

PP attachment ambiguity resolution with corpus-based pattern distributions and lexical signatures

Nuria Gala¹ and Mathieu Lafourcade²

¹DELIC, Univ. Provence - 29, av. R. Schuman - F-13621 Aix de Provence Cedex 1
nuria.gala@up.univ-aix.fr

²LIRMM, Univ. Montpellier - 161, rue Ada -F-34392 Montpellier Cedex 5
mathieu.lafourcade@lirmm.fr

ABSTRACT

We propose a method mixing unsupervised learning of lexical pattern frequencies with semantic information which aims at improving the resolution of PP attachment ambiguity. Using the output of a robust parser, i.e. the set of all possible attachments for a given sentence, we query the Web and obtain statistical information about the frequencies of the attachments distributions as well as lexical signatures of the terms on the patterns. All this information is used to weight the dependencies yielded by the parser and eventually to choose of the most probable attachment.

1. INTRODUCTION

The problem of identifying right PP attachments, especially when there is inherent semantic ambiguity, is a crucial issue for NLP applications, particularly when semantic interpretation is required (e.g. in question-answering, translation systems, etc.). Thus, the following example with the pattern *VNP PP* (or *VNP N*) "He sees a girl with a telescope" can have two different interpretations, depending on the attachment of the PP : (sees (a girl with a telescope)) and (sees (a girl) (with a telescope)).

In recent years, many researchers have been working on the subject of PP attachment ambiguity resolution. A variety of solutions have been proposed, going from the use of semantic information extracted from a dictionary [Jensen and Binot, 87] to probability-based approaches: lexical association scores [Hindle and Rooth, 93], transformation-based learning [Brill and Resnik, 94], etc. Methods already combining probabilistic with semantic information lead to better results [Stetina and Nagao, 97]. However, these methods usually require very large annotated corpora

(i.e. syntactically annotated and semantically disambiguated) often unavailable.

For other languages than English, the number of experiences conducted on this issue is fewer than for English. For French, [Gaussier and Cancedda, 01] propose a statistical model that integrates different resources (including semantic information). [Bourigault and Fabre, 01] present a distributional method to solve the ambiguities of syntactic analysis based on a productivity measure which identifies different levels of lexical dependency. Also, [Aït and Gala, 03] use a weighted subcategorisation lexicon obtained by calculating the frequencies of the PP attachment patterns within the Web.

In the following sections, we describe our approach which combines unsupervised learning of lexical frequencies, as in [Aït and Gala, 03], with semantic information. Section 2 describes the output of the parser and gives an overview of the gathering of statistical information (frequencies of PP attachments). Section 3 presents the lexical signatures related to the terms in the patterns. Before concluding, section 4 discusses the method for scoring the attachments and points out the experiments undertaken.

2. Automatic learning of PP distribution patterns

To obtain the statistical information about the distributions of the patterns in a very large corpora, we query the Web with the PP attachment dependencies yielded by a robust parser.

2.1 The parser output

The parser we use is the Xerox Incremental Parser (XIP), a rule-based incremental parsing framework for the analysis of raw text [Aït-Mokhtar et al, 01]. The grammars for French produce an accurate linguistic analysis with significant precision and recall

rates (i.e. for subject, P=93,45%, R=89,36%).

The output for a given sentence consists on the set of chunks¹ and a list of dependencies. Figure 1 shows the analysis for the sentence "*Elle achète des vêtements pour ses enfants.*" (Eng. She buys clothes for her children.):

```
SUBJ(acheter, il)
OBJ(acheter, vêtement)
VMOD(acheter, enfant)
NMOD(vêtement, enfant)
PREPOBJ(enfant, pour)
DETERM(enfant, son)
DETERM(vêtement, un)
0>GROUPE{SC{NP{il} FV{acheter}} NP{un
vêtement} PP{pour NP{son enfant}} .}
```

Fig.1: XIP output (with lemmas).

A dependency is a syntactic relation between two headwords of two chunks, i.e. a noun and a verb for subject and verb modifier, two nouns or a noun and an adjective for a noun modifier, etc. Dependencies show binary relations; for prepositionnal attachment, the relations with the three elements (X, P, N) can be calculated through *VMOD* or *NMOD* and *PREPOBJ* dependencies. Thus *VMOD(acheter, enfant)* and *PREPOBJ(enfant, pour)* give *(acheter, pour, enfant)*.

The parser is deterministic for calculating all the dependencies (one solution is proposed among the eventual possibilities). Prepositionnal attachment is the only exception because syntactic rules (with very few lexical or semantic information) are not able to take a decision concerning right PP-attachments. In this case, recall is favoured and all the potential attachments are extracted.

