Acquiring Selectional Preferences from Untagged Text for Prepositional Phrase Attachment Disambiguation*

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. Extracting information automatically from texts for database representation requires previously well-grouped phrases so that entities can be separated adequately. This problem is known as prepositional phrase (PP) attachment disambiguation. Current PP attachment disambiguation systems require an annotated treebank or they use an Internet connection to achieve a precision of more than 90%. Unfortunately, these resources are not always available. In addition, using the same techniques that use the Web as corpus may not achieve the same results when using local corpora. In this paper, we present an unsupervised method for generalizing local corpora information by means of semantic classification of nouns based on the top 25 unique beginner concepts of WordNet. Then we propose a method for using this information for PP attachment disambiguation.

1 Introduction

Extracting information automatically from texts for database representation requires previously well-grouped phrases so that entities can be separated adequately. For example in the sentence *See the cat with a telescope*, two different groupings are possible: *See [the cat] [with a telescope]* or *See [the cat with a telescope]*. The first case involves two different entities, while the second case has a single entity. This problem is known in syntactic analysis as prepositional phrase (PP) attachment disambiguation.

There are several methods to disambiguate a PP attachment. Earlier methods, e.g. those described in [1, 2], showed that up to 84.5% of accuracy could be achieved using treebank statistics. Kudo and Matsumoto [3] obtained 95.77% accuracy with an algorithm that needed weeks for training, and Lüdtke and Sato [4] achieved 94.9% accuracy requiring only 3 hours for training. These methods require a corpus annotated syntactically with chunk-marks. This kind of corpora is

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

H. Calvo, A. Gelbukh. Acquiring Selectional Preferences from Untagged Text for Prepositional Phrase Attachment Disambiguation. In: F. Meziane; E. Metais (Eds.). Natural Language Processing and Information Systems (NLDB-2004). Lecture Notes in Computer Science, N 3136, Springer-Verlag, 2004, pp. 207–216.

Table 1. Occurrence examples for some verbs in Spanish

Triplet	Literal English translation	Occurrences	% of total verb occurrences
ir a {actividad}	go to {activity}	711	2.41%
ir a {tiempo}	go to {time}	112	0.38%
ir hasta {comida}	go until {food}	1	0.00%
beber {sustancia}	drink {substance}	242	8.12%
beber de {sustancia}	drink of {substance}	106	3.56%
beber con {comida}	drink with {food}	1	0.03%
amar a {agente_causal}	love to {causal_agent}	70	2.77%
amar a {lugar}	love to {place}	12	0.47%
amar a {sustancia}	love to {substance}	2	0.08%

not available for every language, and the cost to build them can be relatively high, considering the number of person-hours that are needed. A method that works with untagged text is presented in [5]. This method has an accuracy of 82.3, it uses the Web as corpus and therefore it can be slow—up to 18 queries are used to resolve a single PP attachment ambiguity, and each preposition + noun pair found in a sentence multiplies this number.

The algorithm presented in [5] is based on the idea that a very big corpus has enough representative terms that allow PP attachment disambiguation. As nowadays it is possible to have locally very big corpora, we ran experiments to explore the possibility of applying such method without the limitation of an Internet connection. We tested with a very big corpus of 161 million words in 61 million sentences. This corpus was obtained on-line from 3 years of publication of 4 newspapers. The results were disappointing—the same algorithm that used the Web as corpus yielding a recall of almost 90% had a recall of only 36% with a precision of almost 67% using the local newspaper corpus.

