Lexicalising a robust parser grammar using the WWW

Núria Gala Pavia
XRCE and LIMSI-CNRS
Bt. 508 Université de Paris-Sud
91403 Orsay Cedex
gala@limsi.fr

Salah Aït-Mokhtar
XRCE
6 chemin de Maupertuis
38240 Meylan
ait-mokhtar@xrce.xerox.com

Abstract
This paper describes an unsupervised method to lexicalise a robust parser grammar for French in order to improve prepositional phrase (PP) attachment. The ambiguous attachments produced by the parser after a first analysis of an input text are transformed and used in queries to find and download documents from the Web, where the involved words occur. The collected corpus is parsed and, from the parsing results, we acquire statistical information on PP-attachment configurations, hence building a weighted subcategorisation lexicon. This automatically acquired subcategorisation information is used in a second analysis of the input text in order to improve the disambiguation of multiple PP-attachment.

1. Introduction

Many robust parsing systems that can process real-world texts were developed in the nineties (Abney 1996, Grefenstette 1996, Collins 1996, Aït-Mokhtar and Chanod 1997, Charniak 2000). These parsers assign linguistic structures to text sentences and are used with some success in various applications, such as information extraction, question answering systems, word sense disambiguation, etc. Most of them are reported to have good accuracy, sometimes above 90% precision in the recognition of some syntactic structure elements (noun and prepositional phrases, subject-verb and object-verb relations, etc.).

However, the correct handling of attachment ambiguities, especially PP-attachment, is still an issue. The use of traditional subcategorisation lexicons is not of much help. Large-scale lexical subcategorisation information for verbs, nouns and adjectives is available for few languages. When available, it is very often not informative enough to resolve such ambiguities. In particular, in ambiguous situations where the involved headwords share the same subcategorisation properties, it is not possible to induce attachment preferences from traditional subcategorisation lexicons.

In this context, probability-based techniques, be they supervised or unsupervised, have been designed to capture and measure subcategorisation preferences, and have been applied to PP-attachment.

Unsupervised methods exploit regularities in raw or automatically annotated corpora and calculate frequencies of the presence of a word next to another. One of the classic approaches is that of lexical association scores (Hindle and Rooth 1993). This method, which obtains about 80% precision, estimates the probability of association of a preposition with a noun or a verb from their distributions in corpora (previously annotated by a parser). Other unsupervised approaches, such as that of Volk (Volk 2000, Volk 2001), obtain 74% accuracy for all decidable test cases for German using triples (`X,P,N2`).

Supervised methods such as transformation-based learning (Brill and Resnik 1994), memory-based learning (Zavrel et al. 1997), etc. use manually annotated training corpora. The average accuracy results on PP-attachment ambiguities, restricted to the [`Verb NP PP`] configuration, are reported to be around 80-84% for English. The use of semantic classes in the supervised method of (Stetina and Nagao 1997) allows them to reach a precision of 88% for the same configuration.

As for more ambiguous attachments, with multiple alternatives, (Merlo et al. 1997) propose a generalized back-off method for English, dealing with structures such as [`Verb NP PP1 PP2`]. The reported precision rate is 69,6% for `PP2` and 43,6% for `PP3`. According to the authors, the increasing number of configurations to take into account, and a problem of sparse data may justify these figures.

For PP-attachment disambiguation in French, (Gaussier and Cancedda 2001) propose a statistical model that integrates different resources (subcategorisation frames, semantic information). The results reported are around 85% for `[Verb NP PP]` but only 74% for other structures such as `[NP PP PP]`.

To summarize this brief overview on existing methods to resolve PP-attachment, (a) most of the proposed methods are supervised, because it is more efficient to compare the results with corpora created or verified by humans and (b) precision rates are around 84% for the `[Verb NP PP]` configuration but significantly less for more ambiguous configurations.

2. Overview of our approach

Our method for lexicalising the core dependency grammar for PP-attachment resolution combines the results of a first analysis of the parser with statistical information extracted from the Web. We applied an unsupervised method to improve PP-attachment in an existing rule-based dependency system for French based on the XIP parser (Aït-Mokhtar et. al. 2002).