For the previous example, two attachments are thus extracted instead of one:

```
(acheter, pour, enfant)
(vêtement, pour, enfant)
```

Fig.2: Prepositionnal phrase attachments. buy, for, child - cloth, for, child

2.2 Querying the Web

As in [Aït and Gala, 03], ambiguous dependencies (i.e. those where a same noun is attached to two different headwords) are transformed into queries for the

¹SC (sentence clauses), NP (noun phrases), PP (prepositional phrases), AP (adjective phrases), FV (finite verb clauses).

Web and a measure of frequency is calculated for each frame. The three elements of the dependency (X, P, N) are used in the query, that is: X , the potential head of the dependency (a noun or a verb or an adjective); P , the preposition and N , the noun to be attached.

Each dependency concerning the PP attachment is thus transformed into a query for the Web and for each one 10 URLs are automatically retrieved using Google. The result of this process is a new collection of corpora which is parsed to obtain a higher number of PP attachments. The aim of parsing the collected corpora is to avoid wrong configurations when calculating the scores, i.e. words appearing together in a corpus but not linked by a syntactic dependency.

Thus, syntactic co-occurrence probabilities (i.e. weights for a given syntactic pattern) are measured from the frequencies of words co-occurring in the same syntactic dependency relation (attachments already yielded by the parser) coming from the large corpora obtained by harvesting the Web.

This measure, that we call *SCS* (syntactic co-occurrence score), is determined by the ratio between the number of occurrences (in the corpus) of the whole dependency (X, P, N) and the number of occurrences of a subcategorization frame (X, P):

$$SCS(X, P, N) = \frac{\#(X, P, N)}{\#(X, P)} \quad (1)$$

As a result, we obtain a database scoring the probability of co-occurrence of the three words of a pattern. Such a measure permits to significantly increase the precision rate of PP-attachment dependencies, as shown in [Aït and Gala, 03]. However, when there is inherent semantic ambiguity, this probabilistic information is not significant to resolve PP-attachment ambiguity. Especially, with a pattern $X N_1 P N_2$, where N_1 cannot be optional, the probability to find $N_1 P N_2$ would be higher than the one to find $X P N_2$ even though the correct attachment is indeed $X P N_2$ (but cannot or rarely be found as it in the corpus).

Another bottleneck with the *SCS* measure concern particular constructions with very few occurrences in the corpus. In this case, there is not significant statistical information to score the attachments. For instance, we have in French, the sentence "*Le résultat courant exprime la rentabilité de la société en intégrant les excédents dégagés par l'exploitation (...).*" (Eng. The current result shows the profitability of the society by including the surplus obtained by exploiting (...)) where *par l'exploitation* although at-

tached to *dégagés* would be found in the corpus with very few occurrences and the pattern *excédents par l'exploitation* would not be found at all.

SCS(dégagés par l'exploitation) = 240/102.000 = 0.0023

SCS(excédents par l'exploitation) = 0/257 = 0

SCS(intégrant par l'exploitation) = 0/632 = 0

All those reasons make us think that combining this SCS measure with lexical signatures that reflect more thematic proximities between terms (or chunks) would improve PP attachment resolution.

3. Lexical Signatures

A lexical signature of a term t is a set of weighted terms that allows to characterize *thematically* this term. We could roughly consider that the signature describes the semantic field of the term. The signature of a term can be built in several ways, but one approach is to pick up surrounding words in a given corpus. For example, we can have the following signatures (computed from Le Monde corpus).

For the term *enfant* (Eng. child): enfant: ("femme" 2.37) ("personne âgé" 1.62) ("parent" 1.12) ("deux opéra" 1) ("Milhaud" 1.0) ("batelier" 1) ("être en partie carboniser" 1) ("aucun guide touristique" 1) ("me adresser" 1) ("ex-Yougo" 1) ("spectateur contraint" 1) ("jeune" 0.90) ("garde" 0.83) ("vieillard" 0.83) ("Naf - Naf" 0.81) ("deuxième" 0.77) ("vêtement" 0.77) ("le oeil plein" 0.76) ("prodige" 0.76) ("illustre" 0.76) ("tombe" 0.76) ...

