PP Attachment Ambiguity Resolution with Corpus-Based Pattern Distributions and Lexical Signatures

Nuria Gala¹ and Mathieu Lafourcade², Non-members

ABSTRACT

We propose a method mixing unsupervised learning of lexical pattern frequencies with semantic information which aims to 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 \( V \text{ NP PP} \) (or \( V \text{ N P N} \)) “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 chunks 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:)

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 prepositional attachment, the relations with the three ele-
PP Attachment Ambiguity Resolution with Corpus-Based Pattern Distributions and Lexical Signatures

ments \((X, P, N)\) can be calculated through \textit{VMOD} or \textit{NMOD} and \textit{PREPOBJ} dependencies. Thus \textit{VMOD(acheter, enfant)} and \textit{PREPOBJ(enfant, pour)} give \((acheter, pour, enfant)\).

The parser is deterministic for calculating all the dependencies (one solution is proposed among the eventual possibilities). Prepositional 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:

\[(acheter, pour, enfant)\]
\[(vtement, pour, enfant)\]

\textbf{Fig.2: Prepositional phrase attachments. buy, for, child - cloth, for, child}

\subsection*{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 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 obtained by harvesting the Web.

This measure, that we call \textit{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 \((X, P)\):

\[
SCS(X, P, N) = \frac{\#(X, P, N)}{\#(X, P)}
\]

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 \textit{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 “\textit{Le résultat courant exprime la rentabilité de la société en intégrant les excédents de gains par l’exploitation (…).}” (Eng. The current result shows the profitability of the society by including the surplus obtained by exploiting (…)) where \textit{par l’exploitation} although attached to \textit{dgags} would be found in the corpus with very few occurrences and the pattern \textit{excédents par l’exploitation} would not be found at all.

\[
\text{SCS(dgags par l’exploitation)} = \frac{240}{102.000} = 0.0023
\]
\[
\text{SCS(excédents par l’exploitation)} = \frac{0}{257} = 0
\]
\[
\text{SCS(intégrant par l’exploitation)} = \frac{0}{632} = 0
\]

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

\subsection*{3. LEXICAL SIGNATURES}

A lexical signature of a term \(t\) is a set of weighted terms that allows to characterize \textit{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 \textit{enfant} (Eng. child): \textit{enfant}: ("femme" 2.37) ("personne" 1.62) ("parent" 1.12) ("deux opéra" 1) ("Milhaud" 1.0) ("bataille" 1) ("pare de partie carboniser" 1) ("aucun guide touristique" 1) ("me adresser" 1) ("ex-Yougo"
thematic proximity

Intuitively, this function constitutes an evaluation of distance \( D \). Then, we define an angular as the scalar product of their vector divided by the in information retrieval. We can express this function measures between two signatures \( A \) et \( B \), often used

distance \( D \) product of their norm. Then, we define an angular as the scalar product of their vector divided by the in information retrieval. We can express this function measures between two signatures \( A \) et \( B \), often used

distance \( D \) product of their norm. Then, we define an angular as the scalar product of their vector divided by the in information retrieval. We can express this function measures between two signatures \( A \) et \( B \), often used

For the term vtement (Eng. clothes): vtement: ("un marque italien" 1) ("fabricant choletain" 1) ("le sous-vtement" 0.67) ("enfant" 0.41) ("le prt-porter" 0.38) ("coutette" 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 rpartir" 0.18) ("le chaus-sure" 0.16) ("blondinet" 0.16) ("ce entre" 0.16) ("appt" 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) ("dtenir" 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:

\[
D_A(A, B) = \arccos(Sim(A, B))
\]

with

\[
Sim(A, B) = \cos(\overline{A} \cdot \overline{B}) = \frac{A \cdot B}{\|A\| \times \|B\|}
\]

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 \( D_A(A, B) \leq \frac{\pi}{2} \), (i.e. less than 45 degrees) \( A \) and \( B \) are thematically close and share many terms. For \( D_A(A, B) \geq \frac{\pi}{2} \), 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 dictionaries. 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:

\[
D_A(\text{child}, \text{child}) = 0 \quad D_A(\text{clothes}, \text{child}) = 70
\]

\[
D_A(\text{to buy, child}) = 85 \quad D_A(\text{clothes, to buy}) = 76
\]

The first value as a straightforward interpretation due to the reflexivity of the distance. There are more mutual information between \text{clothes} and \text{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:

\[
MIS(X,P,N) = M_S(X,P,N) \quad \text{if } X.P \in C
\]

\[
MIS(X,N) = M_S(X,N) \quad \text{otherwise}
\]

For the sentence "Elle achte des vtements pour ses enfants", we have the following attachments: "acheter pour ses enfants" or "vtements pour ses en-fants". The MIS are respectively:

\[
\text{MIS (buy, child)} = 0.05
\]

\[
\text{MIS (clothes, child)} = 0.22
\]

3.2 Building signatures

For a given word \( w \), we build its signature over the corpus \( C \) the following way. We consider a window of \( \delta \) terms before and \( \delta \) terms after the target word, at the paragraph level, which have been processed beforehand through a chunk analyzer. In our experiments, we empirically set \( \delta \) to 10. The terms before \( w \) are noted \( t_1, \ldots, t_{-\delta} \), those after \( t_\delta \) are noted \( t_{1-\delta}, \ldots, t_{10} \). 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 before the previous NP chunk. For example:

\[
NP(\text{missile}) \ AP(\text{amricain})
\]

adds \( NP(\text{missile amricain}) \)

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" + "tre tellement difficile" 

"jeune" + "tre parfois trs jeune" "saccager" + "avoir saccage"
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 $\mathcal{C}$:

$$V(w) = \sum_{t \in \mathcal{T}} \frac{1}{d} \times \frac{1}{1 + \log(\# t)} \times V_0(t) \quad (4)$$

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

$${\mathcal{C}} = \{ t_1, \ldots, t_i, \ldots, t_n \}$$

$$V_0(t_i) = < 0_1, \ldots, 1_i, \ldots, 0_n > \quad (5)$$

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 that 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 $t$.

$$V_n(t) = < w_1, \ldots, w_i, \ldots, w_n >$$

$$V_{n+1}(t) = \sum_{k \in [1,n]} w_k \times V_n(t_k) \quad (6)$$

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 $\text{score}(a_1)/\text{score}(a_2)$. A value below 0 implies that the second attachment if found as more likely than the first. In the following example, both scorings agree on the second attachment.

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

MIS (acheter, enfant) / MIS(vtement, 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:

$$\text{Conf(score)} = \begin{cases} 1 & \text{if score} < 0 \\ \text{score} & \text{otherwise} \end{cases} \quad (7)$$

For example:

$$\text{SCS}(X_1 P_1, N_1)/\text{SCS}(X_2 P_2, N_2) = 0.3$$

$$\text{Conf}(0.3) = 0.333$$

$$\text{MIS}(X_1, N_1)/\text{MIS}(X_2, N_2) = 2.8$$

$$\text{Conf}(2.8) = 2.8$$

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 ($T_e$) 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 ($L_e$) 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
5. CONCLUSION AND FURTHER WORK

This paper addresses the issue of combining two kinds of information, statistical and lexical, for improving PP attachment disambiguation. We presented a ratio method that allows us to overcome the issue of 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. The increase in the training corpus size, would by itself improve performance but would eventually reach its own limits.

References