as 32%. (33) describes a combination of a pattern-based and a distributional-based approach, also for hyperonymy. (34) describe a whole framework which incorporates terminology extraction

and ontology construction and pruning which takes into account, amongst other things, substring relationships for identifying hyperonyms.

Some of these systems use the web for learning the relationships and patterns (35). They have the advantage that the training corpora can be collected easily and automatically, so they are useful in discovering many different relations from text. Several similar approaches have been proposed (36; 37; 38), with various applications: Question-Answering (38), multi-document Named Entity Coreference (39), and generating biographical information (40). (41) applies a similar, with no seed lists, to extract automatically entailment relationships between verbs

- Systems based on dictionary definitions analysis (42; 43; 44; 45), take advantage of the particular structure of dictionaries in order to extract relationships with which to arrange the concepts in an ontology. Concept definitions and glosses have been found very useful, as they are usually concise descriptions of the concepts and include the most salient information about them (46). There are also several works which extract additional relationships from WordNet glosses, by disambiguating the words in the glosses (46; 47; 48; 49).

Concerning the use of lexical and syntactic patterns, and collocations, they have been employed in several applications like Word Sense Disambiguation (50; 51; 52),Question Answering (53; 54), Terminology Extraction (55), Named Entity Recognition (56; 57; 58; 59; 11), Syntactic/Semantic Annotation (60) or Language Characterisation (61).

3 Procedure

The procedure consists in crawling the Simple English version of Wikipedia, collecting all the entries, disambiguating them, and associating each other with relations. The steps carried out, similar to those described in (28; 31), are the following:

(1) *Entry Sense Disambiguation*: This step consists in preprocessing the Wikipedia definitions and associating each Wikipedia entry to its corresponding WordNet synset, so the sense of the entry is explicitly determined.
(2) *Pattern extraction*: For each entry, the definition is processed looking for words that are connected with the entry in Wikipedia by means of a hyperlink. If there is a relation in WordNet between the entry and any of those words, the context is analysed and a pattern is extracted for that relation.
(3) *Pattern generalisation*: In this step, the patterns extracted in the previous step are compared with each other, and those that are found to be similar

are automatically generalised.
(4) *Identification of new relations*: the patterns are applied to discover new relations other than those already present in WordNet.

The following sections detail all the steps in the procedure:

3.1 Entry sense disambiguation

The goal of this step is to mark each entry in the Wikipedia with its corresponding synset in WordNet. To this aim, the entries are downloaded, and they are processed in the following way:

(1) Those web pages which contain more than one definition are divided in separate files.
(2) Most of the HTML tags are removed.
(3) The definitions are processed with a sentence splitter, a part-of-speech-tagger and a stemmer (62).
(4) For each entry, choose the WordNet synset whose sense is nearer according to the definition.

The disambiguation procedure, described in detail in (63), is mainly based on the Vector Space Model and the dot-product similarity metric, co-occurrence information and some heuristics. Approximately one third of the entries in Wikipedia are not found in WordNet, one third appear with just one sense (they are monosemous), and one third have multiple possible senses (they are polysemous). As indicated in (63), the final accuracy obtained is 91%.

The output of this pre-processing step is a list of Wikipedia disambiguated entries.

3.2 Pattern extraction

In the previous step, every entry from the encyclopedia has been disambiguated using WordNet as the sense dictionary. The aim of this step is the extraction of patterns relating two concepts such that they have already been disambiguated and they share a relation in WordNet. The process is the following:

(1) For each term t in the Wikipedia, with a definition d, we select every term f such that there is a hyperlink within d pointing to f. This assures that f's entry also exists in Wikipedia, and its sense has been disambiguated in the previous step.

The reason why we only select the terms which have an entry in Wikipedia is that we have obtained a higher accuracy disambiguating the entry terms than attempting a disambiguation of every word inside the definitions. In this way, we expect the patterns to be much more accurate.

If a particular entry is not found in the disambiguated set, it is ignored, because it means that either the entry is not yet defined in the Wikipedia [2], or it was not found in WordNet and was not disambiguated previously.

(2) Once we have found a hyperlink to other disambiguated entry, the following process is carried out:
   (a) Look up in WordNet relationships between the two terms.
   (b) If any relation is found, collect the sentence where the hyperlink appears (with part-of-speech tags).
   (c) Replace the hyperlink by the keyword TARGET.
   (d) If the entry term appears in the sentence, replace it by the keyword ENTRY.

This work uses WordNet 1.7, in which there are six possible relationships between nouns. The first four, hyperonymy, hyponymy, holonymy and meronymy have been included in this study. Concerning antonymy, this relationship in WordNet does not always refer to the same feature, as sometimes it relates nouns that differ in gender (e.g. *king* and *queen*), and, other times, in a different characteristic (e.g. *software* and *hardware*), so it would be very difficult to find a consistent set of patterns for it. With respect to synonymy, we found that there are very few sentences in Wikipedia that contain two synonyms together, as they are expected to be known by the reader and they are used indistinctly inside the entries.

For illustration, if the entry for *Lisbon* contains the sentence *Lisbon is part of Portugal*, the pattern produced would be the following: `ENTRY is/VBZ part/NN of/IN TARGET`. Note that the words are annotated with part-of-speech tags, using the labels defined for the Penn Treebank(64).

The output of this step consists of as many lists as relationships under study, each list containing patterns that are expected to model each particular relation for diverse pairs of words.

[2] The Wikipedia is continuously refreshing its contents and growing, and some of the links of the definitions fail to bring to another definition.

3.3 Pattern generalisation (I): Edit distance calculation

In order to generalise two patterns, the general idea is to look for the similarities between them, and to remove all those things that they do not have in common.

