Using Lexical Patterns for Extracting Hyponyms from the Web

Rosa M. Ortega-Mendoza, Luis Villaseñor-Pineda and Manuel Montes-y-Gómez

Laboratorio de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Óptica y Electrónica, México.
{rmortega, villasen, mmontesg}@ccc.inaoep.mx

Abstract. This paper describes a method for extracting hyponyms from free text. In particular it explores two main matters. On the one hand, the possibility of reaching favorable results using only lexical extraction patterns. On the other hand, the usefulness of measuring the instance's confidences based on the pattern's confidences, and vice versa. Experimental results are encouraging because they show that the proposed method can be a practical high-precision approach for extracting hyponyms for a given set of concepts.

1 Introduction

Linguistic resources such as dictionaries, gazetteers and ontologies have a broad range of applications in computational linguistics and automatic text processing [4]. This kind of resources provides significant knowledge about languages, but they are expensive to build, maintain and extend.

At present, most linguistic resources are manually constructed. They contain high precision entries but have very limited coverage. As a result, their usefulness is still restricted to certain domains or specific applications. In order to overcome this problem, recently many researchers have been working on semiautomatic methods for their construction. In particular, there is a special interest in the extraction of synonyms, antonyms and hyponyms from free text documents [6, 7, 5, 8, 9, 10].

This paper focuses on the extraction of hyponyms (*is-a* relations) from free text. Specifically, it proposes a pattern-based method for automatically acquiring hyponyms from the Web. This method mainly differs from previous approaches [5, 8, 9, 10] in that it only considers lexical information. That is, whereas previous methods make use of lexico-syntactic patterns, the proposed approach exclusively employs patterns expressed at lexical level.

Working at lexical level makes the method easily adapted to different languages, but also brings out some additional challenges. For instance, it is necessary to take into consideration a large number of patterns in order to compensate their poor generalization degree. The proposed method confronts this requirement by applying, on the one hand, a text mining technique that allows acquiring many lexical patterns from the Web, and on the other hand, an iterative process for evaluating the confi-

Work done under partial support of CONACYT (project grants 43990 and 61335)

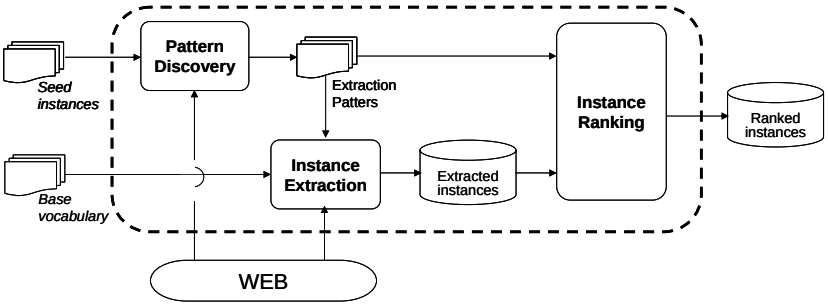

Figure 1. General scheme of the proposed method

dence of the discovered instances (pairs of hyponym-hypernym) to belong to the target relation. This process is supported on the assumption that pertinent instances are extracted by different patterns, and that valuable patterns allow extracting several pertinent instances. This process adopts some ideas proposed elsewhere [9], but modifies the computation of the initial weights of the acquired patterns. This modification allows assigning major importance to those patterns showing a good balance between precision and recall.

The following sections describe the proposed method and present some experimental results on the extraction of hyponyms related to a given base vocabulary.

2 Method at a Glance

Figure 1 shows the general scheme of the proposed method. It consists of three main modules: pattern discovery, instance extraction and instance ranking. The following paragraphs briefly describe the purpose and functionality of each module.

Pattern discovery. This module focuses on the discovery of a set of lexical extraction patterns from the Web. Its objective is to capture most of the writing conventions used to introduce a hyponym relation between two words.

This module adopts the method described in [2]. It mainly uses a small set of seed instances (pairs of hyponym-hypernym such as *apple–fruit*) to collect from the Web an extended set of usage examples of the hyponym relation. Then, it applies a text mining method [3] on the collected examples in order to obtain all maximal frequent word sequences[1]. These sequences express lexical patters highly related to the hyponym relation. Finally, it retains only the patterns satisfying the following regular expressions:

<left-frontier-string> HYPONYM <center-string> HYPERNYM

HYPERNYM <center-string> HYPONYM < right-frontier-string>

[1] A maximal frequent word sequence is a sequence of words that occurs more than a predefined threshold and that is not a subsequence of another frequent sequence.