For the term *vêtement* (Eng. clothes): vêtement: ("un marque italien" 1) ("fabricant choletais" 1) ("le sous-vêtement" 0.67) ("enfant" 0.41) ("le prêt-à-porter" 0.38) ("couette" 0.25) ("se accompagner" 0.23) ("table" 0.23) ("exquis" 0.20) ("spectaculaire" 0.19) ("sentier" 0.19) ("notre culture" 0.19) ("son gamme" 0.19) ("se répartir" 0.18) ("le chaussure" 0.16) ("blondinet" 0.16) ("ce entrée" 0.16) ("appât" 0.14285714285714285) ("Chinois" 0.13) ("le licence" 0.12) ("le enfant" 0.12) ("Naf - Naf" 0.11) ("le brochette" 0.10) ("le firme" 0.08) ("client" 0.07) ("marchandise" 0.07) ("Albert SA" 0.07) ("détenir" 0.04) ...

3.1 Comparing signatures

Let us define $Sim(A, B)$ as one possible *similarity* measures between two signatures A et B, often used in information retrieval. We can express this function as the scalar product of their vector divided by the product of their norm. Then, we define an *angular distance* D_A between two signatures A and B as:

$$D_A(A, B) = \arccos(Sim(A, B))$$

$$\text{with } Sim(A, B) = \cos(\widehat{A, B}) = \frac{A \cdot B}{\|A\| \times \|B\|} \quad (2)$$

Intuitively, this function constitutes an evaluation of the *thematic proximity* and is the measure of the angle between the two signatures. We would generally and quite naively consider that, for a distance $D_A(A, B) \leq \frac{\pi}{4}$, (i.e. less than 45 degrees) A and B are thematically close and share many terms. For $D_A(A, B) \geq \frac{\pi}{4}$, the thematic proximity between A and B would be considered as loose. Around $\frac{\pi}{2}$, they have almost no relation.

In practice, the actual values of the distance function highly depend of the underlying corpus. The distribution of distances might differ drastically if signatures have been computed with a corpus of free texts, or of texts belonging to a specific domain (like technical documentation), or from general dictionnaires. A better practice is to actually compare an angle to the mean angle between objects of the collection.

D_A is a real distance function. As such, it verifies the properties of reflexivity, symmetry and triangular inequality.

We can have, for example, the following angles:

$$\begin{aligned} D_A(\langle child \rangle, \langle child \rangle) &= 0^\circ & D_A(\langle clothes \rangle, \langle child \rangle) &= 70^\circ \\ D_A(\langle to buy \rangle, \langle child \rangle) &= 85^\circ & D_A(\langle clothes \rangle, \langle to buy \rangle) &= 76^\circ \end{aligned}$$

The first value as a straightforward interpretation due to the reflexivity of the distance. There are more mutual information between *clothes* and *child* than between any other two terms. From our corpus, which is not specific, the angle values are generally quite high.

We focus on the angle, because it provide a real mathematical distance (to be opposed to the similarity function). A second reason, is that the angle is more discriminate to small angle variations for high value of mutual information (when the cosine is close to 1).

To ensure a normalized scoring, we do invert the definition domain of the angular distance in a linear way:

$$M_S(A, B) = 1 - \frac{2}{\pi} D_A(A, B) \quad (3)$$

We call *MIS* (mutual information score), the application of the above formula on the dependency X, P ,

N . Depending on the available chunks (either $X P$ or only X) provided by the chunk analyzer, we do have:

$$\begin{aligned} MIS(X, P, N) &= M_S(X.P, N) && \text{if } X.P \in \mathcal{C} \\ MIS(X, N) &= M_S(X, N) && \text{otherwise} \end{aligned} \quad (4)$$

For the sentence "*Elle achète des vêtements pour ses enfants*", we have the following attachments: "*acheter pour ses enfants*" or "*vêtements pour ses enfants*". The MIS are respectively:

$$\begin{aligned} MIS(\langle \textit{buy}, \textit{child} \rangle) &= 0.05 \\ MIS(\langle \textit{clothes}, \textit{child} \rangle) &= 0.22 \end{aligned}$$

3.2 Building signatures

For a given word w , we build its signature over the corpus \mathcal{C} the following way. We consider a window of δ terms before and δ terms after the target word, at the paragraph level, which have been processed beforehand through a chunk analyzer. In our experiments, we empirically set δ to 10. The terms before w are noted t_{-1}, \dots, t_{-10} , those after t are noted t_1, \dots, t_{10} . Those terms are under a lemmatized form, possibly syntactically disambiguated, when several parts of speech are eligible. Terms appearing before and after the target terms are treated symmetrically at the exception of right-hand AP (adjectival phrase) attachments that are collated the previous NP chunk. For example:

NP(missile) AP(américain)
adds *NP (missile américain)*

We then obtain, as elements of indexation, either isolated terms of noun phrases. Dealing with such chunks multiplies the possible items but offers a great increase in precision, especially when confronted with technical compound terms. Chunks can be also complex verbal phrases like:

"difficile" + "être tellement difficile"
"jeune" + "être parfois très jeune" "saccager" + "avoir saccagé"

We have the following notations: \mathcal{T} as the set of all terms that occur in the surrounding of w . The scalar $d \in [1, 10]$ is the distance between $t \in \mathcal{T}$ and w . The scalar $\#t$ is the number of occurrences of t in \mathcal{C} . We construct the signature $V(w)$ as a vector of all lemmatized terms or chunks of the corpus \mathcal{C} : $\langle w_1, \dots, w_n \rangle$.

$$V(w) = \sum_{t \in \mathcal{T}} \frac{1}{d} \times \frac{1}{1 + \log(\#t)} \times V_0(t) \quad (5)$$

If $V(t_i)$ corresponds to the i th term of the corpus, then it is initialized to the boolean vector where all components are 0 but the i th which is 1:

$$\begin{aligned} \mathcal{C} &= \{t_1, \dots, t_i, \dots, t_n\} \\ V_0(t_i) &= \langle 0_1, \dots, 1_i, \dots, 0_n \rangle \end{aligned} \quad (6)$$

A term t participates more to a signature if it is close to the target term, although its weight is tampered if it has many occurrences in the corpus. A very frequent term is less relevant than a rare one.

We shorten signatures to the first highest 500 items. Shortening vectors is due to efficiency consideration, but the loss of information is negligible (less than 2% in average). We obtain signatures that are reminiscent of the saltonian vectors computed for documents [Salton and MacGill 1983]. The main difference here is that that vector are computed for terms (or chunks) of the corpus.

Such a way, we do obtain for each term of the corpus a *first generation* signature. To ensure, that each signature has a higher recall, we iteratively augment then. An augmentation process step from *generation* n to *generation* $a + 1$ is simply a weighted sum of all signatures of the terms contained in the signature of t .

$$\begin{aligned} V_n(t) &= \langle w_1, \dots, w_i, \dots, w_n \rangle \\ V_{n+1}(t) &= \sum_{k \in [1, n]} w_k \times V_n(t_k) \end{aligned} \quad (7)$$

Each vector is normalized between iterations, i.e. all vectors have the same norm and then only the proportion of their components is relevant when comparing two vectors. The process is convergent, and vectors stabilize quickly after roughly 3 iterations. The augmentations process ensures that the probability of having two vectors in the same semantic field but that share no common term is very low. Without the augmentation, semantic fields that are lexically dense and then might have many quasi-synonyms for terms may "produce" vectors with not much in common. The iterative process of augmentation is quite similar in spirit to what happens in LSA [Deerwester et al, 90] when computing proper vectors and then reducing the dimension of vectors.

4. Scoring and Experiments

4.1 Scoring Ratio and Confidence

For the two scoring methods (SCS and MIS), we compare the score for both attachments (a_1 and a_2), and we compute a ratio $score(a_1)/score(a_2)$. A value below 0 implies that the second attachment is found as more likely than the first. In the following example, both scorings agree on the second attachment.

SCS (acheter pour, enfant) / SCS(vêtement pour, enfant) = $0.286 / 0.55 = 0.51$

MIS (acheter, enfant) / MIS(vêtement, enfant) = $0.05 / 0.22 = 0.22$

When the ratio is close to 1, then the scoring is weak as a decision process. The interesting case is when the two scorings do not agree on the same attachment. As an empirical approach, we retain the attachment for which the confidence is the highest. The *confidence* of a given score is defined as follows:

$$\begin{aligned} Conf(score) &= \frac{1}{score} && \text{if } score < 0 \\ Conf(score) &= score && \text{otherwise} \end{aligned} \quad (8)$$

For example :

$$SCS(X_1P_1, N_1)/SCS(X_2P_2, N_2) = 0.3$$

$$Conf(0.3) = 0.333$$

$$MIS(X_1, N_1)/MIS(X_2, N_2) = 2.8$$

$$Conf(2.8) = 2.8$$

In this case, we retain the attachment proposed by the SCS as its confidence value is higher than with the MIS. In this approach, we suppose that both scorings are of equal quality. This is strong assumption which may be a limiting factor for our experiments.

4.2 Experiments

We have conducted our experiments with a test corpus (T_c) from the French newspaper Le Monde of 10.002 words (425 sentences, 98 paragraphs). From this corpus, 2.444 ambiguous attachments have been extracted by the parser and transformed into queries for the Web. An average of 6 attachments per sentence as found by the parser, but not necessarily for the same head. For a given head, we found out around 2.2 attachments in average.