The procedure used to obtain a similarity metric between two patterns, consists of a slightly modified version of the dynamic programming algorithm for *edit-distance* calculation (65). The *edit distance* between two strings A and B is defined as the minimum number of changes (character insertion, deletion or replacement) that have to be done to the first string in order to obtain the second one. The algorithm can be implemented as filling in a matrix $\mathcal{M}$ with the following procedure:

$$\mathcal{M}[0,0] = 0 \tag{1a}$$
$$\mathcal{M}[i,0] = M[\ \ i-1,0] + 1 \tag{1b}$$
$$\mathcal{M}[0,j] = M[0,j-1] + 1 \tag{1c}$$
$$\begin{aligned}\mathcal{M}[i,j] = min(&\mathcal{M}[i-1,j-1] + d(A[i],B[j]),\\ &\mathcal{M}[i-1,j] + 1,\\ &\mathcal{M}[i,j-1] + 1)\end{aligned} \tag{1d}$$

where $i\epsilon[1...|A|], j\epsilon[1...|B|]$

and

$$d(A[i],B[j]) = \begin{cases} 0 & \text{if } A[i] = B[j] \\ 1 & \text{otherwise} \end{cases}$$

In these equations, $\mathcal{M}[i,j]$ will contain the edit distance between the first i elements of A and the first j elements of B. Equation (1a) indicates that, if A and B are both empty strings, the edit distance should be 0. Equations (1b) and (1c) mean that the edit distance between an empty string, and a string with N symbols must be N. Finally, equation (1d) uses the fact that, in order to obtain a string [3] $A\sigma$ from a string $B\gamma$, we may proceed in three possible choices:

- We may obtain $A\gamma$ from $B\gamma$, and next substitute γ by σ. If γ and σ are the same, no edition will be required.
- We may obtain $A\sigma\gamma$ from $B\gamma$, and next delete γ at the end.
- We may obtain A from $B\gamma$, and next insert the symbol σ in the end.

[3] $A\sigma$ represents the concatenation of string A with character σ.

A: `It is a kind of`
B: `It is nice of`

$\mathcal{M}$	**0**	**1**	**2**	**3**	**4**
0	0	1	2	3	4
1	1	0	1	2	3
2	2	1	0	1	2
3	3	2	1	1	2
4	4	3	2	2	2
5	5	4	3	3	2

$\mathcal{D}$	**0**	**1**	**2**	**3**	**4**
0		I	I	I	I
1	R	E	I	I	I
2	R	R	E	I	I
3	R	R	R	U	I
4	R	R	R	R	U
5	R	R	R	R	E

Fig. 1. Example of the edit distance algorithm. A and B are two word patterns; $\mathcal{M}$ is the matrix in which the edit distance is calculated, and $\mathcal{D}$ is the matrix indicating the choice that produced the minimal distance for each cell in $\mathcal{M}$.

In the end, the value at the rightmost lower position of the matrix is the edit distance between both strings. The same algorithm can be implemented for word patterns, if we consider that the basic element of each pattern is not a character but a whole token.

At the same time, while filling matrix $\mathcal{M}$, it is possible to fill in another matrix $\mathcal{D}$, in which we record which of the choices was selected as minimum in equation (1d). This can be used afterwards in order to have in mind which were the characters that both strings had in common, and in which places it was necessary to add, remove or replace characters. We have used the following four characters:

- `I` means that it is necessary to insert a token, in order to transform the first string into the second one.
- `R` means that it is necessary to remove a token.
- `E` means that the corresponding tokens are equal, so it is not necessary to edit them.
- `U` means that the corresponding tokens are unequal, so it is necessary to replace one by the other.

Figure 1 shows an example for two patterns, A and B, containing respectively 5 and 4 tokens. The first row and the first column in $\mathcal{M}$ would be filled during the initialisation, using Formulae (1b) and (1c). The corresponding cells in matrix $\mathcal{D}$ are filled in the following way: the first row is all filled with `I`'s, indicating that it is necessary to insert tokens to transform an empty string into B; and the first column is all filled with `R`'s indicating that it is necessary to remove tokens to transform A into an empty string. Next, the remaining cells would be filled by the algorithm, looking, at each step, which is the

choice that minimises the edit distance. $\mathcal{M}(5,4)$ has the value 2, indicating the distance between the two complete patterns. For instance, the two editions would be replacing `a` by `nice`, and removing `kind`.

3.4 Pattern generalisation (II): Algorithm

After calculating the edit distance between two patterns A and B, we can use matrix $\mathcal{D}$ to obtain a generalised pattern, which should maintain the common tokens shared by them. The procedure used is the following:

(1) Initialise the generalised pattern G as the empty string.
(2) Start at the last cell of the matrix $\mathcal{D}(i,j)$. In the example, it would be $\mathcal{D}(5,4)$.
(3) While we have not arrived to $\mathcal{D}(0,0)$,
   (a) If ($\mathcal{D}(i,j) =$ `E`), then the two patterns contained the same token A[i]=B[j].
      - Set $G = A[i]\ G$
      - Decrement both i and j.
   (b) If ($\mathcal{D}(i,j) =$ `U`), then the two patterns contained a different token.
      - $G = A[i]|B[j]\ G$, where `|` represents a disjunction of both terms.
      - Decrement both i and j.
   (c) If ($\mathcal{D}(i,j) =$ `R`), then the first pattern contained tokens not present in the other.
      - Set $G = *\ G$, where $*$ represents any sequence of terms.
      - Decrement i.
   (d) If ($\mathcal{D}(i,j) =$ `I`), then the second pattern contained tokens not present in the other.
      - Set $G = *\ G$
      - Decrement j

If the algorithm is followed, the patterns in the example will produced the generalised pattern

```
It is a kind   of
It is nice     of
-------------------
It is a|nice * of
```

This pattern may match phrases such as *It is a kind of*, *It is nice of*, *It is a hyperonym of*, or *It is a type of*. As can be seen, the generalisation of these two rules produces one that can match a wide variety of sentences, and which may be indicating different kinds of relationships between concepts.

