Semantic Pattern Learning Through Maximum Entropy-based WSD technique*

Maximiliano Saiz-Noeda
Depto. de Lenguajes y
Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
max@dlsi.ua.es

Armando Suárez
Depto. de Lenguajes y
Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
armando@dlsi.ua.es

Manuel Palomar
Depto. de Lenguajes y
Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
mpalomar@dlsi.ua.es

Abstract

This paper describes a Natural Language Learning method that extracts knowledge in the form of semantic patterns with ontology elements associated to syntactic components in the text. The method combines the use of EuroWordNet's ontological concepts and the correct sense of each word assigned by a Word Sense Disambiguation(WSD) module to extract three sets of patterns: subject-verb, verb-direct object and verb-indirect object. These sets define the semantic behaviour of the main textual elements based on their syntactic role. On the one hand, it is shown that Maximum Entropy models applied to WSD tasks provide good results. The evaluation of the WSD module has revealed a accuracy rate of 64% in a preliminary test. On the other hand, we explain how an adequate set of semantic or ontological patterns can improve the success rate of NLP tasks such us pronoun resolution. We have implemented both modules in C++ and although the evaluation has been performed for English, their general features allow the treatment of other languages like Spanish.

1 Introduction

Semantic patterns, as defined in this method, configure a system to add a new information source to Natural Language Processing (NLP) tasks. To obtain these semantic patterns, it is necessary to count on different tools. On the one hand, a full parser must make a syntactic analysis of the text. This parsing will allow the selection of the different syntactic functional elements such as subject, direct object (*DObj*) and indirect object (*IObj*). On the other hand, a WSD tool must provide the correct sense in order to ensure the appropriate selection of the ontological concept associated to each word. Finally, with the parsing and the correct sense of each word, the pattern extraction method will form and store ontological pairs that define the semantic behaviour of each sentence.

2 Full parsing

The analyzer used for this work is the Conexor's FDG Parser (Pasi Tapanainen and Timo Järvinen, 1997). This parser tries to provide a build dependency tree from the sentence. When this is not possible, the parser tries to build partial trees that often result from unresolved ambiguity. One visual example of this dependency trees is shown in Figure 1 where the parsing tree of sentence (1) is illustrated.

(1) `The minister gave explanations to the Government.`

As seen in Figure 2, the analyzer assigns to each word a text token (second column), a base form (third column) and functional link

*This paper has been partially supported by the Spanish Government (CICYT) project number TIC2000-0664-C02-02.

```
0
1  The           the          det:>2     @DN> DET SG/PL
2  minister      minister     subj:>3    @SUBJ N NOM SG
3  gave          give         main:>0    @+FMAINV V PAST
4  explanations  explanation  obj:>3     @OBJ N NOM PL
5  to            to           dat:>3     @ADVL PREP
6  the           the          det:>7     @DN> DET SG/PL
7  Government    government   pcomp:>5   @<P N NOM SG/PL
   .             .
```

Figure 2: FDG Analyser's output example

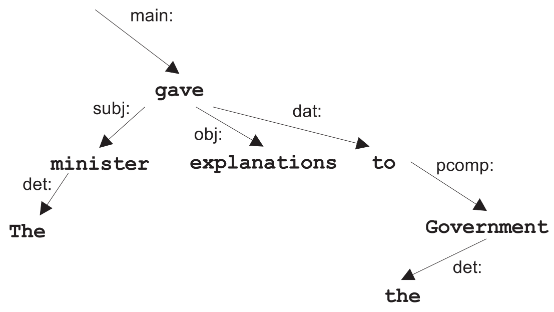

Figure 1: *Parsing tree*

names, lexico-syntactic function labels and parts of speech (fourth column). Figure 1 shows the parsing tree related to this output. These elements are enough for the pattern extraction method to be applied to NLP tasks.

Regarding to the evaluation of the parser, the authors report an average precision and recall of 95% and 88% respectively in the detection of the correct head. Furthermore, they report a precision rate between 89% and 95% and a recall rate between 83% and 96% in the selection of the functional dependencies.