2. 1 The parser

XIP is an incremental parsing framework that allows the extraction of deep dependency relations at sentence and inter-sentential level. A dependency is, in this approach, a syntactic relation holding between the headwords of two chunks, e.g. a noun and a verb, or in the case of PP-attachment, between two headwords linked by a preposition. We represent the PP-attachment dependency with `A(X,Prep,N)` where `X` is a noun, a verb or an adjective, `Prep` a preposition and `N` a noun, for instance `A(impact,sur,comptes)` *(impact on accounts).*

Several grammars have been designed and implemented for chunking and dependency extraction in a two-tier approach (Gala 2001). Figure 1 gives an idea of the overall architecture of the system.

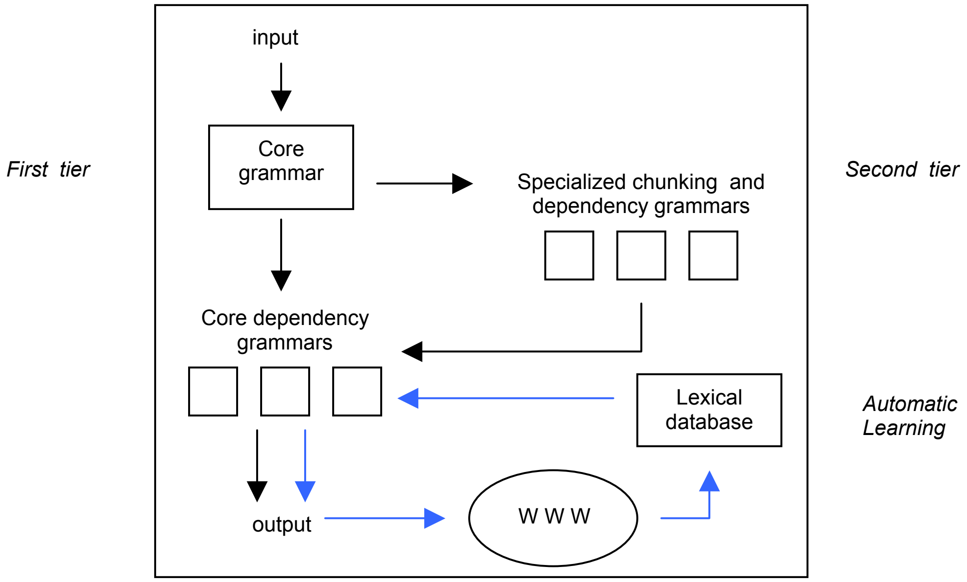

Figure 1: Architecture of our parsing approach.

The grammars (a core grammar and some specialized modules) are applied depending on the linguistic and structural characteristics of each input sentence (Gala 2003). For chunking, the core grammar identifies main chunks (NPs, PPs, etc.) whereas specialized chunking modules mark particular constructions such as lists, enumerations, etc. For dependency extraction, specialized modules calculate dependencies in those constructions. Core dependency grammars extract main syntactic functions (subject, object, etc.) as well as PP-attachment.

The output of this model is a chunked sentence with a list of dependency relations. Given the French sentence "La méthode d'amortissement choisie aura un impact significatif sur les comptes." *("The* *selected depreciation method will have a significant impact on the accounts.”*), the parser gives the following output[1]:

```
SUBJ(aura,méthode)
OBJ(aura,impact)
NADJ(impact,significatif)
NADJ(méthode,choisie)
A(méthode,de,amortisement)
A(a,sur,comptes)
A(impact,sur,comptes)
A(significatif,sur,comptes)

0>MAX{NP{La méthode} PP{d' NP{amortissement}} AP{choisie} FV{aura}
 NP{un impact} AP{significatif} PP{sur NP{les comptes}} .}
```

2. 2 Automatic learning of subcategorisation

To improve PP-attachment some of the dependencies extracted after a first analysis are transformed into subcategorisation frames and a measure of frequency for each frame (a weight) is calculated using a big collection of documents. These weights are integrated into the system and later used to disambiguate controversial attachments in a second analysis.

Unlike Volk’s approach (Volk 2001), the subcategorisation weights here are *syntactic co-occurrence* probabilities, as opposed to *textual* co-occurrence probabilities, based on the frequencies of words occurring in the same text within a maximal linear distance, i.e. with some maximal number of words in between. Instead, in our approach syntactic co-occurrence probabilities are measured from the frequencies of words co-occurring in the same syntactic dependency relations (i.e. attachments already yielded by the parser in a first analysis). The approach is unsupervised since the dependency relations are automatically produced by the existing parser itself, which may suggest for each PP one or several possible attachments.