The previous example shows that, when two patterns are combined, sometimes the result of the generalisation is far too general, and matches a wide variety of sentences that don't share the same meaning. Therefore, in order to restrict the kinds of patterns that can combine to produce a generalisation, the algorithm has been extended to handle part-of-speech tags. Now, a pattern will be a sequence of terms, and each term will be annotated with a part-of-speech tag, as in the following examples:

(a) `It/PRP is/VBZ a/DT kind/NN of/IN`

(b) `It/PRP is/VBZ nice/JJ of/IN`

(c) `It/PRP is/VBZ the/DT type/NN of/IN`

The edit distance algorithm is modified in the following way: the system only allows replacement actions if the words from the two patterns A and B belong to the same general part-of-speech (nouns, verbs, adjectives, adverbs, etc.). Also, if this is the case, we consider that there is no edit distance between the two patterns. In this way, two patterns that do not differ in the part-of-speech of any of their words will be considered more similar than other pairs of patterns differing in the part-of-speech of one word. The d function, therefore, is redefined as:

$$d(A[i], B[j]) = \begin{cases} 0 & \text{if } PoS(A[i]) = PoS(B[j]) \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

The insertion and deletion actions are defined as before. Therefore, patterns (a) and (b) above would have an edit distance of 2, and the result of their generalisation is:

`It/PRP is/VBZ * of/IN`

On the other hand, the patterns (a) and (c) would have an edit distance of 0, and the result of their generalisation would be the following:

`It/PRP is/VBZ a|the/DT kind|type/NN of/IN`

Once the generalisation procedure has been defined, the following algorithm is used in order to generate the final set of generalised patterns:

(1) Collect all the patterns from the Wikipedia entries in a set $\mathcal{P}$.
(2) For each possible pair of patterns, calculate the edit distance between them.
(3) Take the two patterns with the smallest edit distance, p_i and p_j.

(4) If the edit distance between them exceeds a threshold θ, stop and return the result of all the generalisations.
(5) Otherwise,
    (a) Remove them from $\mathcal{P}$.
    (b) Calculate the more general pattern p_g from them.
    (c) Add p_g to $\mathcal{P}$.
(6) Go back to step 2.

The previous algorithm is repeated for each relationship (e.g. hyponymy or meronymy). The output of the algorithm is the set containing all the rules that have been obtained by combining pairs of original rules. The purpose of the parameter θ is the following: if we set no limit to the algorithm, ultimately all the rules can be generalised to a single generalisation containing just one asterisk, which would match any text. Thus, it is desirable to stop merging rules when the outcome of the merge is too general and would be source of a large quantity of errors. The value of θ was tuned empirically through the tests and evaluation described in Section 4.

The best threshold may depend on the particular application, whether the focus is set on the accuracy or on maximising the number of results. As is shown later, values of θ ranging from 1 to 3 provide accuracies higher or equal to 60%. For higher values of θ, the system tries to generalise very different rules, resulting in rules with many asterisks and few lexical terms.

3.6 Identification of new relations

Finally, given a set of patterns for a particular relation, they can be applied to all the entries in the Wikipedia corpus. Whenever a pattern matches, the target word is identified, and a candidate relationship is produced.

In this step, we took into account the fact that most relations of holonymy and meronymy are either between instances or between concepts, but not between an instance and a concept. For instance, it is correct to say that *Lisbon* is part of *Portugal*, but it does not sound correct to say that *Lisbon* is part of the concept *country*, even though Portugal is a country. Therefore, all the results obtained for holonymy and meronymy in which one of the two concepts related was an instance and the other was a concept were removed from the results. We have used the classification of WordNet synsets as instances or concepts provided by (66).

The output of this step is a list of extracted related pairs of entries for each relation.

4 Evaluation and Results

The algorithm has been evaluated with the whole Simple English Wikipedia entries, as available on September 27, 2005. Each of the entries was disambiguated using the procedure described in (63). An evaluation of 360 entries, performed by two human judges, indicates that the precision of the disambiguation is 92% (87% for polysemous words). The high figure should not come as a surprise, given that, as can be expected, it is an easier problem to disambiguate the title of an encyclopedia entry (for which there exist much relevant data) than a word inside unrestricted text.

The next step consisted in extracting, from each Wikipedia entry e, a list of sentences containing references to other entries f which are related with e inside WordNet. This resulted in 485 sentences for hyponymy, 213 for hyperonymy, 562 for holonymy and 509 for meronymy. When analysing these patterns, however, we found that, both for hyperonymy and meronymy, most of the sentences extracted only contained the name of the entry f (the target of the relationship) with no contextual information around it. The reason was unveiled by examining the web pages:

- In the case of hyponyms and holonyms, it is very common to express the relationship with natural language, with expressions such as *A dog is a mammal*, or *A wheel is part of a car*.
- On the other hand, when describing hyperonyms and meronyms, their hyponyms and holonyms are usually expressed with enumerations, which tend to be formatted as HTML bullet lists. Therefore, the sentence splitter chunks each hyponym and each holonym as belonging to a separate sentence.

All the results in these experiments have been evaluated by hand by two judges. The total inter-judge agreement reached 95%. In order to unify the criteria, in the doubtful cases, similar relations were looked inside WordNet, and the judges tried to apply the same criteria as shown by those examples. The cases in which the judges disagree have not been taking into consideration for calculating the accuracy.

