Enriching EWN with Syntagmatic Information by means of WSD

Iulia Nica[♣♠], Mª Antònia Martí[♣], Andrés Montoyo[♦] and Sonia Vázquez[♦]

[♣] CLiC - Centre de Llenguatge i Computació
Department of General Linguistics
University of Barcelona, Spain
iulia@clic.fil.ub.es, amarti@ub.edu
[♠] Department of General Linguistics
University of Iasi, Romania

[♦] Research Group of Language Processing and Information Systems
Department of Software and Computing Systems
University of Alicante, Spain
{montoyo,svazquez}@dlsi.ua.es

Abstract

Word Sense Disambiguation confronts with the lack of syntagmatic information associated to word senses. In the present work we propose a method for the enrichment of EuroWordNet with syntagmatic information, by means of the WSD process itself. We consider that an ambiguous occurrence drastically reduces its ambiguity when considered together with the words it establishes syntactic relations in the sentence: the claim of "quasi one sense per syntactic relation". On this hypothesis, we obtain sense-tagged syntactic patterns for an ambiguous word intensively using the corpus, with the help of EWN and of associated WSD algorithms. For an occurrence disambiguation, we also consider the whole sentential context where we apply the same WSD algorithms, and combine the sense proposals from the syntactic patterns with the ones from the sentential context. We evaluate the hole WSD method on the nouns in the Spanish Senseval-2 exercise and also the utility of the syntactic patterns for the sense assignment. The annotated patterns we obtain in the WSD process are incorporated into EWN, associated to the synset of the assigned sense. As the syntactic pattern repeat themselves in the text, if sense-tagged, they are a valuable information for future WSD tasks.

1. Introduction

WordNet (WN) has become the standard lexical device in the area of NLP, both for applications and for intermediary tasks. We consider the relation between WN and these processes a dialectical one. In the present work we investigate the interaction between WN and WSD (Word Sense Disambiguation): the varied exploitation of WN leads to improvements in the WSD process, and the information so acquired, if incorporated into the lexicon, can augment the quality of future sense assignments.

One of the limitations of WSD is the unavailability of syntagmatic information associated to senses. We propose an automatic method for the enrichment of WN with this kind of information, by means of a WSD process. The method is based on an approach to WSD that exploits both a corpus and WN, in its original form and in an own adaptation.

The annotated patterns we obtain are registered into a special lexical device where the patterns are connected to the synset in EWN of the assigned sense. Our proposal is useful both for knowledge-based WSD systems and corpus-based systems. The syntactic patterns repeat themselves in the text, thus, if sense-tagged, they are a valuable information for future WSD tasks. They are used here inside a knowledge-based WSD method.

We carry out this investigation for Spanish, so our object of study is the Spanish component of EuroWordNet (EWN). We limit here to nouns, but the proposal can be extended to other categories. The experimentation is performed on the nouns in Senseval-2, in order to evaluate the contribution of the sense-tagged patterns to sense assignment.

The method requires a corpus and a POS-tagger, a WN component and very little syntactic knowledge: a list of syntactic relations between the POS categories and of their textual realisations. Thus, it is easily transportable to other languages that dispose of these devices.

The paper has the following structure: the related previous work (section 2); the proposal for obtaining sense-tagged syntactic patterns (section 3); its application to WSD (section 4); conclusions and future work (section 5).

2. Similar Previous Work in Sense-Tagging and WSD

In one of the first methods for the automatic creation of sense-tagged materials, Gale *et al.* (1992) use an aligned French-English bilingual corpus to discriminate occurrences of a English word with different senses by means of the different translations that the word has into French. In (Yarowsky, 1992), the Roget's categories are exploited to collect contexts from Grolier's Encyclopedia for the different senses of a word that is common to more thesaurus categories: he looks for sentences with words in each of these categories. Yarowsky (1995), in a bootstrapping approach, augments a small set of labelled seed collocations by locating examples containing the seeds, from these extracting new patterns, then looking again into the corpus to find sentences with these patterns. Leacock et al. (1998) find examples for the different senses of a word using the monosemous words in synsets related to these senses in WordNet. Mihalcea and Moldovan (1999) try to overcome these limitations with the help of other information, as the glosses in WordNet, and by substituting the corpus with Internet.

There are also WSD methods that meet ours from one point of view or other. The syntactic information exploited for WSD has been limited generally to verb-subject and verb-object relations (Ng, 1996; Martínez *et al*., 2002, etc.), with few exceptions (Lin, 1997; Stetina *et al*., 1998), on corpora annotated at the syntactic and semantic levels. The use of functional words that are contiguous to the ambiguous occurrence have been done especially from an example-based approach, and so it has been related to and dependent on a sense tagged corpus (Pedersen, 2001). Our work is closer to corpus-based methods as (Montemagni *et al*., 1996; Federici *et al*., 2000), defined as "Paradigm-driven Approach" to WSD, and (Agirre and Martínez, 2001). In these methods there are combined paradigmatic variants for the two lexical content positions of what we call syntactic pattern. The combination is performed only for verb-argument relations and on syntactic patterns already tagged, at syntactic and sense levels.