Our hypothesis is that PP-attachments produced by the parser are more reliable than rough textual co-occurrences, because the parser disregards all the cases where the words co-occurring in text clearly cannot be syntactically linked. Such situations include words occurring in distinct sentences, in distinct finite clauses, etc.

3. Initial and baseline grammars of the parser

The grammar that produces the dependencies for PP-attachment (the baseline grammar that we call G2) is made of two initial grammars. The first one (G0) gives priority to recall (the rules present few constraints in order to extract as many dependencies as possible) while the other grammar (G1) produces fewer dependencies but with a high degree of precision.

The corpus used for the experiments is a heterogeneous corpus from different domains: scientific (46,8% of the corpus), economic (41%), and legal (12,2%). It contains 2.493 words (89 sentences) with an average of 28 words per sentence. The evaluation of PP-attachment using G0 and G1 grammars gives the following results:

Grammar	Number of rules	Precision	Recall	F	Number of dependencies
G0	16	41,88 %	**94,28 %**	58,00 %	2.438
G1	11	**87,49 %**	77,04 %	**81,93 %**	964

Table 1: Results of the initial grammars.

[1] SUBJ is a subject relation, OBJ a direct object, NADJ an adjective modifying a noun, A a PP attachment. MAX is the maximal group, NP a noun phrase, PP a prepositional phrase, AP an adjective phrase, FV a finite verb phrase.

The baseline G2 grammar is an optimisation of the two initial grammars. It is made of "reliable" G1 rules (those rules that, individually, have more than 93% precision[2]) and the remaining G0 rules, only applied after verifying that a PP has not been attached by a reliable G1 rule. The overall results of G2 are the following:

Grammar	Number of rules	Precision	Recall	F	Number of dependencies
G2	21	**71,34 %**	**92,10 %**	**80,40 %**	1.413

Table 2: Results of the baseline grammar.

While the precision rate decreases compared to G1, the recall rate increases significantly. The F-score (average measure) remains a little bit higher than 80%.

The output provided by G2 gives information on the reliability of a PP-attachment (we use the features MF1 and MF2, MF stands for *mesure de fiabilité* –reliability measure). For the previous example[3]:

```
A_MF1(méthode,de,amortissement)
A_MF2(avoir,sur,compte)
A_MF2(impact,sur,compte)
A_MF2(significatif,sur,compte)
```

The dependency relations with a "less reliable" mark (MF2) are used to build a database of subcategorisation patterns enriched with statistical measures obtained from the Web.

4. Building a subcategorisation database with the WWW

The use of the World Wide Web as a huge database of examples for different tasks related to NLP is quite a recent idea: for translating compound nouns (Grefenstette 1999), for acquiring named entities (Jacquemin and Bush 2000), for PP-attachment resolution (Volk 2001). In our approach, less reliable PP-attachments produced by the baseline grammar on a given document are used to query the Web in order to obtain a huge collection of documents. Our hypothesis is that occurrences of the attachments to validate should be present in this corpus (or not at all if they are wrong).

4.1 The queries

After a first analysis of the initial corpus with G2, the 869 dependencies having the feature MF2 have been transformed into queries for the Web. Each query contains the three words involved in the PP-attachment dependency `A(X,Prep,N)`, where `X` is a verb, a noun or an adjective. The two first tokens are what we call the subcategorisation frame; the third one is the word to attach.

We have chosen `Altavista` (www.Altavista.com) with the advanced search option, which permits the use of Boolean operators such as `NEAR` (allowing a distance of at most ten words between two given words). The queries are of the form `(X Prep) NEAR N`. For the dependency `A_MF2(impact,sur,comptes)` the query is the following one.

```
(impact sur) NEAR comptes
```

This query may select documents containing the following configurations:

Document 1	… **impact sur** les **comptes** …
Document 2	… **Impact sur** l'analyse des **comptes** …
Document 3	… **impact sur** la fermeture de vos **comptes** …

[2] An evaluation of the 11 rules of G1 has shown that 5 of them have a precision rate going from 100% to 93,72% (while the others from 89,79% to 41,67%). These 5 rules where P > 93% are called "reliable".