Extraction of hyponymy relations

Table 1 shows the results obtained for several values of the threshold θ that governs when to stop generalising the patterns. With threshold 1, only patterns that have an edit distance less or equal to 1 can be merged. The system output consisted of 19 merged patterns. Note that all the patterns that had not merged with any other are discarded for the result of the generalisation. The 19 patterns extracted a total of 1965 relationships, out of which 681 were already present in WordNet, and the remaining 1284 were evaluated by hand,

Threshold	No. of patterns	Known Relations	New relations	Prec.
1	19	681	1284	72.43%
3	26	951	2162	65.12%
5	23	700	2095	53.13%
7	26	721	2158	52.78%
9	30	729	2217	51.87%

Table 1
Results obtained when extracting hyponymy relationships using different thresholds to stop the generalisation of the rules.

with an overall precision of 72.43%.

As can be seen, as the threshold increases, more rules can be merged, because their edit distance becomes lower than the threshold, so we obtain a larger set of generalised rules. Also, because more rules have been generalised, the number of results increases with threshold 3, and remains rather stable for higher thresholds. On the other hand, as can be expected, the precision drops as we generalise the rules more and more, because we obtain rules with fewer content words that can apply in other contexts not related to hyponymy.

Table 2 describes some of the rules extracted with the threshold 3, which were evaluated separately. The pattern that applied most often is the classical hyponymy copular expression, `ENTRY is a TARGET`, which relates a concept with its hyperonym (rules 7, 8 and 10). There are several versions of this pattern, allowing for extra tokens before and in between, and providing a long list of adjectives that may appear in the definition.

Secondly, there are also patterns which have been extracted because of the characteristics of Wikipedia. For instance, there are several entries about months in the years, and all of them contain a variant of the sentence *XXX is the n-th month in the year*. Therefore, rule 5 shows a pattern extracted from those sentences. Other example is that of colours, and all of which contain the same sentence, *List of colors*, in their definition.

Finally, rules 25 and 26 have been displayed as examples of too specific rules that, because they can only match in very particular contexts, have not been able to identify any hyponymy relationship apart from those that were already in WordNet (rule 15). In the training corpus, every entry containing that sentence is a hyponym of the concept *color*.

Amongst the most common mistakes produced by these rules we may cite the following:

- Errors due to the choice of a modifying PP rather than taking the NP to

No.	Match	Prec.	Rule
1	6	1.0	ENTRY/NN is/VBZ a/DT type/NN of/IN TARGET
2	1	1.0	ENTRY/NNP is/VBZ the/DT */* common|largest/JJS TARGET on/IN Earth|earth/NNP
3	1	1.0	The/DT ENTRY/NNP are|is/VBZ */* big/JJ TARGET in/IN eastern/JJ North/NNP America/NNP
4	1	1.0	ENTRY|Isotopes|Jupiter|Neptune|Saturn|Uranus|Venus/NNS are|is/VBP */* different|eighth|fifth|first|second|seventh|sixth|small/JJ TARGET from|in|of/IN the/DT */* Ocean|Sun|element|sun|year/NN
5	152	0.92	*/* is|was/VBD a|an/DT British|English|alcoholic|non-metal|old/JJ TARGET
6	6	0.83	The/DT ENTRY/NNP is/VBZ a|the/DT TARGET around|for|in|of/IN the/DT */* Party|Pole|States|Yorkshire|tree/NNP
7	574	0.79	ENTRY/NN is/VBZ a/DT TARGET
8	579	0.74	*/* ENTRY/NN is/VBZ a|an/DT TARGET
9	29	0.66	*/* ENTRY/NN is/VBZ a/DT */* branch|drink|piece|sheet|type/NN of/IN TARGET
10	639	0.49	ENTRY/NNP is/VBZ a|the/DT TARGET for|in|of|that/IN */*
11	7	0.43	ENTRY/NN came|is/VBZ */* a|an/DT TARGET drink|family/NN
12	36	0.42	TARGET of/IN the/DT Year/NN
13	35	0.17	Earth/NNP 's/POS TARGET
14	78	0.17	*/* is|use/VBP coins|part/NNS as|of/IN TARGET
15	18	0.0	TARGET List/NN of/IN colors/NNS
(9 more rules)			
25	0	n/a	The/DT language/NN called/VBD */* is/VBZ one/CD of/IN the/DT language|languages/NNS that/WDT came/VBD from/IN the/DT TARGET language/NN
26	0	n/a	A|An|The/DT ENTRY/NNP is/VBZ a|the/DT TARGET that/WDT connects|has|helps|lets/VBZ */* computers|letter|plants|run/NNS */*

Table 2
Some of the rules obtained for the relation of hyponymy (threshold 3). Columns indicate the number of the rules, the new results produced by each rule, its precision and the text of the rule.

which it modifies. For example, from the sentence *the man with the telescope is the leader*, the word *telescope* would be chosen as hyponym of *leader*. To correct these errors, the patterns should also include syntactic information.

- Invalid information obtained from erroneous sentences, such as *the U.K. is a communist republic*. The Wikipedia is a supervised Encyclopedia, but the erroneous information introduced by the authors may persist for a few days

Threshold	No. of patterns	Known Relations	New relations	Precision
1	1	1	0	n/a
3	4	1	0	n/a
5	5	2	16	50%
7	9	9	28	32.14%
9	10	15	77	27.27%

Table 3
Results obtained when extracting hyperonymy relationships using different thresholds to stop the generalisation of the rules.

before it is noticed and removed.

- Typographic errors, e.g. *Swedish* is classified as a hyponym of *launge* from the text:

  > Swedish is a person or a object that comes from the country Sweden. It's like English and England. It can also be the *launge* that is spoken in Sweden

Some of the new words that have been classified in WordNet correctly are *Rochdale F. C.* as a *club*, *Ijtihad* as a *war*, *Bambuco* as a *music*, and *Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch* as a *village*. Some existing words for which new relationships have been added are *Paris* and *Athens*, as the capital towns in France and Greece, which appear in WordNet as hyponyms of *capital* and now have a new hyponymy relationship to *city*.