Therefore, our hypothesis is that we need to generalize the information contained in the local newspaper corpus to maximize recall and precision. A way for doing this is using selectional preferences: a measure of the probability of a com-

{food}:	breakfast, feast, cereal, beans, milk, etc.
{activity}:	abuse, education, lecture, fishing, hurry, test
{time}:	dawn, history, Thursday, middle age, childhood
{substance}:	alcohol, coal, chocolate, milk, morphine
{name}:	John, Peter, America, China
{causal_agent}:	lawyer, captain, director, intermediary, grandson
{place}:	airport, forest, pit, valley, courtyard, ranch

Figure 1. Examples of words for categories shown in Table 1

Table 2. Examples of Semantic Classifications of Nouns

Word	English translation	Classification
rapaz	predatory	activity
rapidez	quickness	activity
rapiña	prey	shape
rancho	ranch	place
raqueta	racket	thing
raquitismo	rickets	activity
rascacielos	skyscraper	activity
rasgo	feature	shape
rastreo	tracking	activity
rastro	track	activity
rata	rat	animal
ratero	robber	causal agent
rato	moment	place
ratón	mouse	animal
raya	boundary	activity
	manta ray	animal
	dash	shape
rayo	ray	activity
raza	race	grouping
razón	reason	attribute
raíz	root	part
reacción	reaction	activity
reactor	reactor	thing
real	real	grouping
realidad	reality	attribute
realismo	realism	shape
realización	realization	activity
realizador	producer	causal agent

plement to be used for certain verb, based on the semantic classification of the complement. This way, the problem of analyzing *I see the cat with a telescope* can be solved by considering *I see {animal} with {instrument}* instead.

For example, to disambiguate the PP attachment for the *Spanish* sentence *Bebe de la jarra de la cocina* '(he) drinks from the jar of the kitchen' selectional preferences provide information such as *from {place}* is an uncommon complement for the verb *bebe* 'drinks', and thus, the probability of attaching this complement to the verb *bebe*, is low. Therefore, it is attached to the noun *jarra* yielding *Bebe de [la jarra de la cocina]* '(he) drinks [from the jar of the kitchen]'.

Table 1 shows additional occurrence examples for some verbs in Spanish. From this table it can be seen that the verb *ir* 'to go' is mainly used with the complement *a {activity}* 'to {activity}'. Less used combinations have almost zero occurrences, such as *ir hasta {food}* lit. ‘go until food’. The verb *amar* ‘to love’ is often used with the preposition *a* ‘to’.

In this paper, we propose a method to obtain selectional preferences information such as that shown in Table 1. In Section 2, we will discuss briefly related work on selectional preferences. Sections 3 to 5 explain our method. In Section 6, we present an experiment and evaluation of our method applied to PP attachment disambiguation, and finally we conclude the paper.

2 Related work

The terms *selectional constraints* and *selectional preferences* are relatively new, although similar concepts are present in works such as [6] or [7]. One of the earliest works using these terms was [8], where Resnik considered selectional constraints to determine the restrictions that a verb imposes on its object. Selectional constraints have rough values, such as whether an object of certain type can be used with a verb. Selectional preferences are graded and measure, for example, the probability that an object can be used for some verb [9]. Such works use a shallow parsed corpus and a semantic class lexicon to find selectional preferences for word sense disambiguation.

Another work using semantic classes for syntactic disambiguation is [10]. In this work, Prescher *et al.* use an EM-Clustering algorithm to obtain a probabilistic lexicon based in classes. This lexicon is used to disambiguate target words in automatic translation.

A work that particularly uses WordNet classes to resolve PP attachment is [2]. In this work, Brill and Resnik apply the Transformation-Based Error-Driven Learning Model to disambiguate the PP attachment, obtaining an accuracy of 81.8%. This is a supervised algorithm.

As far as we know, selectional preferences have not been used in unsupervised models for PP attachment disambiguation.

3 Sources of Noun Semantic Classification

A semantic classification for nouns can be obtained from existing WordNets, using a reduced set of classes corresponding to the unique beginners for WordNet nouns described in [11]. These classes are: activity, animal, life_form, phenomenon, thing, causal_agent, place, flora, cognition, process, event, feeling, form, food, state, grouping, substance, attribute, time, part, possession, and motivation. To these unique beginners, *name* and *quantity* are added. Name corresponds to capitalized words not found in the semantic dictionary and Quantity corresponds to numbers.

Since not every word is covered by WordNet and since there is not a WordNet for every language, the semantic classes can be alternatively obtained automatically from Human-Oriented Explanatory Dictionaries. A method for doing this is explained in detail in [12]. Examples of semantic classification of nouns extracted from the human-oriented explanatory dictionary [13] using this method are shown in Table 2.