3 WSD based on Maximum Entropy

A WSD module is applied to this parser's output, in order to select the correct sense of each entry.

Maximum Entropy(ME) modeling is a framework for integrating information from many heterogeneous information sources for classification (Manning and Schütze, 1999). This WSD system is based on conditional ME probability models. The system implements a supervised learning method consisting of the building of word sense classifiers through training on a semantically tagged corpus. A classifier obtained by means of a ME technique consist of a set of parameters or coefficients estimated by an optimization procedure. Each coefficient associates a weight to one feature observed in the training data. A feature is a function that gives information about some characteristic in a context associated to a class. The basic idea is to obtain the probability distribution that maximizes the entropy, that is, maximum ignorance is assumed and nothing apart of training data is considered. As advantages of ME framework, knowledge-poor features applying and accuracy can be mentioned; ME framework allows a virtually unrestricted ability to represent problem-specific knowledge in the form of features (Ratnaparkhi, 1998).

Let us assume a set of contexts X and a set of classes C. The function $cl : X \rightarrow C$ that performs the classification in a conditional probability model p chooses the class with the highest conditional probability: $cl(x) = \arg\max_c p(c|x)$. The features have the form expressed in equation (1), where $cp(x)$ is some observable characteristic[1]. The conditional probability $p(c|x)$ is defined as in equation (2) where α_i are the parameters or weights of each feature, and $Z(x)$ is a constant to ensure that the sum of probabilities for each possible class in this context is equal to 1.

$$f_{c'}(x,c) = \begin{cases} 1 & \text{if } c' = c \text{ and } cp(x) = true \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$p(c|x) = \frac{1}{Z(x)} \prod_{j=1}^{K} \alpha_j^{f_j(x,c)} \tag{2}$$

The features defined on the present system are,

[1] This is the kind of features used in the system due to it is required by the parameter estimation procedure, but the ME approach is not limited to binary funtions.

basically, collocations of content words and POS tags of function words around the target word. With only this information the system obtains results comparable to other well known methods or systems. For training, DSO sense tagged English corpus (Hwee Tou Ng and Hian Beng Lee, 1996) is used. The DSO corpus is structured in files containing tagged examples of some word. The tags correspond to the correct sense in WordNet 1.5 (FellBaum, 1998). The examples were extracted from articles of the Brown Corpus and Wall Street Journal.

The implemented system has three main modules: the Feature Extractor (FE), the Generalized Iterative Scaling (GIS), and the Classification module. Each word has its own ME model, that is, there will be a distinct classifier for each one. The FE module automatically defines the features to be observed on the training corpus depending on the classes (senses) defined in WordNet for a word. The GIS module performs the parameter estimation. Finally, the Classification module uses this set of parameters in order to disambiguate new occurrences of the word.

3.1 Evaluation and results

Some evaluation results over a few terms of the aforementioned corpus are presented in Table 1. The system was trained with features that inform of content words in the sentence context (w_{-1}, w_{-2}, w_{-3}, w_{+1}, w_{+2}, w_{+3}), multi-word expressions ((w_{-2}, w_{-1}), (w_{-1}, w_{+1}), (w_{+1}, w_{+2}), (w_{-3},w_{-2},w_{-1}), (w_{-2}, w_{-1},w_{+1}), (w_{-1},w_{+1},w_{+2}), (w_{+1}, w_{+2},w_{+3})), and POS tags (p_{-1}, p_{-2}, p_{-3}, p_{+1}, p_{+2}, p_{+3}). For each word, the training set is divided in 10 folds, 9 for training and 1 for evaluation; ten tests were accomplished using a different fold for evaluation in each one (10-fold cross-validation). The accuracy results are the average accuracy on the ten tests for a word.