As it can be seen, the Web is a very valuable resource for this step; however, it restricts the method to applications that do not require constructing the hyponym catalog on the fly. In our particular case, the discovery of the extraction patterns as well as the entire construction of the hyponym catalog are defined as off-line processes. Therefore, the quality of the resultant resource is much more relevant than the efficiency of its construction.

Instance extraction. In this module, the patterns discovered in the previous stage are applied over a target document collection in order to locate several text segments that presumably contain an instance of the hyponym relation. The result is a set of candidate hyponym-hypernym pairs.

To locate as many as possible hyponym relations our implementation of the module considers using the Web as target document collection. In addition, in order to extract as many as possible correct relations, it uses a user-given vocabulary to instantiate the patterns (i.e., to construct the Web queries).

For instance, having the pattern "*the HYPONYM is one of the HYPERONYM*" and the target concept *stone*, our method constructs the query "*the HYPONYM is one of the stones*". Using the instantiated pattern it is possible to extract the hyponym-hypernym pair *diamond–stone* from the snippet "the diamond is one of the stones associated with Aries and Leo…", but also it is possible to discover incorrect instances such as *privatization–stone* (from the snippet "the privatization is one of the stones of market development of regional economy…"). Therefore, to differentiate between right and wrong instances it is necessary to incorporate an extra module for evaluating and ranking patterns and instances.

Instance ranking. This module evaluates the confidence of the extracted instances to belong to the hyponym relation. Its purpose is to rank the instances in such a way that those with higher probability of being correct locate at the very first positions.

This evaluation bases on the idea that pertinent instances are extracted by different patterns, and that valuable patterns allow extracting several pertinent instances. In particular, we defined an *iterative evaluation process* where instance's confidences are calculated based on pattern's confidences, and vice versa.

The following section gives details on the iterative evaluation process; especially it defines the evaluation functions used to estimate the confidence of instances and patterns.

3 Iterative Evaluation Process

As we mentioned before, the iterative evaluation process calculates the confidence of instances and patterns in accordance with their association: i.e., an instance has a greater probability of being correct if it is associated to (was extracted by) several confidence patterns, and a pattern is more relevant if it is associated to (allow extracting) several confidence instances.

The most direct approach for measuring the confidence of a pattern (its association with the extracted instances) is through its precision and recall[2]. However, these measures are impossible to assess due to the lack of information on the extension of the relation at hand. In other words, it is impossible to known in advance the whole set of pairs hyponym-hypernym existing at the Web. Therefore, it is common to evaluate the association degree between patterns and instances (hyponym-hypernym pairs) using a pointwise mutual information metric [1]. In particular, we consider three different well-known metrics to compute the mutual information between a pattern p and an extracted instance $i = (x, y)$:

$$pmi_1(p,i) = \log \frac{P(x,p,y)}{P(*,p,*)P(x,*,y)} \tag{1}$$

$$pmi_2(p,i) = \frac{|x,p,y|}{|x,*,y|} \tag{2}$$

$$pmi_3(p,i) = \log \frac{|x,p,y||*,p,*|}{|x,p,*||*,p,y|} \tag{3}$$

where $|x,p,y|$ and $P(x,p,y)$ indicate the absolute and relative frequencies of the pattern p instantiated with terms x and y, and the asterisk (*) represents a wildcard.

Based on any given association metric, we compute the confidence values of patterns and instances as proposed by [9]:

$$c_\pi(p) = \frac{\sum_{i \in I'} \left(\frac{pmi(p,i)}{\max_{pmi}} \times c_\sigma(i) \right)}{|I|} \tag{4}$$

$$c_\sigma(i) = \frac{\sum_{p \in P'} \left(\frac{pmi(p,i)}{\max_{pmi}} \times c_\pi(p) \right)}{|P|} \tag{5}$$

where $c_\pi(p)$ and $c_\sigma(i)$ are the confidences of pattern p and instance i respectively, max_{pmi} indicates the maximum pointwise mutual information between all patterns and all instances, I' is the set of instances extracted by pattern p, and P' is the set of patterns that extract the instance i.

3.1 Computing the Initial Confidence of Patterns

It is noticeable from formulas (4) and (5) that the confidence values of patterns and instances are recursively defined. In this scheme, the initial confidence of patterns is commonly estimated by (4) using $c_\sigma(i) = 1$ for the manually supplied seed instances.

---

[2] *Precision* indicates the percentage of correct instances extracted by the pattern. On the other hand, *recall* is the percentage of all relevant instances (in the target document collection) that were actually extracted by the pattern.

