Unsupervised Learning of Ontology-Linked Selectional Preferences*

Hiram Calvo [1] and Alexander Gelbukh [1,2]

[1] Center for Computing Research, National Polytechnic Institute,
Av. Juan de Dios Bátiz s/n, esq. Av. Mendizábal, México, D.F., 07738. México
hcalvo@sagitario.cic.ipn.mx,
gelbukh@gelbukh.com; www.gelbukh.com

[2] Department of Computer Science and Engineering, Chung-Ang University,
221 Huksuk-Dong, DongJak-Ku, Seoul, 156-756, Korea

Abstract. We present a method for extracting selectional preferences of verbs from unannotated text. These selectional preferences are linked to an ontology (e.g. the hypernym relations found in WordNet), which allows for extending the coverage for unseen valency fillers. For example, if *drink vodka* is found in the training corpus, a whole WordNet hierarchy is assigned to the verb *to drink* (*drink liquor*, *drink alcohol*, *drink beverage*, *drink substance*, etc.), so that when *drink gin* is seen in a later stage, it is possible to relate the selectional preference *drink vodka* with *drink gin* (as *gin* is a co-hyponym of *vodka*). This information can be used for word sense disambiguation, prepositional phrase attachment disambiguation, syntactic disambiguation, and other applications within the approach of pattern-based statistical methods combined with knowledge. As an example, we present an application to word sense disambiguation based on the Senseval-2 training text for Spanish. The results of this experiment are similar to those obtained by Resnik for English.

1 Introduction

Selectional Preferences are patterns that measure the degree of coupling of an argument (direct object, indirect object and prepositional complements) with a verb. For example, for the verb *to drink*, the direct objects *water*, *juice*, *vodka*, and *milk* are more probable than *bread*, *ideas*, or *grass*.

In order to have a wide coverage of possible complements for a verb, it is necessary to have a very big training corpus, so that every combination of a verb and a complement be found in such a training corpus. However, even for a corpus of hundreds of millions of words, there are word combinations that do not occur in it; sometimes these word combinations are not used very frequently, or sometimes they are used often but they are not seen in certain training corpora.

* Work done under partial support of Mexican Government (CONACyT, SNI, PIFI-IPN, CGEPI-IPN), Korean Government (KIPA), and RITOS-2. The second author is currently on Sabbatical leave at Chung-Ang University.

Table 1. Non-common usages (lower occurrence values) and common usages (higher occurrence values) of word combinations of verb + WordNet synset

verb	synset	Literal English gloss	Weighted occurrences
leer	*fauna*	'read fauna'	0.17
leer	*comida*	'read food'	0.20
leer	*mensaje*	'read message'	27.13
leer	*escrito*	'read writing'	28.03
leer	*objeto_inanimado*	'read inanimate_object'	29.52
leer	*texto*	'read text'	29.75
leer	*artículo*	'read article'	37.20
leer	*libro*	'read book'	41.00
leer	*comunicación*	'read communication'	46.17
leer	*periódico*	'read newspaper'	48.00
leer	*línea*	'read line'	51.50
beber	*superficie*	'drink surface'	0.20
beber	*vertebrado*	'drink vertebrate'	0.20
beber	*lectura*	'drink reading'	0.20
beber	*sustancia*	'drink substance'	11.93
beber	*alcohol*	'drink alcohol'	12.50
beber	*líquido*	'drink liquid'	22.33
tomar	*artrópodo*	'take arthropod'	0.20
tomar	*clase_alta*	'take high_class'	0.20
tomar	*conformidad*	'take conformity'	0.20
tomar	*postura*	'take posture'	49.83
tomar	*resolución*	'take resolution'	89.50
tomar	*control*	'take control'	114.75
tomar	*acción*	'take action'	190.18

A solution for this problem is to use word classes. In this case, *water, juice*, *vodka* and *milk* belong to the class of *liquid* and can be associated with the verb *to drink*. However, not all verbs have a single class that is associated with them. For example the verb *to take* can have arguments of many different classes: *take a seat*, *take place*, *take time*, etc. On the other hand, each word can belong to more than one class. This depends not only on the sense of the word, but the main feature that has been taken into account when assigning it to a class. For example, if we consider the color of the objects, *milk* would belong to the class of white objects. If we consider physical properties, it may belong to the class of fluids or liquids. *Milk* can be *basic_food* too, for example. We can say then that the relevant classification for a word depends both on its use and the classification system being used.

To find a correlation between the usage of a noun, its sense, and the selectional preferences for the verbs, the following kind of information is needed: (1) Ontological information for a word —a word is not linked to a single class, but a whole hierarchy, and (2) information of the usage of the word in a sentences, given a verb and its specific position in the ontology.

In this paper we propose a method to extract selectional preferences that are linked to an ontology. This information is useful to solve several problems following the approach of pattern-based statistical methods combined with knowledge [1, 2].