We propose an alternative for the creation of sense-tagged examples: the labelling process is executed for words integrated into syntactic patterns. Our method is independent on a syntactically and semantically tagged corpus, and it uses different types of syntactic relations involving nouns. The method works with real examples from texts, and from there it obtains substitutes for the focalised word into the syntactic pattern. Thus it is independent of the existence of related monosemous word in (E)WN. Furthermore, it works in a good percentage on the local context, with more syntactic patterns, so it limits the data-sparseness problem affecting the methods that consider all the sentence.

3. Proposal: Sense-Tagged Syntactic Patterns

In our approach to WSD, we consider that the sense of an ambiguous occurrence is mainly determined by means of its syntactic relations. For the formal treatment of the context from this perspective, we introduce the term of syntactic pattern: a triplet X-R-Y, formed by two lexical content units X and Y (nouns, adjectives, verbs, adverbs) and a relational element R, which corresponds to a syntactic relation between X and Y. Examples: [*grano*-N *de*-PREP *azúcar*-N], [*pasaje*-N *subterráneo*-ADJ][1].

We see the integration of an ambiguous occurrence into a syntactic pattern as a first approximation to its sense, as a pre-sense tagging: the WSD process will be first developed for an ambiguous occurrence in relation with its syntactic patterns, and after in relation with the sentence in which it occurs. At the basis of our approach it lies the hypothesis that inside a syntactic pattern a word will reduce its polysemy and will tend to be monosemous: the "quasi one sense per syntactic pattern" claim.

The integration allows us to identify, into the corpus, information both of paradigmatic and syntagmatic type associated to the word into the pattern. For the sense assignment, we apply on this information a WSD algorithm that uses an adaptation of EWN.

The enriching method we propose consists thus of three phases for a given noun X in EWN:

1º. Identification of nouns occurrences and of their syntactic patterns.

2º. Sense disambiguation for X inside a syntactic pattern.

3º. Registration of the syntactic patterns previously sense tagged in a lexical device where the patterns are connected to the synsets in EWN of the assigned senses.

3.1. Identification of syntactic patterns

In order to identify occurrences for a noun X in EWN and for their syntactic patterns, we work on a POS-tagged corpus (EFE)[2]; we call it "search corpus". The identification of the syntactic patterns is done following criteria of structure and of frequency. We predefine a list of basic patterns: [N ADJ], [ADJ N], [N PART], [PART N], [N CONJ N], [N PREP N], [N, N]. As the quality of the sense assignment for X inside the syntactic pattern is highly dependent on the syntactic patterns identification, we introduce some filters on the patterns we obtain: we impose the condition on the potential patterns to repeat into the corpus and we eliminate the ones with more than 1000 substitutes for the word to be disambiguated. The substitutes are obtained by fixing the rest of the pattern and letting variable the position of the focalised word at lemma level (the set S1 in 2.2). In this way, we obtain syntactic patterns P_k for the noun X. For the 688 occurrences with verifiable sense assignation in terms of EWN senses (from all 799) in the Senseval-2 test corpus, we obtain 318 syntactic patterns corresponding to 294 occurrences (that is a coverage with patterns of 42,73%).

3.2. Sense-tagging inside the syntactic patterns

For every syntactic pattern P_{k0} of X, we extract from corpus both paradigmatic and syntagmatic information related to X inside P_{k0}: the sets S1 and S2 below.

[1] We use here the following POS-tags: N=noun, ADJ=adjective, PART=past participe, CONJ=conjunction, PREP=preposition, DET=determinant, ADV=adverb.

[2] The POS-tagger for Spanish is the one from (Civit, 2003).

-**S1** is the set of the most frequent 20 nouns in the paradigm corresponding to the position X into the considered syntactic pattern P_{k0}. The paradigm is obtained by fixing the syntactic pattern P_{k0} at lemma and morphosyntactic levels, and letting variable only the position of X at lemma level.

-**S2** is the set of the most 10 frequent nouns in sentences with the syntactic pattern P_{k0}.

We obtain thus the sets S_{1k} and S_{2k} corresponding to X inside the syntactic patterns P_k.

For the sense assignment to the word inside the syntactic patterns, we use several WSD heuristics. For a given syntactic pattern P_{k0}, the heuristics apply a WSD algorithm (A1 or A2 below) on one of the sets S_{1k0}, S_{2k0}.