Results comparison with previous work is difficult because there is different approaches to the WSD task (knowledge based methods, supervised and unsupervised statistical methods...) (Mihalcea and Moldovan, 1999) and many of them focus on a different set of words and sense definitions. Furthermore, the training corpus seems to be critical to the application of the learning to a specific domain (Escudero et al., 2000b).

	ocurrences	accuracy	standard deviation
age,N	48,2	0,584	0,134
art,N	38,0	0,623	0,090
car,N	136,7	0,963	0,048
child,N	105,1	0,809	0,073
church,N	35,8	0,625	0,126
cost,N	143,2	0,895	0,051
fall,V	143,7	0,759	0,242
head,N	83,3	0,714	0,125
interest,N	147,8	0,619	0,173
know,V	143,3	0,421	0,087
line,N	132,8	0,529	0,154
set,V	126,1	0,537	0,139
speak,V	51,1	0,729	0,080
take,V	138,0	0,264	0,042
work,N	118,9	0,530	0,175
Overall		0,637	

Table 1: Evaluation results from DSO-WSJ

In the experiment presented here, the selection of the target words and the corpus used are the same that (Escudero et al., 2000a) where a Boosting method is proposed. In this paper a comparison between some WSD methods is shown. Boosting is the most successful method with a 68.1 % accuracy. Our method obtains lower accuracy but this is a first implementation and a better feature selection is expected to improve our results.

4 Semantic Pattern Learning

Once the WSD phase has been performed, the semantic pattern extraction module can be executed. This module extracts head word pairs with subject-verb, verb-*DObj* and verb-*IObj* roles in the sentence and convert them into patterns formed by ontological concepts extracted from EuroWordNet.

4.1 EuroWordNet's ontology

EuroWordNet (Vossen, 2000) is a multilingual lexical database representing semantic relations among basic concepts for West European languages. In our case, we are going to work with isolated WordNets, it means, we won't take advantage of its multilingual feature, although we will use the ontology defined on it.

EuroWordnet's ontology consists of 63 higher-level concepts and distinguishes three types of entities:

- *1stOrderEntity*: any concrete entity (publicly) perceivable by the senses and located at any point in time, in a three-dimensional space, e.g.: vehicle, animal, substance, object.

- *2ndOrderEntity*: any Static Situation (property, relation) or Dynamic Situation, which cannot be grasped, heard, seen, felt as an independent physical thing. They can be located in time and occur or take place rather than exist, e.g.: happen, be, have, begin, end, cause, result, continue, occur..

- *3rdOrderEntity*: any unobservable proposition which exists independently of time and space. They can be true or false rather than real. They can be asserted or denied, remembered or forgotten, e.g.: idea, thought, information, theory, plan.

These ontological concepts, associated to each synset from EuroWordNet, give semantic properties to these synsets that can be used, as we will see in the nexts sections, for improving the information source in Natural Language Processing tasks.

4.2 The Learning Process

From each clause, the module extracts the verb and (if exists) its subject, its direct object and its indirect object. With these elements, three possible pairs can be formed using the verb and the noun head of the aforementioned syntactic components. The verb head and the noun head are looked up in EuroWordNet's ontology using the correct sense previously selected. This query generates three possible ontological pairs that define, for each clause, the semantic concept associated to the main syntactic elements.

Sentence (2) corresponds to a fragment extracted from a training corpus in English.

(2) `The minister`^{4} `gave`^{7} `explanations`^{2} `to the Government`^{2}`.`

As shown in section 2, the output of the parser generates the next functional entities:

```
Verb:          give
Subject head:  minister
D.Obj.  head:  explanations
I.Obj.  head:  Government
```

The superscripts indicate the correct sense in EuroWordNet for each word. After consulting EuroWordNet the semantic patterns formed are:

```
Subj|V:   Human,Occupation|Communication
V|DObj:   Communication|Agentive,Mental
V|IObj:   Communication|Group,Human
```

These patterns will be stored in their corresponding files in order to be consulted later by the NLP task.

This process is completely automatic and the error rate in the pattern extraction come from the aforementioned errors in the WSD and parsing phases.