Given that the set of seed instances is –in the majority of the cases– very small, this kind of "probabilistic" estimation tends to favor patterns with very high precision or very high recall, but not necessarily gives preferentiality to those patterns showing a good balance between both measures. In order to achieve this balance we propose to compute the initial confidence of patterns using a kind of *F*-measure metric[3]:

$$c_\pi(p) = \frac{F(p)}{\max_{\forall l \in P}\{F(l)\}}, \text{ where} \tag{6}$$

$$F(p) = \frac{2 \times E(p) \times R(p)}{E(p) + R(p)} \tag{7}$$

Here, $E(p)$ indicates the proportion of seed instances extracted by pattern p (i.e., a kind of precision of p), $R(p)$ is the quotient between the number of seed instances and the whole set of instances extracted by p (i.e., a kind of recall of p), and finally $\max\{F(l)\}$ is a normalization factor.

4 Experimental Results

To evaluate the proposed method we approached the discovery of hyponyms for a given set of concepts in *Spanish* language. The following passages present the achieved results at each module of the method.

For pattern discovery, we considered a set of 25 seed instances and used Google to retrieve 500 text segments per seed instance. This way, we constructed a corpus of 12,500 segments expressing the hyponym relation. From this corpus we extracted 43 lexical extraction patterns. Some of these patterns are shown in table 1. It is noticeable that the quality of the discovered patterns is very diverse. Some are too specific and precise but not so applicable, whereas some others are too general, but guarantee a high coverage.

Table 1. Some lexical patterns for extracting hyponyms

Original extraction patterns (in Spanish)	***Extraction patterns (English translation)***
el HYPONYM es un HYPERNYM que	*the HYPONYM is a HYPERNYM that*
el HYPONYM es el único HYPERNYM	*the HYPONYM is the single HYPERNYM*
el HYPONYM es uno de los HYPERNYM mas	*the HYPONYM is one of the HYPERNYM more*
las HYPONYM son una HYPERNYM	*the HYPONYM are a HYPERNYM*
El uso de la HYPONYM como HYPERNYM	*the use of HYPONYM as HYPERNYM*

For instance extraction, we firstly instantiated the acquired patterns using a given set of concepts (hypernyms). In particular, we considered the following five terms: *banco* (bank), *enfermedad* (disease), *felino* (feline), *profesión* (profession) and *roca* (stone). The selection of these terms allow studying the use of the patterns in very different topics, and therefore to obtain a general notion about their applicability. Using the instantiated patterns as web queries, we extracted 851 candidate hyponym

[3] *F-measure* is the weighted harmonic mean of precision and recall.

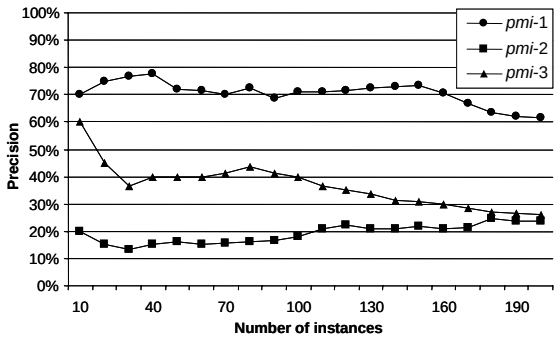

Figure 2. Comparing three pointwise mutual information metrics

instances: 193 related to bank, 307 to disease, 9 to feline, 226 to profession, and 116 to stones.

Finally, we ranked the list of hyponyms by applying the evaluation process described in section 3. In a first experiment, we compared the results obtained by using three different well-known metrics for computing the mutual information between a pattern and an instance (see formulas 1-3). Figure 2 shows the achieved precision curves after three iterations. It is clear that pmi_1 is the best approach for this specific problem. In average, it was 50% points better than pmi_2 and 30% superior to pmi_3. It is important to mention that in this experiment the initial confidence of patterns were estimated as usually, that is, using $c_\sigma(i) = 1$ for the manually supplied seed instances.

In a second experiment, we evaluated the impact of applying the proposed method for computing the initial confidence of patterns (refer to formula 4). The following figures present the evaluation results on the 200-top ranked instances. On the one hand, figure 3 shows the results obtained by the proposed method for the first three iterations of the evaluation process. In this case, we used the *F*-measure metric to compute the initial pattern's confidences. On the other hand, figure 4 compares the precision curves after three iterations using two different metrics for computing the initial confidences, the traditional one based on the pmi_1 metric and the proposed one based on the *F*-measure. The achieved results corroborate the relevance of the idea of iteratively computing the confidence of instances and patterns in accordance with