Table 1 presents an example of the kind of information we obtain with our method. The table shows the values of argument's co-occurrence with the verb for three Spanish verbs using the WordNet hierarchy. These numbers were obtained following the methodology that is described in detail in Section 3. Note that synsets that have greater chance of being an argument for a verb have a greater value, such as *drink liquid*. In contrast, lower values indicate that a synset is less likely to be an argument for the corresponding verb (v. gr. *drink reading*, *read food* or *drink surface*). These combinations were found due to mistakes in the training corpus or due to several unrelated senses of a word. For example, *gin* can be also a *trap* that in turn is a *device*. This may lead to **drink device*. When big corpora are used for training, this noise is substantially reduced in contrast with correct patterns, allowing for disambiguation of word senses based on the sentence's main verb.

Table 1 also shows that synsets located higher in WordNet hierarchy have higher values, as they accumulate the impact of the hyponym words that are below them (see for example *communication*, *liquid* or *action*). A simple ad-hoc strategy of weighting values in WordNet's hierarchy will be described also in Section 3.

In the following sections we will show how we obtain information like that shown in Table 1, and then we will illustrate the usefulness of our method applying this information to word sense disambiguation (WSD).

2 Related Work

One of the first works on selectional preference extraction linked to WordNet senses was Resnik's [3]. It is devoted mainly to word sense disambiguation in English. Resnik assumed that a text annotated with word senses was a resource difficult to obtain, so he based his work on text tagged only morphologically. Subsequently, Agirre and Martinez [4, 5] worked linking verb usage with their arguments. In contrast with Resnik, Agirre and Martínez assumed the existence of a text annotated with word senses: Sem-Cor, in English. Other supervised WSD systems include JHU [6], which won the Senseval-2 competition, and a maximum entropy WSD system by Suarez and Palomar [7]. The first system combined, by means of a voting-based classifier, several WSD subsystems based on different methods: decision lists [8], cosine-based vector models, and Bayesian classifiers. The second system selected a best-feature selection for classifying word senses and a voting system. These systems had a score around 0.70 on the Senseval-2 tests.

We take into account that a resource such as Sem-Cor is currently not available for many languages (in particular, Spanish), and the cost of building it is high. Accordingly, we follow Resnik's approach, in the way of assuming that there is not enough quantity of text annotated with word senses. Furthermore, we consider that the WSD process must be completely automatic, so that all the text we use is automatically tagged with morphological and part-of-speech (POS) tags. Accordingly, our system is fully unsupervised.

Table 2. Selected combinations extracted from VCC

	verb	relation	noun	English gloss
1	*contar*	*con*	*permiso*	'to have permission'
2	*pintar*	<	*pintor*	'painter paints'
3	*golpear*	>	*balón*	'kick ball'
4	*solucionar*	>	*problema*	'solve problem'
5	*dar*	>	*señal*	'give signal'
6	*haber*	>	*incógnita*	'there is unknown quantity'
7	*poner*	*en*	*cacerola*	'put in pan'
8	*beber*	*de*	*fuente*	'drink from source'
9	*beber*	>	*vodka*	'drink vodka'

Previous work on unsupervised systems has not achieved the same performance as with supervised systems: Carroll and McCarty [9] present a system that uses selectional preferences for WSD obtaining 69.1% precision and 20.5% recall; Agirre and Martinez [10] present another method, this time unsupervised. They use recall as the only performance measure, reporting 49.8%; Resnik [3] achieves 40% correct disambiguation.

In the next sections we describe our method and measure its performance.

3 Methodology

In order to obtain the selectional preferences linked to an ontology, we used the hypernym relations of Spanish EuroWordNet[1] 1.0.7 (S-EWN) as ontology, and the corpus described in [11] as a training corpus (VCC). This corpus of 38 million words is supposed to combine the benefits of a virtual corpus (e.g. the web as corpus), with those of a local corpus, see details in [11].

The text was morphologically tagged using the statistical tagger TnT by Thorsten Brants [12] trained with the corpus CLiC-TALP. This tagger has a performance of over 92%, as reported in [13].

After the text was tagged morphologically, several combinations were extracted for each sentence: (1) verb + noun to the left (subject), (2) verb + noun to the right (object), and (3) verb NEAR preposition + noun. Here, + denotes adjacency, while NEAR denotes co-occurrence within a sentence. Table 2 shows an example of the information obtained in this way. The symbol > means that the noun is to the right of the verb; the symbol < means that the noun appears to the left of the verb.