4 Preparing Sources for Extracting Selectional Preferences

Journals or newspapers are common sources of great amounts of medium to good quality text. However, usually these media exhibit a trend to express several ideas in little space.

Y ahora, cuando
(el mundo) **está** gobernado por (las leyes del mercado), cuando
(lo determinante en la vida) **es**
comprar o
vender, sin
fijarse en <los que
carecen de todo>,
son fácilmente **comprensibles** <las razones de
<la ola de publicidad global que
convenció <a los posibles compradores de servicios y regalos > de que
había (grandes razones) para
celebrar> y
como les **pareciese** poco (el fin de año)
se **lanzaron** a
propagar (el fin del siglo y del milenio)

Literal English translation:

And now, when
the world is governed by market's laws, when
what **determines** life **is**
to buy or
to sell without
taking into account those that
don't **have** anything,
easily understandable **are** the reasons for
the global publicity wave that
convinced the possible buyers of services and gifts that
there were great reasons to
celebrate, and
as the end of the year **was** not enough for them,
they **launched** themselves
to propagate the end of the century and the millennium

Figure 2. Example of a very long sentence in a style typically found in journals.
( ) surround simple NPs; < > surround NP subordinate clauses, **verbs** are in boldface.

```
PREP V ,
V ADV que
, PRON V
V PREP N , N V
V PREP N , N PRON V
V PREP N V
V de que
CONJ PRON V
PREP DET que N
N que V
, donde
N , N
CONJ N N V
CONJ N PRON V
CONJ N V
PREP DET V
, N V
N , que V
N , CONJ que
N que N PRON V
CONJ PRON que V V
```

Figure 3. Delimiter patterns.
V: verb, PREP: preposition, CONJ: conjunction, DET: determiner,
N: noun, lowercase are strings of words

This causes sentences to be long and full of subordinate sentences, especially for languages in which an unlimited number of sentences can be nested. Because of this, one of the first problems to be solved is to break a sentence into several sub-sentences. Consider for example the sentence shown in Figure 2—it is a single sentence, extracted from a Spanish newspaper.

We use two kinds of delimiters to separate subordinate sentences: delimiter words and delimiter patterns. Examples of delimiter words are *pues* 'well', *ya que* 'given that', *porque* 'because', *cuando* 'when', *como* 'as', *si* 'if', *por eso* 'because of that', *y luego* 'and then', *con lo cual* 'with which', *mientras* 'in the meantime', *con la cual* 'with which' (feminine), *mientras que* 'while'. Examples of delimiter patterns are shown in Figure 3. These patterns are POS based, so the text was shallow-parsed before applying them.

The sentence in Figure 2 was separated using this simple technique so that each sub-sentence lies in a different row.

5 Extracting Selectional Preferences Information

Now that sentences are tagged and separated, our purpose is to find the following syntactic patterns:

1. Verb $_{\text{NEAR}}$ Preposition $_{\text{NEXT\_TO}}$ Noun
2. Verb $_{\text{NEAR}}$ Noun
3. Noun $_{\text{NEAR}}$ Verb
4. Noun $_{\text{NEXT\_TO}}$ Preposition $_{\text{NEXT\_TO}}$ Noun

Patterns 1 to 3 will be referred henceforth as *verb patterns*. Pattern 4 will be referred as a *noun* or *noun classification pattern*. The $_{\text{NEAR}}$ operator implies that there might be other words in-between. The operator $_{\text{NEXT TO}}$ implies that there are no words in-between. Note that word order is preserved, thus pattern 2 is different of pattern 3. The results of these patterns are stored in a database. For verbs, the lemma is stored. For nouns, its semantic classification, when available through Spanish WordNet, is stored. As a noun may have several semantic classifications, due to, for example, several word senses, a different pattern is stored for each se-