Extraction of hyperonymy relations

Concerning hyperonymy, as commented before, it is usually expressed in the Wikipedia with enumerations, that are not handled properly by the pattern-matching procedure. Consequently, there were very few patterns to use, and those available were very specific. Table 3 shows the results of the evaluation for five threshold values. As can be seen, with thresholds 1 and 3, the obtained patterns can just identify one already-known relationship. Using thresholds 5, 7 and 9, the system produced several new results, but with a low precision. It will be necessary to represent enumerations in the patterns in order to apply them to extract hyperonymy relationships.

Extraction of holonymy relations

The case of holonymy is similar to that of hyponymy. The results are shown in Table 4. As can be seen, as we increase the threshold on the edit distance so that two rules are allowed to be merged, we obtain more general rules that can extract more results, but with a lower precision. Depending on the desired

Threshold	No. of patterns	Known Relations	New relations	Precision
1	19	134	79	70.89%
3	22	207	336	59.82%
5	14	304	1746	50.63%
7	15	307	2979	33.43%
9	21	313	3300	31.67%

Table 4
Results obtained when extracting holonymy relationships using different thresholds to stop the generalisation of the rules.

accuracy and number of results a different threshold can be chosen.

Table 5 shows some of the rules for holonymy. Most of the *member part-of* and *substance part-of* relations were rightly extracted by the first few rules in the table, which match sentences such as *X is in Y* or *X is a part of Y*. However, they also extracted some wrong relations.

Interestingly, most of the patterns focused on locations, as we can see in rules 1, 3, 5, 6, 7, 8, 9, 11, 12, 13 and 14. A possible explanation is the large number of entries describing villages, cities and counties in the Simple English Wikipedia.

In the case of holonymy, we have identified several common errors:

- An important source of errors was the lack of a multiword expression recogniser. Many of the part-of relations that appear in Wikipedia are relations between instances, and a large portion of them have multi-word names. For instance, the application of the set of patterns to the sentence

  Oahu is the third largest of the Hawaiian Islands

  returns the relation *Oahu is part of Islands*, because *Hawaiian_Islands* has not been previously identified as a multi-word named entity. Other erroneous examples are: (a) *kidney* as part of *system*, and not *urinary system*; and (b) *Jan Peter Balkenende* as part of *party* rather than *Christian CDA party*.
- Other errors were due to orthographic errors in the Wikipedia entry (e.g. *Lourve* instead of *Louvre*) and relations of holonymy which held in the past, but which are not true by now, such as *New York City is part of Holland* or *Caribbean Sea is part of Spain*.
- Finally, some errors are also due to the polysemy of the words in the pattern. For instance, the following pattern,

  `ENTRY/NNP is/VBZ a|the/DT capital|city|country|province|state/NN in|of/IN TARGET`

  extracts erroneously, from the following sentences:

  (1) Plasma is a state of matter when the bonds between molecular particles are broken and subatomic particles are all lumped in together.

No.	Match	Prec.	Rule
1	1	1.0	ENTRY/NNP */* the/DT */* capital|city/NN of/IN TARGET ,/, */* Japan|city/NN
2	2	1.0	Some/DT TARGET also/RB have/VBP hair/NN like/IN this/DT ,/, and/CC people/NNS sometimes/RB also/RB call/VB this/DT hair/NN a/DT ENTRY/NN
3	32	0.75	ENTRY/NNP is/VBZ a/DT city|province/NN in/IN TARGET
4	331	0.73	ENTRY/NNP is/VBZ a|an|the/DT */* in|of/IN the/DT TARGET
5	104	0.60	*/* is|makes|means|was/VBZ */* a|the/DT */* States|corner|countries|country|layer|part|parts|planet/NNS in|of|that/IN */* a|the/DT TARGET
6	18	0.56	*/* Countries|city|country|follower/NNS in|of/IN */* East|Southeast|Southeastern|West|faith|world/NN TARGET
7	851	0.45	*/* ENTRY|South|capital|city|continent|country| county|fact|state/NN */* as|in|of/IN TARGET
8	396	0.41	*/* Things|city|member|north|part|planets|state/NNS in|of/IN the/DT TARGET
9	5	0.4	*/* ENTRY/NNP is|was/VBZ a/DT */* country|part|river/NN in|of/IN */* eastern|north|northern/JJ TARGET
10	5	0.4	It/PRP is/VBZ part/NN of/IN the/DT TARGET
11	1	0.0	It/PRP is/VBZ in/IN central|southwest/JJ TARGET
12	0	n/a	ENTRY/NNP is/VBZ a|the/DT capital|country/NN */* between|of/IN TARGET and/CC */* Europe|city/NNP
13	0	n/a	The/DT */* Kingdom|Republic|part/NN of/IN */* ENTRY/NNP is/VBZ */* a|the/DT country|middle/NN in|of/IN the/DT continent|middle|southwest/NN of/IN TARGET
14	0	n/a	ENTRY/NNP (/(Cornish|German|Icelandic|Welsh/NNP */* Bayern|Caerdydd|Kernow|sland/NNP)/) is/VBZ a|the/DT */* city|country|county|part/NN in|of/IN TARGET

Table 5
Rules obtained for the relation of holonymy (threshold 5), ordered by precision. Columns indicate the rules' number, number of new results found, precision and pattern.

(2) Weather is the state of the atmosphere at any given time

the relationships between *Plasma* and *matter*, and between *weather* and *atmosphere*. This error stems from the fact that *state* is not used with the sense of territorial division, but with the senses, respectively, of *state of matter* and *the way something is with respect to its main attributes*.