The WSD algorithms we use are:

A1: Commutative Test (CT) (Nica *et al.*, 2003). The algorithm exploits an adaptation of EWN we have obtained in the following way: for every sense Xi of a given word X in EWN, we extract the set Di of nouns related to it in EWN along the different lexical-semantic relations. We eliminate then the common elements between the sets, obtaining so disjunctive sets Di. As the elements of the set Di are related exclusively with the sense Xi, they become sense discriminators for Xi.

For example, in EWN, *órgano* has five senses[3]:

órgano_1: 'part of a plant';

órgano_2: 'governmental agency, instrument';

órgano_3: 'functional part of an animal';

órgano_4: 'musical instrument' ;

órgano_5: 'newspaper'.

Correspondingly, we obtain from the EWN hierarchy the following Sense Discriminators sets:

D_1: {*flor, pera, manzana, bellota, hinojo, semilla, …*}

D_2: {*agencia, unidad administrativa, banco central, …*}

D_3: {*músculo, riñón, oreja, ojo, glándula, dedo, …*}

D_4: {*instrumento de viento, aparato, teclado, pedal, …*}

D_5: {*periódico, publicación, serie, serial, número, …*}

The CT algorithm applies on a set S of nouns related to a word X in a given syntactic pattern. The algorithm intersects S with every set Di; if it obtains a not empty intersection between S and Di, then it concludes that X can have the sense Xi in the syntactic pattern.

A2: Specificity Mark algorithm (Montoyo and Palomar, 2000). The algorithm works on the original form of EWN. The intuitive base of this algorithm is that the more common information two concepts share the more related they will be. In EWN, the common information shared by two concepts corresponds to the father concept of both in the hierarchy, called Specificity Mark (SM) by the authors. The heuristic takes as input a noun set and looks for the SM in EuroWordNet with the bigger density of input words in its subtree. It chooses as correct for every input word the sense situated in the sub-tree of the SM so identified, and it lets undisambiguated the words without senses in this subtree.

We have thus four WSD heuristics: CT on S1 (**H1**); CT on S1; CT on S2 (**H2**); ME on S1 (**H3**); ME on S2 (**H4**). For the final assignment, we consider the heuristics with equal weights, and we sum the votes for every sense from all the heuristics. We finally select the most voted sense over an established limit. From the 318 patterns previously identified with verifiable sense assignation, for the strict limit of at least 75% of votes, we obtain 30 accurate sense-tagged patterns with 93,33% of precision (variant A), and for the relaxed limit of 62,5%, we obtain 62 patterns with 83,87% of precision (variant B).

3.3. Enriching EWN with sense-tagged syntactic patterns

The sense-tagged patterns we obtain in 2.2. have the format [**lemma0-CAT** lemma1-CAT lemma2-CAT], where the bold marks the lemma and the category of the focused word. They also have associated the list of the senses of X with their probability to be assigned to the pattern. We register the annotated patterns we obtain (in the strict variant, A) in a special lexical device where the patterns are connected to the synset(s) in EWN of the assigned sense(s). The format of this sense-tagged pattern database is shown in table 1. We take into account the categories and the position of the words in patterns with respect to the focalised word: "-k" indicates the position k on its left side and "+k" the position k on its right side.

Noun	Pattern type (categories on positions)					Examples (lemma level)	S1 (02831270n)	S2 (03650737n)	S3 (05302115n)	S4 (07977350n)	S5 (02604665n)	Majority sense
	-2	-1	0	+1	+2							
órgano	N	S	N			*informe de ~*		100%				2
			N	A		*~ afectado*		12,5%	62,5%	25%		3

Table 1. Format for the obtained syntagmatic information

We also incorporate in this device, associated to each pattern, the corpus examples where it appears. Thus we enrich (E)WN with broader examples for senses too.

4. Application to WSD

We performed a WSD experiment using the sense-tagged patterns previously obtained, by integrating the patterns in a WSD system with two groups of heuristics:

-Heuristics I: heuristics based on the syntactic patterns (H1, H2, H3 and H4 in section 2.2.)

-Heuristics II: heuristics based on the sentence. They consist in applying one of the two algorithms from section 2.2. on the nouns set in the sentence of the occurrence.

[3] The pseudo-definitions are ours.

For the final sense assignment, we first apply the heuristics I and after the heuristics II. Repeating the Senseval-2 exercise, we obtained the results in table 2.