[3] The dependencies can also be displayed with the tokens and their word number: `A_MF2(<a^avoir:5>,<sur^sur:9>,<comptes^compte:11>)`.

Some of the selected documents may be inappropriate, that is, the occurrences present in the document are not syntactic co-occurrences. In the previous example, the occurrence in document 3 is not appropriate because the right attachment is "*fermeture de comptes*" and not "*impact sur comptes*". Before computing the frequencies of subcategorisation patterns, we parse the documents and only those occurrences that yield a PP-attachment candidate are considered. We think that the huge size of the WWW should prevent any sparse data problem that such restrictions might otherwise yield.

4.2 The corpus for the database

Each of the 869 queries (created from the 869 MF2 dependencies) retrieves 20 URLs. To perform this task we have used a set of `perl` scripts combined with the Unix command `wget`. The result of this process is a collection of about 17.000 documents (the sizes of the corpora obtained for one single query can be very different; it is also possible to harvest fewer than 20 URLS for a given query).

After removing all the HTML tags from the corpora, we have a new corpus (we call it the Web corpus) of 38.242.073 words and 1.368.903 sentences. To create the database, we have parsed this corpus and extracted the overall dependencies for PP-attachment. Here, we were interested in the *quantity* of the dependencies (about 4 million) and not in the *quality* (produced by a reliable or less reliable rule). Therefore no difference has been made between MF1 and MF2 dependencies.

4.3 The database

The dependencies `A(X,Prep,N)` from the above mentioned corpus have been automatically transformed into subcategorisation frames `(X Prep)` using the lemmas. As in (Volk 2001) a measure of co-occurrence (`MC`) has been automatically calculated. This measure is determined by the frequency of a word (`X`) and the frequency of the co-occurrence `(X Prep)`:

```
MC(X,Prep) = freq(X Prep) / freq(X)
```

As a result, we obtain a database that has entries of the following form:

```
0.0008 avoir | sans
0.0005 avoir | sous
0.0118 avoir | sur
…
0.0005 impact | selon
0.0626 impact | sur
…
0.0323 significatif | à
```

This weighted subcategorisation lexicon contains about 100.000 subcategorisation frames obtained from the Web corpus, each one having a co-occurrence value. The database is incorporated as a resource in the parser and used to resolve PP-attachment ambiguities.

5. Disambiguation of PP-attachment dependency relations

For PP-attachment resolution, we have applied a disambiguation algorithm to all "less reliable" (MF2) dependencies from the initial corpus. This algorithm aims at finding the correct attachment for a given headword when the parser has yielded multiple attachments after the first analysis step.

Each PP-attachment candidate is compared to the patterns in the weighted subcategorisation lexicon. Among the set of attachment candidates for a given PP (Prep N), only the one that corresponds to the best-weighted subcategorisation pattern is selected. For the previous example, the dependency `A(impact,sur,compte)` is kept while `A(avoir,sur,compte)` is wiped away (see the scores above). Whenever the scores are equal, as the algorithm does not have any other information to resolve the ambiguity, both dependencies are kept.

Otherwise, if a frame is not in the database, we suppose its score to be null. In this case, the comparison to an existing pattern does not take place and the existing subcategorisation frame is kept. In the example, this means eliminating `A(significatif,sur,compte)`. When none of the patterns exists in the database, both dependencies are kept.

The final result for the example above is:

```
A(méthode,de,amortissement)
A(impact,sur,compte)

0>MAX{NP{La méthode} PP{d' NP{amortissement}} AP{choisie} FV{aura}
 NP{un impact} AP{significatif} PP{sur NP{les comptes}} .}
```

Note that the first PP-attachment is yielded after a first analysis by a reliable rule while the second one is produced in a second round, after applying the disambiguation algorithm using the information on the lexical database.

The overall results on the initial corpus are the following:

Grammar	Precision	Recall	F	Number of dependencies
G2	71,37 %	92,10 %	80,40 %	1.413
Lexicalised G2	**83,21 %**	85,12 %	**84,16 %**	1.120

Table 3: Results of the baseline grammar before and after lexicalisation.

As shown in the table, we get an attachment precision of 83,21%, which means an 11,84% increase compared to the results obtained only with the rule-based baseline grammar. The F-score also increases (+3,76%). Unlike most of the evaluations reported on section 1, these figures concern all kinds of configurations for PP-attachment in French `[V NP+ AP+ PP1 PP2 … PPn]`.