Extraction of meronymy relations

Concerning the last relationship studied, meronymy, even though it is also represented quite often with enumerations in the Wikipedia, the results are

Threshold	No. of patterns	Known Relations	New relations	Precision
1	8	32	10	100%
3	10	74	124	62.90%
5	10	78	147	56.46%
7	14	84	473	40.59%
9	18	95	494	40.89%

Table 6
Results obtained when extracting meronymy relationships using different thresholds to stop the generalisation of the rules.

No.	Match	Prec.	Rule
1	2	1.0	`TARGET (/( Bayern|Thringen/NNP )/)`
2	1	1.0	`TARGET (/( city-state/JJ )/)`
3	1	1.0	`TARGET is/VBZ in/IN the/DT north|south/NN`
4	2	1.0	`A|An/DT ENTRY/NN is/VBZ a/DT unit/NN of/IN time/NN ,/, it/PRP is/VBZ equal/JJ to/TO 60/CD TARGET`
5	93	0.74	`*/* ENTRY|capital|city/NNP is/VBZ TARGET`
6	6	0.17	`Winnipeg|mangrove|volcano/NNP */* in/IN TARGET`
7	19	0.11	`Cape|Sun/NNP Horn|Moon/NNP */* Chile|Comets/NNPS TARGET`
8	0	n/a	`ENTRY|Ireland/NNP contains|is/VBZ a|the/DT Republic|gas/NNP */* of/IN */* Ireland|nitrogen/NNP and/CC TARGET`
9	0	n/a	`The/KT capital/NN */* city/NN in|of/IN ENTRY|Georgia/NNP is/VBZ TARGET`
10	0	n/a	`Calgary|Edmonton|Montreal|Vancouver/NNP ,/, in/IN TARGET ;/,`

Table 7
Rules obtained for the relation of meronymy (threshold 3), ordered by precision. Columns indicate the rules' number, number of new results found, precision and pattern.

rather better than those of hyperonymy. The results are shown in Table 6. The number of results is lower than the case of hyponymy and holonymy, but the accuracy, for the different threshold values, follows a similar behaviour. The accuracy is very high with precision 1 (although the number of new results is very low), and decreases as the threshold increases.

Table 7 shows the patterns obtained with threshold 3. In contrast to the relationships of hyponymy and holonymy, in which most of the patterns obtained were according to our intuition (variations of *An X is a Y; X is part of Y*), in this case that behaviour is not so clear. That is due to the fact that most of these patterns will only apply correctly inside Encyclopedic text, but do not

indicate meronymy in general texts. To illustrate this point, let us consider, pattern 3, that infers that X is a part of the defined entry if the sentence

(3) X is in the north

appears inside the entry. This inference will be probably wrong if we are not processing an encyclopedia.

5 Conclusions and future work

This work addresses the problem of automatically identifying semantic relationships in free text. Some of the conclusions that can be drawn from this work are the following:

- A new algorithm for generalising lexical patterns has been described, implemented and evaluated. It is based on the edit distance algorithm, which has been modified to take into account the part-of-speech tags of the words. This algorithm is fully automatic, as it requires no human supervision.
- The set of patterns which has been found automatically from the Wikipedia entries is able to extract new relations from text for each of the four relationships: hyperonymy, hyponymy, meronymy and holonymy. More than 2600 new relationships have been provided using thresholds from 3 to 5 in the generalisation step.
- The precision of the generated patterns is similar to that of patterns written *by hand* (although they are not comparable, as the experimental settings differ). The kind of hyponymy lexico-syntactic patterns as described by (28) were evaluated, in different settings, by (32) and (22), who report a precision of 0.68 and 0.39, respectively. (30) reports a 0.55 accuracy for a set of patterns that identify holonyms. Only (31) reports much higher accuracies (0.72, 0.92 and 0.93), when identifying relationships of merging between companies.

This work opens the following research lines:

- To extract other kinds of relations, such as *location, instrument, telic* or *author*. Newer versions of WordNet (2.1) include more specific relations than the version used for this work, like differentiating instance-of and subconcept-of hyponymy relationships. WordNet also includes relationships between verbs, adjectives and adverbs that have not been studied in the present experiments.
- To generalise the experiment to other ontologies and encyclopedias, and to apply it to fully unrestricted texts collected from the web. Using a multilingual ontology, this procedure can be applied to other languages, provided a

suitable base ontology, ontology and Linguistic Processing tools. WordNet, for instance, can be found in several languages [4]. There are, also, several non-English Wikipedias available [5]. We would like to check whether this approach works equally for highly inflectional languages, or for those with free word order.

- To extend the formalism used to represent the patterns, so they can encode syntactic features as well. A very straightforward extension of the pattern generalisation procedure consists in processing the encyclopedic entries with shallow parsing to detect phrasal nouns and simple subject-verb-object relations, and contemplate this information in the generalisation algorithm, in a way similar to what was done with part-of-speech. The possibility of using deeper syntactic analysis can also be studied.

References

[1] R. Baeza-Yates, Excavando la web, El profesional de la información 13 (1) (2004) 4–10.

[2] P. Castells, La Web Semantica. In C. Bravo and M. A. Redondo (Eds.), Sistemas Interactivos y Colaborativos en la Web, Ediciones de la Universidad de Castilla-La mancha, 2003, pp. 195–212.

[3] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web - a new form of web content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American 284 (5) (2001) 34–43.

[4] D. Fensel, C. Bussler, Y. Ding, V. Kartseva, M. Klein, M. Korotkiy, B. Omelayenko, R. Siebes, Semantic web application areas, in: 7th. International Workshop on Application of Natural Language to Information Systems, Stockholm, Sweden, 2002.

[5] Y. Ding, D. Fensel, M. C. A. Klein, B. Omelayenko, The semantic web: yet another hip?, Data Knowledge Engineering 41 (2-3) (2002) 205–227.