Heuristics	Precision	Recall	Coverage
I (variant A)	92,59%	3,63%	3,92%
I (variant B)	82,45%	6,81%	8,29%
II	31,63%	30,66%	96,94%
I (variant A) + II	33,28%	32,26%	96,94%
I (variant B) + II	35,02%	34,01%	97,09%

Table 2. Final results

The evaluation indicates a low level of performance of our method. The causes are: no use of patterns with verbs; insufficiently adequate filters on patterns; limited search corpus (70 millions words). The results indicate that there can be done WSD using only syntactic patterns and that the use of syntactic patterns improves the WSD level. For the iterative syntactic patterns in the Senseval-2 test corpus, we have also verified the "*quasi one sense per syntactic pattern*" hypothesis. Even data is very limited, it seems that there is a tendency of the syntactic patterns to associate with a unique word sense: 49 cases on 53 (92,4% of the patterns); in the other 4 cases, the word ambiguity inside the syntactic patterns reduces to two senses. This is a very partial confirmation of our strategy to integrate the ambiguous occurrences into syntactic patterns as a first step towards their disambiguation.

5. Conclusions and Future Work

We propose a method for sense-tagging words inside syntactic patterns and the enrichment of EWN with this syntagmatic information associated to senses. The method requires only a minimal preprocessing phase (POS-tagging) and very little grammatical knowledge. It can be used both to sense tag corpora and to enrich (E)WN. The enrichment of (E)WN is continuous as we find new syntactic patterns with the words from the lexicon. Furthermore, the method allows to map (E)WN and corpora by means of the syntactic patterns incorporated into (E)WN, and so to enrich (E)WN with broader examples for senses. An experiment performed in the conditions of Spanish Senseval-2 exercise reveals the utility of the sense-tagged patterns for the WSD process.

As future work, we are investigating ways to improve the patterns identification and filtering, as well as the extraction of the related information associated to the word integrated into a pattern, in order to enlarge the acquisition process of accurate sense-tagged patterns. We are also analysing the further use of the sense-tagged patterns for the acquisition of disambiguation clues and the combination of the sense-tagged patterns with other WSD methods. A necessary future step is the extension to other morphosyntactic categories and to other languages. The transfer of the method supposes only the change of the lexical and grammatical modules related to a language: the EWN component, the corpus, the POS-tagger, the very little syntactic knowledge. Thus the method is transportable with minimal costs to other languages that dispose of these devices.

6. Bibliography

• Agirre, E. and D. Martinez, 2000. Exploring automatic WSD with decision lists and the Web. In: *Proceedings of the COLING Workshop on Semantic Annotation and Intelligent Content*, Saabrücken

• Agirre, E. and D. Martínez, 2001. Learning class-to-class selectional preferences. In: *Proceedings of the ACL CONLL'2001 Workshop,* Tolouse

• Civit, M., 2003. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*, Ph.D. Thesis, University of Barcelona

• Gale, W.A., K.W. Church, and D. Yarowsky, 1992. "One sense per discourse". In: *Proceedings of DARPA speech and Natural Language Workshop,* Harriman, NY

• Leacock, C., M. Chodorow and G.A. Miller, 1998. Using Corpus Statistics and WordNet Relations for Sense Identification, *Computational Linguistics. Special Issue on Word Sense Disambiguation,* **24 (1)**

• Lin, D., 1997. Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In: *Proceedings of ACL and EACL'97,* San Francisco

• Martínez D., E. Agirre E. and L. Màrquez L, 2002. Syntactic Features for High Precision Word Sense Disambiguation. In: *Proceedings of COLING'02,* Taipei

• Mihalcea, R. and D. Moldovan, 1999. An Automatic Method for Generating Sense Tagged Corpora. In: *Proceedings of AAAI '99,* Orlando

• Montemagni, S., S. Federici and V. Pirelli, 1996. Example-based Word Sense Disambiguation: a Paradigm-driven Approach. In: *Proceedings of EURALEX'96,* Göteborg

• Montoyo, A. and M. Palomar, 2000. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. In: *Proceedings of DEXA'00*, Greenwich

• Ng, H.T. and H.B. Lee, 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. In: *Proceedings ACL'96,* Santa Cruz, CA

• Nica, I., M.A. Martí and A. Montoyo, 2003. Automatic sense (pre-)tagging by syntagmatic patterns. In: *Proceedings of RANLP-03,* Borovets

• Pedersen, T., 2001. A decision tree of bigrams is an accurate predictor of word sense. In: *Proceedings of NAACL 2001*, Pittsburg

• Stetina, J., S. Kurohashi and M. Nagao, 1998. General WSD Method based on a Full Sentential Context. In: *Proceedings of COLING-ACL Workshop,* Montreal

• Yarowsky, D., 1992. WSD using statistical models of Roget's categories trained on large corpora, *Proceedings of COLING-92,* Nantes

• Yarowsky, D., 1995. Unsupervised WSD rivalising Supervised Methods. In: *Proceedings of ACL'95,* Cambridge