Table 3. Semantic patterns information extracted from Sentence in Figure 2

Words	Literal translation	Pattern
gobernado, *por*, *ley*	governed, by, law	*gobernar*, *por*, cognition
gobernado, *de*, *mercado*	governed, of, market	*gobernar*, *de*, activity thing
es, *en*, *vida*	is, in, life	*ser*, *en*, state life_form causal_agent attribute
convenció, *a*, *comprador*	convinced, to, buyer	*convencer*, *a*, causal_agent
convenció, *de*, *servicio*	convinced, of, service	*convencer*, *de*, activity process possession thing grouping
pareciese, *de*, *año*	may seem, of, year	*parecer*, *de*, cognition time
lanzaron, *de*, *año*	released, of, year	*lanzar*, *de*, cognition time
propagar, *de*, *siglo*	propagate, of, century	*propagar*, *de*, cognition time
propagar, *de*, *milenio*	propagate, of, millennium	*propagar*, *de*, cognition time
ley, *de*, *mercado*	law, of, market	cognition, *de*, activity thing
ola, *de*, *publicidad*	wave, of, publicity	event, *de*, activity cognition
comprador, *de*, *servicio*	buyer, of, service	causal_agent, *de*, activity process possession thing grouping
fin, *de*, *año*	end, of, year	place cognition event time, *de*, cognition time
fin, de, *siglo*	end, of, century	place cognition event time, *de*, cognition time

mantic classification. For example, see Table 3. This table shows the information extracted for the sentence of Figure 2.

Once this information is collected, the occurrence of patterns is counted. For example, the last two rows in Table 3, *fin, de, año* and *fin, de, siglo* add 2 of each of the following occurrences: place of cognition, cognition of cognition, event of cognition, time of cognition, place of time, cognition of time, event of time, and time of time. An example of the kind of information that results from this process is shown in Table 1. This information is used then as a measure of the selectional preference that a noun has to a verb or to another noun.

6 Experimental Results

The procedure explained in the previous sections was applied to a corpus of 161 million words comprising more than 3 years of articles from four different Mexican newspapers. It took approximately three days on a Pentium IV PC to obtain 893,278 different selectional preferences for verb patterns (patterns 1 to 3) for 5,387 verb roots, and 55,469 different semantic selectional preferences for noun classification patterns (pattern 4).

6.1 PP Attachment disambiguation

In order to evaluate the quality of the selectional preferences obtained, we tested them on the task of PP attachment disambiguation. Consider the first two rows of Table 4, corresponding to the fragment of text *governed by the laws of the market*. This fragment reported two selectional preferences patterns: *govern by {cognition}* and *govern of {activity/thing}*. With the selectional preferences obtained, it is possible to determine automatically the correct PP attachment: values of co-occurrence for *govern of {activity/thing}* and *{cognition} of {activity/thing}* are compared. The highest one sets the attachment.

Formally, to decide if the noun N_2 is attached to its preceding noun N_1 or is attached to the verb V of the local sentence, the values of frequency for these attachments are compared using the following formula [14]:

$$freq(X,P,C_2) = \frac{occ(X,P,C_2)}{occ(X) + occ(C_2)}$$

where X can be V, a verb, or C_1, the classification of the first noun N_1. P is a

Table 4. Results of the PP attachment disambiguation using selectional preferences

file	#sentences	words	average words per sentence	kind of text	precision	recall
n1	252	4,384	17.40	news	80.76%	75.94%
t1	74	1,885	25.47	narrative	73.01%	71.12%
d1	220	4,657	21.17	sports	80.80%	81.08%
total:	546	10,926		**average**:	78.19%	76.04%

preposition, and C_2 is the classification of the second noun N_2. If $freq(C_1, P, C_2) > freq(V, P, C_2)$, then the attachment is decided to the noun N_1. Otherwise, the attachment is decided to the verb V. The values of $occ(X, P, C_2)$ are the number of occurrences of the corresponding pattern in the corpus. See Table 1 for examples of verb occurrences. Examples of noun classification occurrences taken from the Spanish journal corpus are: *{place} of {cognition}*: 354,213, *{place} with {food}*: 206, *{place} without {flora}*: 21. The values of *occ(X)* are the number of occurrences of the verb or the noun classification in the corpus. For example, for *{place}* the number of occurrences is 2,858,150.