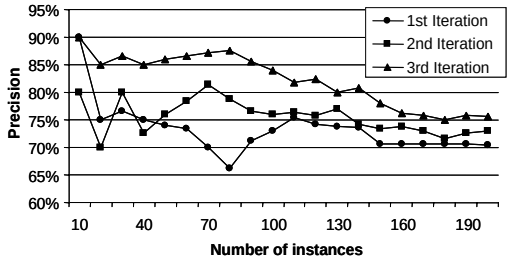

Figure 3. Results of the proposed method

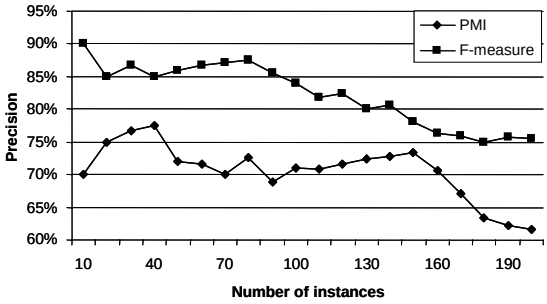

Figure 4. Pmi_1 vs. *F-measure* for computing the initial pattern confidences

their association. In addition, they show the impact of the initial pattern's confidences over the final ranking, and demonstrate the convenience of giving more importance to those patterns showing a good balance between precision and recall.

5 Conclusions

This paper proposed a new method for extracting hyponyms from free text. The method consists of three main modules. The first one focuses on the acquisition of extraction patterns from the Web, the second one uses the acquired patterns to extract a set of instances that presumably belong to the hyponym relation, finally, the third module considers the ranking of the extracted instances.

In particular, the proposed method differs from previous approaches in that: (*i*) it only considers lexical information, whereas the rest of the works make use of lexico-syntactic patterns, (*ii*) it uses all discovered patterns –specific and general– for instance extraction, and (*iii*) it applies a new metric, based on *F*-measure, for computing the initial confidence of the extraction patterns.

The presented experimental results demonstrated the feasibility of using lexical patterns to extract hyponyms from free texts as well as the pertinence of the proposed metric for computing the initial pattern's confidence.

Future work will be focused on concluding about the language independence and large-scale performance of the proposed method. In particular, our current results have demonstrated that the method is adequate for Spanish (a language with moderate complex morphology), therefore, we expect that it will be also useful for dealing with other romance languages or other languages with relative simple morphology such as English. However, it is very possible that the method will not be pertinent for dealing with languages having a complex morphology, for instance, agglutinative languages such as German or Arabic.

On the other hand, we plan to use the method to extract hyponyms for a large number of concepts (i.e., using a greater base vocabulary). Using a large number of concepts will not be a problem for the method; on the contrary, we believe that having more information will allow to obtain an accurate estimation of the confidences of instances and patterns.

References

1. Blohm S., Cimiano P. Learning Patterns from the Web - Evaluating the Evaluation Functions - Extended Abstract. OTT'06 - Ontologies in Text Technology: Approaches to Extract Semantic KnowledInformation, Osnabrück, Germany, 2006.
2. Denicia C., Montes M., Villaseñor L., García R. A Text Mining Approach for Definition Question Answering. 5th International Conference on Natural Language Processing (FINTAL-06). Turku, Finland, 2006.
3. García-Hernández, R., Martínez-Trinidad F., and Carrasco-Ochoa A. (2006). A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. International Conference on Computational Linguistics and text Processing, CICLing-2006. Mexico City, Mexico, 2006.
4. Gelbukh A. and Sidorov G. Procesamiento automático del español con enfoque en recursos léxicos grandes. IPN, 2006, 240 p.
5. Hearst M. Automatic acquisition of hyponyms from large text corpora. Conference on Computational Linguistics (COLING-92). Nantes, France, 1992.
6. Lin D., Zhao S., Qin L., and Zhou M. Identifying synonyms among distributionally similar words. International Joint Conference of Artificial Intelligence (IJCAI-2003). Acapulco, Mexico, 2003.
7. Lucero C., Pinto D., Jiménez H. A Tool for Automatic Detection of Antonymy Relations. Workshop on Herramientas y Recursos Linguísticos para el Español y el Portugués. IBERAMIA-2004. Puebla, Mexico, 2004.
8. Mann G. S. Fine-Grained Proper Noun Ontologies for Question Answering. SemaNet-02: Building and Using Semantic Networks. Taipei, Taiwan, 2002.
9. Pantel P., Pennacchiotti M. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. Conference on Computational Linguistics/Association for Computational Linguistics (COLING/ACL-06). Sydney, Australia, 2006.
10. Ravichandran D., Pantel P., Hovy E. The Terascale Challenge. Proceedings of KDD Workshop on Mining for and from the Semantic Web. Seattle, WA, USA, 2004.