This strategy defined just as it has been done is, in principle, a little bit naive. Obviously, this is the single basis for the approach, but depending on the application, it can be combined with more sophisticated methods to improve its effectiveness. In this way, it is possible to make more elaborated combinations of ontological concepts to form new branches in the ontology defined by EuroWordNet.

5 Applying the method to anaphora resolution

Since the aforementioned semantic patterns reveal the semantic behaviour of the main textual elements, this Natural Language learning process can be applied to any task that involves text understanding.

One possible application in this way could be the anaphora resolution problem, one of the most active research areas in Natural Language Processing.

The comprehension of anaphora is an important process in any NLP system, and it is among the toughest problems to solve in Computational Linguistics and NLP. According to Hirst (Hirst, 1981): *"Anaphora, in discourse, is a device for making an abbreviated reference (containing fewer bits of disambiguating information, rather than being lexically or phonetically shorter) to some entity (or entities) in the expectation that the receiver of the discourse will be able to disabbreviate the reference and, thereby, determine the identity of the entity."*

The reference to an entity is generally called an anaphor (e.g. a pronoun), and the entity to which the anaphor refers is its referent or antecedent. For instance, in the sentence *"John$_i$ ate an apple. He$_i$ was hungry"*, the pronoun *he* is the anaphor and it refers to the antecedent *John*.

Traditionally, some of the most relevant approaches to solve anaphora have been those called *poor-knowledge approaches*. They use limited knowledge (lexical, morphological and syntactic information sources) for the detection of the correct antecedent. These proposals have report high success rates for English (89.7%) (Mitkov, 1998) and for Spanish (83%) (Ferrández et al., 1999). Taking this basis, it is possible to improve the results of a resolution method adding other sources such us semantic, pragmatic, world-knowledge or indeed statistical information.

We have explored the use of semantic information extracted from an ontology and its application to the anaphora resolution proccess. This additional source has give good results on restricted texts (Azzam et al., 1998). Nevertheless, its application on unrestricted texts has not been so satisfactory, mainly due to the lack of adequate and available lexical resources. Due to this, we consider that the pattern learning can complement the semantic source in order to establish additional criteria in the antecedent selection. In addition, we believe that an adequate selection of patterns can improve the success rate in anaphora resolution on unrestricted texts.

Each pattern contributes a compatibility feature between two syntactic elements. The whole set of patterns is a knowledge tool that can be consulted in order to define the compatibility between a pronoun and a candidate according to their syntactic role (subject, direct object and indirect object) and their relation with the verb. So, looking up the concepts associated to the antecedents of the pronoun and the verb, and using the syntactic relation between the pronoun and its verb, the semantic patterns can provide a compatibility degree to help the selection of the antecedent. A method oriented to anaphora resolution that uses these kinds of patterns extracted from two ontologies is detailed in (Saiz-Noeda and Palomar, 2000).

The benefit of this approach is shown in a classical example shown in (3).

(3) [The monkey]$_i$ climbed [the tree]$_j$ to get [*a coconut*]$_k$ when [the sun]$_l$ was rising. *It*$_k$ was ripe.

In this example, there are four possible antecedents of the pronoun '*it*'. Basing the resolution only in morpho-syntactic information, it is not possible to solve it correctly. None of the candidates would be rejected regarding to their morphological features (all of them are masculine and singular). The classical approaches would determine that 'the monkey', for having the same subject role as the pronoun, or 'the sun', for being the closest to the pronoun, could be the correct antecedent. Nevertheless, it is clear that the correct one in this case is 'the coconut'. Only a semantic pattern applied to this method could give additional information to solve it correctly.

If we would extract ontological concepts for all the candidates, we would be able to compare the compatibility degree with the pronoun. One possible output could be the one in next table:

Subject	*concept*	*verb*
monkey	animal	be ripe
tree	plant	be ripe
coconut	fruit	be ripe
sun	star	be ripe

Examining this table it is easy to notice that, when applying this additional information, the suggestion of the system would be the correct antecedent, mainly based on a good previous pattern learning.