Once the combinations have been extracted, the noun for each combination was looked up in WordNet and an occurrence for the corresponding synset (with every sense) was recorded. Also the occurrence was recorded for each hyperonym of each its sense. A weighting factor was used so that words higher in the hierarchy (up to the

[1] S-EWN was Developed jointly by the University of Barcelona (UB), the Nacional University of Open Education (UNED), and the Polytechnic University of Catalonia (UPC), Spain.

root *entity*) have lower impact than the words in the lower part of the hierarchy. We used the weighting factor $\frac{1}{level}$. For example, for *drink vodka* found in the text, an occurrence of the combination *drink vodka* is recorded with the weight 1, also occurrences of *drink liquor* with the weight 0.5, *drink alcohol* with 0.33, etc. are recorded. For each combination, the weights of its occurrences are accumulated (summed up).

Currently we have acquired 1.5 million of selectional preferences patterns linked to the WordNet synsets. Each pattern consists on a verb, a preposition (in some cases), and a synset. An example of the information obtained can bee seen in Figure 1. *Channel* has 6 senses listed by WordNet: *way*, *conduit*, *clear*, *conduit* (anatomic), *transmission*, *depression*, and *water*. The sense marked with the highest number of occurrences is *conduit*, while the one with fewer occurrences is *transmission*, in the sense of *channel of transmission* or *TV channel*, for example; one cannot *cross* a TV channel. Now consider *libro* 'book'; this Spanish word has five senses: *stomach*, *product*, *section*, *publication* and *work / play*. The first sense refers to the name in Spanish for an internal part of body. We can see that this is the sense with fewer occurrences (one cannot *read* an *organ*). The sense with the greatest number of occurrences is that related to *written_language*. This information can be used to disambiguate the sense of the word, given the verb with which it is used. In the next section we describe an experiment we ran to measure the performance of this method in the task of WSD.

4 Application to WSD

Senseval is a series of competitions aimed to evaluation of word sense disambiguation programs, organized by the ACL-SIGLEX. The last competition took place in 2001 (the next one being scheduled for 2004). The data for this competition are available on-line. This competition included, among 10 languages, Spanish data, to which we applied our method. The evaluation set comprises slightly more than 1,000 sentences. Each sentence contains one word, for which the correct sense, among those listed for it in WordNet, is indicated.

Our evaluation showed that 577 of 931 cases were resolved (a recall of ~62%). Of those, 223 corresponded in a fine-grained way to the sense manually annotated (precision ca. 38.5%). These results are similar to those obtained by Resnik [3] for English, who obtained on average 42.55% for the relations verb—subject and verb—object only. Note that these results are much better than random selection of senses (around 28% as reported in [3]).

4.1 Discussion

Our results are lower than those of some other WSD systems. For example, Suarez and Palomar [7] report a score of 0.702 for noun disambiguation for the same evaluation set of Senseval-2. However, their system is supervised, whereas ours is

atravesar canal: 'cross channel'

02342911n → **way** 3.00 → trough 8.83 → artifact 20.12 → unanimated_obect 37.10 → entity 37.63

02233055n → **conduit** 6.00 → way 3.00 → trough 8.83 → artifact 20.12 → unanimated_object 37.10 → entity 37.63

03623897n → **conduit** 5.00 → anatomic_structure 5.00 → body_part 8.90 → part 7.22 → entity 37.63

04143847n → **transmission** 1.67 → communication 3.95 → action 6.29

05680706n → **depression** 2.33 → geological_formation 2.83 → natural_object 14.50 → unanimated_object 37.10 → entity 37.63

05729203n → **water** 4.17 → unanimated_object 37.10 → entity 37.63

leer libro: 'read book'

01712031n → **stomach** 3.50 → internal_organ 3.00 → organ 3.08 → body_part 3.75 → part 4.35 → entity 41.51

02174965n → **product** 14.90 → creation 13.46 → artifact 34.19 → unanimated_object 36.87 → entity 41.51

04214018n → **section** 23.33 → writing 33.78 → written_language 25.40 → communication 55.28 → social_relation 43.86 → relation 42.38 → abstraction 44.18

04222100n → **publication** 16.58 → work 7.95 → product 14.90 → creation 13.46 → artifact 34.19 →unanimated_object 36.87 → entity 41.51

04545280n → **play** 4.50 → writing 33.78 → written_language 25.40 → communication 55.28 →social_relation 43.86 → relation 42.38 → abstraction 44.18

Figure 1. Ontology with usage values for the combinations in Spanish *atravesar canal* 'cross channel' and *leer libro* 'read book'. Synsets labels were translated here from Spanish to English for the reader's convenience.

unsupervised. In comparison with existing unsupervised WSD systems (i.e. [3, 9, 10]) our method has a better recall, though lower precision in some cases. The latter is due the strategy of our method that considers only verb—noun relations, when sometimes the word sense is strongly linked to the preceding noun. This is particularly true for pairs of nouns that form a single prepositional phrase. For example, in the training text the following sentence appears: *La prevalecía del principio de libertad frente al principio de autoridad es la clave de Belle Epoque* 'The prevalence of the liberty principle in contrast with the authority principle is the key of Belle Epoque'. In this case, the sense of *autoridad* 'authority' is restricted more strongly by the preceding noun, *principio* 'principle', in contrast with the main verb: *es* 'is'. To determine the sense of *autoridad* by means of the combinations *is* < *authority* and *is of authority* is not the best strategy to disambiguate the sense of this word.