6.2 Evaluation

The evaluation was carried on 3 different files of LEXESP corpus [15], containing 10,926 words in 546 sentences. On average, this method achieved a precision of 78.19% and a recall of 76.04%. Details for each file processed are shown in Table 4.

7 Conclusions and Future Work

Using selectional preferences for PP attachment disambiguation yielded a precision of 78.19% and a recall of 76.04%. These results are not as good as the ones obtained with other methods, such as an accuracy of 95%. However, this method does not require any costly resource such as an annotated corpus, nor an Internet connection (using the web as corpus); it does not even need the use of a semantic hierarchy (such as WordNet), as the semantic classes can be obtained from Human-Oriented Explanatory Dictionaries, as it was discussed in Section 3.

We found also that, at least for this task, applying techniques that use the Web as corpus to local corpora reduces the performance of these techniques in more than 50%, even if the local corpora are very big.

In order to improve results for PP attachment disambiguation using selectional preferences, our hypothesis is that instead of using only 25 fixed semantic classes, intermediate classes can be obtained by using a whole hierarchy. In this way, it would be possible to have a flexible particularization for terms commonly used together, i.e. collocations, such as *fin de año* 'end of year', while maintaining the power of generalization. Another point of further developments is to add a WSD module, so that not every semantic classification for a single word is considered, as it was described in Section 5.

References

1. Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos. A Maximum Entropy Model for Prepositional Phrase Attachment. In *Proceedings of the ARPA Human Language Technology Workshop*, 1994, pp. 250-255
2. Eric Brill and Philip Resnik. A Rule-Based Approach To Prepositional Phrase Attachment Disambiguation, In *Proceedings of COLING-1994*, 1994.

3. T. Kudo and Y. Matsumoto. Use of Support Vector Learning for Chunk Identification. In *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, Portugal, 2000

4. Dirk Lüdtke and Satoshi Sato. Fast Base NP Chunking with Decision Trees - Experiments on Different POS Tag Settings. In Gelbukh, A. (ed) *Computational Linguistics and Intelligent Text Processing*, Springer LNCS, 2003, pp. 136-147

5. Hiram Calvo and Alexander Gelbukh. Improving Prepositional Phrase Attachment Disambiguation Using the Web as Corpus, In A. Sanfeliu and J. Shulcloper (Eds.) *Progress in Pattern Recognition*, Springer LNCS, 2003, pp. 604-610

6. Weinreich, Uriel. *Explorations in Semantic Theory*, Mouton, The Hague, 1972.

7. Dik, Simon C., *The Theory of Functional Grammar. Part I: The structure of the clause*. Dordrecht, Foris, 1989.

8. Philip Resnik. Selectional Constraints: An Information-Theoretic Model and its Computational Realization, *Cognition*, 61:127-159, November, 1996.

9. Philip Resnik. Selectional preference and sense disambiguation, presented at the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, held April 4-5, 1997 in Washington, D.C., USA in conjunction with ANLP-97.

10. Detlef Prescher, Stefan Riezler, and Mats Rooth. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarland University, Saarbrücken, Germany, July-August 2000. ICCL.

11. Miller, G. WordNet: An on-line lexical database, In *International Journal of Lexicography*, 3(4), December 1990, pp. 235-312

12. Calvo, H. and A. Gelbukh. Extracting Semantic Categories of Nouns for Syntactic Disambiguation from Human-Oriented Explanatory Dictionaries, In Gelbukh, A. (ed) *Computational Linguistics and Intelligent Text Processing*, Springer LNCS, 2004.

13. Lara, Luis Fernando. *Diccionario del español usual en México*. Digital edition. Colegio de México, Center of Linguistic and Literary Studies, 1996.

14. Volk, Martin. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceeding of Corpus Linguistics* 2001. Lancaster: 2001

15. Sebastián, N., M. A. Martí, M. F. Carreiras, and F. Cuestos. *Lexesp, léxico informatizado del español*, Edicions de la Universitat de Barcelona, 2000