[6] T. R. Gruber, A translation approach to portable ontologies, Knowledge Acquisition 5 (2) (1993) 199–220.

[7] W. Degen, B. Heller, H. Herre, B. Smith, Gol: Towards an axiomatized upper-level ontology, in: Proceedings of the International Conference on Formal Ontology in Information Systems, FOIS-2001, 2001.

[8] A. Gómez-Pérez, D. M. Macho, E. Alfonseca, R. N. nez, I. Blascoe, S. Staab, O. Corcho, Y. Ding, J. Paralic, R. Troncy, Ontoweb deliverable 1.5: A survey of ontology learning methods and techniques (2003).

[9] A. Maedche, S. Staab, Ontology learning for the semantic web, IEEE Intelligent systems 16 (2).

[10] G. A. Miller, WordNet: A lexical database for English, Communications of the ACM 38 (11) (1995) 39–41.

[4] http://www.globalwordnet.org/gwa/wordnet_table.htm

[5] http://en.wikipedia.org/wiki/Main_Page

[11] E. Alfonseca, M. Ruiz-Casado, Learning sure-fire rules for named entities recognition, in: Proceedings of the International Workshop in Text Mining Research, Practice and Opportunities, in conjunction with RANLP conference, Borovets, Bulgary, 2005.

[12] P. Resnik, Selection and Information: A Class-Based Approach to Lexical Relationships, Ph.D. thesis. Dept. of Computer and Information Science, Univ. of Pennsylvania, 1993.

[13] D. Faure, C. Nédellec, A corpus-based conceptual clustering method for verb frames and ontology acquisition, in: LREC workshop on Adapting lexical and corpus resources to sublanguages and applications, Granada, Spain, 1998.

[14] A. Mikheev, C. Grover, M. Moens, Description of the ltg system used for muc-7, in: Proceedings of 7th Message Understanding Conference (MUC-7), 1998.

[15] K. Church, W. Gale, P. Hanks, D. Hindle, Using Statistics in Lexical Analysis. In U. Zernik (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1991, Ch. 6, pp. 115–164.

[16] J. Hockenmaier, G. Bierner, J. Baldridge, Providing robustness for a ccg system, in: Proceedings of the Workshop on Linguistic Theory and Grammar Implementation, Hong Kong, 2000.

[17] T. Briscoe, J. Carroll, Automatic extraction of subcategorization from corpora, in: Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97), Washington DC, USA, 1997.

[18] F. Xia, Extracting tree adjoining grammars from bracketed corpora, in: In Fifth Natural Language Processing Pacific Rim Symposium (NLPRS-99), Beijing, China, 1999.

[19] I. Dagan, A. Itai, U. Schwall, Two languages are more informative than one, in: proceedings of ACL-91, Berkeley, California, 1991, pp. 130–137.

[20] R. Basili, R. Catizone, M. T. Pazienzia, M. Stevenson, P. Velardi, M. Vindigni, Y. Wilks, An empirical approach to lexical tuning, in: Proceedings of the Workshop Adapting Lexical and Corpus Resources to Sublanguages and Applications, LREC First International Conference on Language Resources and Evaluation, Granada, Spain, 1998.

[21] L. Lee, Similarity-Based Approaches to Natural Language Processing, Ph.D. thesis. Harvard University Technical Report TR-11-97, 1997.

[22] P. Cimiano, S. Staab, Clustering concept hierarchies from text, in: Proceedings of LREC-2004, 2004.

[23] P. M. Hastings, Automatic acquisition of word meaning from context, University of Michigan, Ph. D. Dissertation, 1994.

[24] U. Hahn, K. Schnattinger, Towards text knowledge engineering, in: AAAI/IAAI, 1998, pp. 524–531.
URL `citeseer.nj.nec.com/43410.html`

[25] V. Pekar, S. Staab, Word classification based on combined measures of distributional and semantic similarity, in: Proceedings of Research Notes

of the 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, 2003.

[26] E. Alfonseca, S. Manandhar, Extending a lexical ontology by a combination of distributional semantics signatures, in: Knowledge Engineering and Knowledge Management, Vol. 2473 of Lecture Notes in Artificial Intelligence, Springer Verlag, 2002, pp. 1–7.

[27] A. Maedche, S. Staab, Discovering conceptual relations from text, in: Proceedings of the 14th European Conference on Artifial Intelligence, 2000.

[28] M. A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proceedings of COLING-92, Nantes, France, 1992.

[29] M. A. Hearst, Automated Discovery of WordNet Relations. In Christiane Fellbaum (Ed.) WordNet: An Electronic Lexical Database, MIT Press, 1998, pp. 132–152.

[30] M. Berland, E. Charniak, Finding parts in very large corpora, in: Proceedings of ACL-99, 1999.

[31] M. Finkelstein-Landau, E. Morin, Extracting semantic relationships between terms: supervised vs. unsupervised methods, in: Proceedings of the International Workshop on Ontologial Engineering on the Global Information Infrastructure, 1999.

[32] J. Kietz, A. Maedche, R. Volz, A method for semi-automatic ontology acquisition from a corporate intranet, in: Workshop "Ontologies and text", co-located with EKAW'2000, Juan-les-Pins, French Riviera, 2000.

[33] E. Alfonseca, S. Manandhar, Improving an ontology refinement method with hyponymy patterns, in: Language Resources and Evaluation (LREC-2002), Las Palmas, 2002.

[34] R. Navigli, P. Velardi, Learning domain ontologies from document warehouses and dedicated websites, Computational Linguistics 30 (2).

[35] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates, Unsupervised named entity extraction from the web: An experimental study, Artificial Intelligence 165 (1) (2005) 91–134.