This pronoun resolution system with additional information provided by the semantic patterns has been evaluated on a corpus formed by a set of texts containing news regarding the common topics in a newspaper (national, international, sports, society, economy, ...). Results obtained in the preliminary evaluation of this pronoun resolution reveal a success rate of 79.3% anaphors correctly solved. Although it has not been mentioned before, it is very important to have in mind that this method provides a fully automatic anaphora resolution process. Methods previously mentioned apply the resolution process over supervised steps to achieve such high rates. When the process is automated, the success rate decrease dramatically up to less than 55% (Mitkov, 2001).

6 Conclusions and outstanding work

In this paper we have presented a semantic pattern learning system driven by a WSD method based on Maximum Entropy models. These semantic patterns have been applied to the anaphora resolution through the construction of ontological patterns. The adding of this pattern learning improve, as it can be seen, the anaphora resolution process. We have pointed out the main advantages of this approach comparing it with other.

The WSD method is based on conditional Maximum Entropy probability models. It is a supervised learning method that uses a semantically annotated corpus for training. ME models are used in order to estimate functions that performs a sense classification of nouns, verbs and adjectives. The learning phase has been made with simple features with no deep linguistic knowledge. Preliminary results indicate that the accuracy of the model is comparable to other learning methods.

The main problem in the addition of this kind of knowledge is the lack of appropriate resources to deal with these tasks. In our research work we are trying to apply these techniques both in English and Spanish. The WSD method have been mainly developed in English, but one of our main goals is the design of a complete anaphora resolution system for Spanish. In this way, the main problem is the short available resources regarding to semantically tagged corpora in Spanish (unlike in English). This lack affects the correct development of tasks belonging to the research line shown in this paper, such us the pattern learning and the anaphora resolution. Nevertheless, this shortage opens the door to new research lines that join English resources and multilingual techniques for the generation of patterns in other languages from the learned English patterns.

References

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1998. Coreference Resolution in a Multilingual Information Extraction System. In *Proceedings of the Workshop on Linguistic Coreference. First Language Resources and Evaluation Conference (LREC'98).*, pages 74–78.

Gerard Escudero, Lluis Màrquez, and German Rigau. 2000a. Boosting applied to word sense disambiguation. In *Proceedings of the 12th Conference on Machine Learning ECML2000*, Barcelona, Spain.

Gerard Escudero, Lluis Màrquez, and German Rigau. 2000b. On the portability and tuning of supervised word sense disambiguation systems. In *Proceedings of the Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, China.

Christiane FellBaum. 1998. *WordNet, an electronic lexical database*. Fellbaum, C. eds. MIT Press.

Antonio Ferrández, Manuel Palomar, and Lidia Moreno. 1999. An empirical approach to Spanish anaphora resolution. *Machine Traslation*, 14(3/4):191–216.

Graeme Hirst. 1981. *Anaphora in Natural Language Understanding*. Springer-Verlag, Berlin.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Senses: An Exemplar-Based Approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, ACL 1996*.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Rada Mihalcea and Dan I. Moldovan. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic*, pages 152–158, College Park, Maryland, USA, June.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 869–875, Montreal (Canada), August.

Ruslan Mitkov. 2001. Outstanding issues in anaphora resolution. In *Proceedings of the Second International Conference on Intelligent Text Processing and Computational Linguistics CICLing2001.*, Lectures Notes In Computer Science. Springer-Verlag, pages 110–125, Mexico City (Mexico), February. Springer Verlag.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 64–71, April.

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.

Maximiliano Saiz-Noeda and Manuel Palomar. 2000. Semantic Knowledge-driven Method to Solve Pronominal Anaphora in Spanish. In Springer Verlag, editor, *NLP'2000 Filling the gap between theory and practice*, Lectures Notes In Artificial Intelligence. Springer-Verlag, pages 204–211, Patras, Greece, June.

Piek Vossen. 2000. EuroWordNet: a Multilingual Database with WordNets in 8 languages. *The ELRA Newsletter*, 5(1):9–10.