We have also used a learning corpus (L_c) from Le Monde of 510.969 words (21.048 sentences, 2.178

paragraphs). This corpus has been used to extract the signatures.

Experiments are still under way, but we can already estimate the following figures. The precision for PP attachment with the first statistical method only is around 75%. With both method, the percentage of ambiguous attachments when the scoring is divergent is roughly of 8%. The choice based on the confidence allows to select in 70% of the cases the proper attachment. Thus, the precision increases from 75% to 80.6% ($75 + 8 \times 0.7$) by combining both methods.

5. Conclusion and further work

This paper addresses the issue of combining two kind of information, statistical and lexical, for improving PP attachment disambiguation. We presented a ratio method that allow us to overcome the issue is different scoring distribution or value domain. In particular, we define a simple evaluation of the confidence that can be attached to the scoring to be able to select (as a heuristic) the proper attachment when scorings are divergent. As such, our approach presents a general framework that can be extended to more scoring methods. Among other criteria that should be addressed, a specific task of WSD (Word Sense Disambiguation) that would be undertaken holistically with the attachment resolution could highly improve the system performance for highly polysemous terms. Adding semantic features to terms and evaluating agreements might certainly be another research path, but by itself the construction of such resources is difficult. The increase in the training corpus size, would by itself improve performance but would eventually reach its own limits.

References

- [Ait-Mokhtar and Chanod and Roux2001] Aït-Mokhtar S., J.-P. Chanod, C. Roux 2001. *A Multi-Input Dependency Parser*. In Proceedings of the Seventh IWPT (International Workshop on Parsing Technologies), Beijing, China, 2001.
- [Ait-Mokhtar and Gala2003] Aït-Mokhtar, S. and Gala, N. 2003. *Lexicalising a robust parser grammar using the WWW*. In Proc. of Conference on Corpus Linguistics, Lancaster, UK.
- [Deerwester et al1990] Scott C. Deerwester and Susan T. Dumais and Thomas K. Landauer and George W. Furnas and Richard A. Harshman 1990. *Indexing by Latent Semantic Analysis*. In Journal of the American Society of Information Science, 6/40, pp. 391-407.

- [Fabre and Bourigault2001] Fabre, C. and Bourigault, D. 2001. *Linguistic clues for corpus-based acquisition of lexical dependencies*. In Proc. Corpus Linguistics, Lancaster, UK.
- [Brill and Resnik1994] Brill, E. and Resnik, P. 1994. *A rule-based approach to prepositional phrase attachment disambiguation*. In Proc. 15th International Conference on Computational Linguistics, COLING-94, Kyoto, Japan.
- [Gala and Lafourcade 2005] Gala, N. and Lafourcade, M. 2005. *Combining corpus-based pattern distributions with lexical signatures for PP attachment ambiguity resolution*. Proc. SNLP-05, 6th Symposium on Natural Language Processing. Chiang Rai, Thailand.
- [Gale 1992] Gale W., K. W. Church, and D. Yarowsky, 1992. *A Method for Disambiguating Word Senses in a Large Corpus.*, Computers and the Humanities, 26:415-43.
- [Gaussier and Cancedda2001] Gaussier, E. and Cancedda, N. 2001. *Probabilistic models for PP attachment resolution and NP-analysis*. In Proc. 15th ACL-01, Computational Natural Language Learning Workshop, CoNLL-01, pp. 45-52, Toulouse, France.
- [Hartrumpf1999] Hartrumpf, S. 1999. *Hybrid Disambiguation of Prepositional Phrase Attachment and Interpretation*. In Proc. of the joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99), pp. 111-120. College Park, Maryland, USA.
- [Hindle and Rooth1993] Hindle, D. and Rooth, M. 1993. *Structural ambiguity and lexical relations*. Computational Linguistics, 19(1): 103-120.
- [Jensen and Binot1987] Jensen, K. and Binot, J. L. 1987. *Disambiguating Prepositional Phrase Attachments by Using On-Line Dictionary Definitions*. Computational Linguistics, 13(3-4): 251-260.
- [Salton and MacGill 1983] Salton G., MacGill M.J., 1983. *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.
- [Stetina and Nagao1994] Stetina, J. and Nagao, M. 1997. *Corpus-based PP Attachment Resolution with a Semantic Dictionary*. In Proceedings of the 5th Workshop on Very Large Corpora, VLC-97, eds. J. Zhou and K. Church, pp. 66-80, Beijing and Hongkong.