[36] S. Brin, Extracting patterns and relations from the World Wide Web, in: Proceedings of the WebDB Workshop at EDBT'98, 1998.

[37] E. Agichtein, L. Gravano, Snowball: Extracting relations from large plain-text collections, in: Proceedings of ICDL, 2000, pp. 85–94.

[38] D. Ravichandran, E. Hovy, Learning surface text patterns for a question answering system, in: Proceedings of ACL-2002, 2002, pp. 41–47.

[39] G. S. Mann, D. Yarowsky, Unsupervised personal name disambiguation, in: CoNLL-2003, 2003.

[40] G. S. Mann, D. Yarowsky, Multi-field information extraction and cross-document fusion, in: Proceedings of ACL 2005, 2005.

[41] I. Szpektor, H. Tanev, I. Dagan, B. Coppola, Scaling web-based acquisition of entailment relations, in: Proceedings of EMNLP 2004, 2004.

[42] Y. Wilks, D. C. Fass, C. M. Guo, J. E. McDonald, T. Plate, B. M. Slator,

Providing machine tractable dictionary tools, Journal of Computers and Translation 2.

[43] G. Rigau, Automatic Acquisition of Lexical Knowledge from MRDs, PhD Thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, 1998.

[44] S. D. Richardson, W. B. Dolan, L. Vanderwende, MindNet: acquiring and structuring semantic information from text, in: Proceedings of COLING-ACL'98, Vol. 2, Montreal, Canada, 1998, pp. 1098–1102.

[45] W. Dolan, L. Vanderwende, S. D. Richardson, Automatically deriving structured knowledge bases rfon on-line dictionaries, in: PACLING 93 Pacific Association for Computational Linguistics, 1993, pp. 5–14.

[46] S. Harabagiu, D. I. Moldovan, Knowledge processing on an extended wordnet, in: WordNet: An Electronic Lexical Database, MIT Press, 1998, pp. 379–405.

[47] S. Harabagiu, G. Miller, D. Moldovan, Wordnet 2 - a morphologically and semantically enhanced resource, in: Proc. of the SIGLEX Workshop on Multilingual Lexicons, ACL Annual Meeting, University of Maryland, 1999.

[48] A. Novischi, Accurate semantic annotation via pattern matching, in: Proceedings of FLAIRS-2002, 2002.

[49] M. DeBoni, S. Manandhar, Automated discovery of telic relations for wordnet, in: Poceedings of the First International Conference on General WordNet, Mysore, India, 2002.

[50] M. A. Hearst, Noun homograph disambiguation using local context in large text corpora, in: Proceedings of the 7th Annual Conference of the Centre for the New OED and Text Research: Using Corpora, Oxford, UK, 1991, pp. 1–22.

[51] D. Yarowsky, One sense per collocation, in: Proceedings of ARPA Human Language Technology Workshop, Princeton, NJ, 1993, pp. 266–271.

[52] J. Batali, The negotiation and acquisition of recursive grammars as a result of competition among exemplars (1998).

[53] J. V, J. Mur, M. de Rijke, Information extraction for question answering: Improving recall through syntactic patterns, in: Proceedings of COLING-04, 2004.

[54] M. Soubbotin, S. Soubbotin, Use of Patterns for Detection of Answer Strings: A systematic Approach In Vorhees and Buckland (Ed.) [7], 2002.

[55] A. P. nas, F. Verdejo, J. Gonzalo, Corpus-based terminology extraction applied to information access, in: Proceedings of Corpus Linguistics 2001, 2001.

[56] M. E. Califf, Relational Learning Techniques for Natural Language Extraction, Ph.D. thesis. University of Texas at Austin, 1998.

[57] S. Soderland, Learning information extraction rules for semi-structured and free text, Machine Learning 34 (1–3) (1999) 233–272.

[58] D. Freitag, Machine learning for information extraction in informal domains, Machine Learning 34.

[59] M. Arevalo, M. Civit, M. Marti, Mice: A module for named entity recognition and classification, International Journal of Corpus Linguistics 9 (2004) 53–68.
[60] M. Santini, Building on syntactic annotation: Labelling of subordinate clauses, in: Proceedings of the Workshop on Exploring Syntactically Annotated Corpora, Corpus Linguistics, Birmingham, UK, 2005.
[61] S. Bonzi, Syntactic patterns in scientic sublanguages: A study of four disciplines, Journal of The American Society for Information Science 41.
[62] E. Alfonseca, Wraetlic user guide version 1.0 (2003).
[63] M. Ruiz-Casado, E. Alfonseca, P. Castells, Automatic assignment of wikipedia encyclopedic entries to wordnet synsets, in: press, 2005.
[64] M. P. Marcus, B. Santorini, M. A. Marcinkiewicz, Building a large annotated corpus of english: the penn treebank, Computational Linguistics 19 (2) (1993) 313–330.
[65] R. Wagner, M. Fischer, The string-to-string correction problem, Journal of Assoc. Comput. Mach. 21.
[66] E. Alfonseca, S. Manandhar, Distinguishing instances and concepts in wordnet, in: Poceedings of the First International Conference on General WordNet, Mysore, India, 2002.

Maria Ruiz-Casado is a Ph.D. student at the Computer Engineering Department, Universidad Autonoma de Madrid. Her thesis topic is Ontology Learning and Population from free text.

Enrique Alfonseca is Assistant Lecturer at the Universidad Autonoma de Madrid since 2001. His main interests are focused on several Natural Language Processing topics such as Information Extraction and summarisation, and their application to e-learning.

Pablo Castells is Associate Professor at the Universidad Autonoma de Madrid since 1999, where he leads and participates in several national and international projects in the areas of the Semantic-Based Knowledge Systems. His current research interests are focused on Information Retrieval, Personalisation technologies, and Semantic Web Services.