6. Conclusions

This paper presents an unsupervised method to improve PP-attachment resolution for French in an existing rule-based dependency parser. The initial system, made of different grammars, is enriched with a weighted lexical database containing subcategorisation frames and weights (syntactic co-occurrence probabilities) automatically acquired from the Web.

To construct the lexical database, a measure of reliability yielded by the parser after a first analysis permits the identification of some of the syntactic dependencies that will be used to query the Web.

Compared to the initial rule-based grammar, the results obtained prove that the use of both lexical and statistical information significantly increases the precision rate of the PP-attachment dependencies.

Further refinements can be foreseen to improve the overall results. One suggestion is the use of the three elements of a dependency (`X Prep N`) to calculate the co-occurrence weight. This solution would allow more precise configurations to be used during the disambiguation process. Another point would be the generalisation of the method using semantic classes encoded as features: the concordance of two classes would be given priority during the PP-attachment resolution.

Finally, this approach could be applied to other problems of structural ambiguity such as coordination or complex named entity recognition.

References

Abney, S. P. 1996 Partial Parsing Via Finite-State Cascades. In *ESSLLI'96 Workshop on Robust Parsing*, Prague, Czech Republic.

Aït-Mokhtar, S. and Chanod, J. P. and Roux, C. 2002 Robustness beyond shallowness: incremental deep parsing. In *Natural Language Engineering*, 8(2/3), pp. 121-144, Cambridge University Press, UK.

Aït-Mokhtar, S. and Chanod, J.-P. 1997 Incremental Finite-State Parsing. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, pages 72-79, Washington, DC, USA.

Bourigault, D. and Fabre, C. 2000 Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire* (25): 139-151.

Brill, E. and Resnik, P. 1994 A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, Kyoto, Japan.

Charniak, E. 2000 A maximum entropy inspired parser. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 132-139, Seattle, Washington.

Collins, Michael J. 1996 A new statistical parser based on bi-gram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 184-191, Santa Cruz, USA.

Gala Pavia, N. 2003 *Un modèle d'analyseur syntaxique robuste basé sur la modularité et la lexicalisation de ses grammaires.* PhD thesis. Université de Paris-Sud (forthcoming).

Gala Pavia, N. 2001 A two-tier corpus-based approach to robust syntactic annotation of unrestricted corpora, *TAL* 42 (2): 381-411.

Gaussier E. and Cancedda, N. 2001 Probabilistic models for PP-attachment resolution and NP-analysis. In *Proceedings of ACL-2001, Computational Natural Language Learning Workshop, CoNLL-2001*, pp. 45-52, Toulouse.

Grefenstette, G. 1999 The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of Aslib Conference on Translation and the Computer 21*, London.

Grefenstette, G. 1996 Light Parsing as Finite-State Filtering. In *Proceedings of the ECAI'96 workshop on extended finite state models of language*, Budapest.

Hindle, D. and Rooth, M. 1993. Structural ambiguity and lexical relations. In *Computational Linguistics*, 19(1), pp.103-120, the MIT Press.

Jacquemin, C. and Bush, C. 2000 Combining lexical and formatting cues for named entity acquisition from the Web. In Schutze H. editor, *Proceedings of joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP-VLC-2000*, Hong Kong.

Merlo P., Crocker M. W. and Berthouzoz C. 1997 Attaching multiple prepositional phrases: Generalized Backed-off Estimation. In Cardie C. and Weischedel, R. editors, *Proceedings of the second conference on Empirical Methods in Natural Language Processing, EMNLP-97*, pp. 149-155, Providence, R.I.

Stetina, J. and Nagao, M. 1997 Corpus Based PP-Attachment Ambiguity Resolution with a Semantic Dictionary. In *Proceedings of the Fifth Workshop on Very Large Corpora*, pp. 66-80, Beijing and Hong Kong.

Volk, M. 2000 Scaling up. Using the WWW to resolve PP attachment ambiguities. In *Proceedings of Konvens-2000*, pp. 151-155, Ilmenau, Germany.

Volk, M. 2001 Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of Corpus Linguistics 2001*, pp. 601-606, Lancaster, UK.

Zavrel, J., Daelemans, W. and Veenstra, J. 1997 Resolving PP Attachment Ambiguities with Memory-Based Learning. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-97)*, Madrid.