In order to improve our method, it is necessary to include information on the usage of combinations of nouns. This is part of our future work.

5. Other Applications

Besides WSD, the information of selectional preferences obtained by this method can be used to solve important problems, such as syntactic disambiguation. For example, consider the phrase in Spanish *Pintó un pintor un cuadro*, lit. 'painted a painter a painting' meaning 'a painter painted a painting'. In Spanish it is possible to put the subject to the right of the verb. There is ambiguity, as it is not possible to decide which noun is the subject of the sentence. As Spanish is a language with rather free word order, even *Pintó un cuadro un pintor*, lit. 'painted a painting a painter' has the same meaning.

To decide which word is the subject (*painting* or *painter*) it is possible to consult the ontology linked with selectional preferences constructed with the method presented in this paper. First, we find statistically that the subject appears to the left of the verb in 72.6% of the times [14]. Then, searching for *un pintor pintó* 'a painter painted' returns the following chain of hypernyms with occurrence values: *painter* → *artist* 1.00 → *creator* 0.67 → *human_being* 2.48 → *cause* 1.98. Finally, the search of *un cuadro pintó* 'a painting painted' returns *scene*→ *situation* 0.42 → *state* 0.34. That is, *painter* (1.00) is more probable as subject than *painting* (0.42) for this sentence. A large-scale implementation of this method is a topic of our future work.

6. Conclusions

We have presented a method to extract selectional preferences of verbs linked to an ontology. It is useful to solve natural language text processing problems that require information about the usage of words with a particular verb in a sentence. Specifically, we presented an experiment that applies this method to disambiguate word senses. The results of this experiment show that there is still a long way to improve unsupervised WSD methods using selectional preferences; however, he have identified specific points to improve our method under the same line of pattern-based statistical methods combined with knowledge.

References

1. P. Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. TesisDoctoral, University of Pennsylvania, December (1993)
2. P. Resnik. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61 (1996) 127—159
3. P. Resnik. Selectional preference and sense disambiguation, *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C., USA, April 4-5 (1997)
4. E. Agirre, D. Martinez. Learning class-to-class selectional preferences. In: *Proceedings of the Workshop Computational Natural Language Learning* (CoNLL-2001), Toulousse, France, 6-7 july (2001)
5. E. Agirre, D. Martinez. Integrating selectional preferences in WordNet. In: *Proceedings of the first International WordNet Conference*, Mysore, India, 21-25 January (2002)

6. D. Yarowsky, S. Cucerzan, R. Florian, C. Schafer, R. Wicentowski. 2001. The Johns Hopkins SENSEVAL-2 System Description. In: Preiss and Yarowsky, eds.: *The Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, (2001) 163–166
7. A. Suárez, M. Palomar. A Maximum Entropy-based Word Sense Disambiguation System. In: Hsin-Hsi Chen and Chin-Yew Lin, eds.: *Proceedings of the 19th International Conference on Computational Linguistics*, COLING 2002, Taipei, Taiwan, vol. 2 (2002) 960—966
8. D. Yarowsky. Hierarchical decision lists for word sense disambiguation. In *Computers and the Humanities*, 34(2) (2000) 179–186
9. J. Carroll, D. McCarthy. Word sense disambiguation using automatically acquired verbal preferences. In *Computers and the Humanities*, 34(1-2), Netherlands, April (2000)
10. E. Agirre E, D. Martínez. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* EMNLP, Barcelona, Spain (2004)
11. A. Gelbukh, G. Sidorov, L. Chanona. Corpus virtual, virtual: Un diccionario grande de contextos de palabras españolas compilado a través de Internet. In: Julio Gonzalo, Anselmo Peñas, Antonio Ferrández, eds.: *Proc. Multilingual Information Access and Natural Language Processing, International Workshop,* in IBERAMIA-2002, VII Iberoamerican Conference on Artificial Intelligence, Seville, Spain, November 12-15, (2002) 7–14
12. T. Brants. TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, Washington, USA (2000)
13. R. Morales-Carrasco, A. Gelbukh. Evaluation of TnT Tagger for Spanish. In *Proc. Fourth Mexican International Conference on Computer Science*, Tlaxcala, Mexico, September 08-12 (2003)
14. J. Monedero, J. González, J. Goñi, C. Iglesias, A. Nieto. Obtención automática de marcos de subcategorización verbal a partir de texto etiquetado: el sistema SOAMAS. In *Actas del XI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* SEPLN 95, Bilbao, Spain (1995) 241—254