Probabilistic Models for PP-attachment Resolution and NP Analysis

Eric Gaussier
XRCE
6, Chemin de Maupertuis
38240 Meylan, France
fname.lname@xrce.xerox.com

Nicola Cancedda
XRCE
6, Chemin de Maupertuis
38240 Meylan, France
fname.lname@xrce.xerox.com

Abstract

We present a general model for PP attachment resolution and NP analysis in French. We make explicit the different assumptions our model relies on, and show how it generalizes previously proposed models. We then present a series of experiments conducted on a corpus of newspaper articles, and assess the various components of the model, as well as the different information sources used.

1 Introduction

Prepositional phrase attachment resolution and noun phrase analysis are known to be two difficult tasks. Traditional context-free rules for example do not help at all in selecting the good parse for a noun phrase, since all valid parses are *a priori* correct. Subcategorization information can help solve the problem, but the amount of information necessary to be encoded in such lexicons is huge, since in addition to subcategorization frames, one should encode the different senses of words and rules on how these senses can be combined, as well as the different units (single or multi word expressions) a language makes use of. It has been recognized in different works that part of this knowledge can be (semi-)automatically acquired from corpora and used in the decision process. Several models have been proposed for PP-attachment resolution and NP analysis, built on various building blocks and making use of diverse information. Most of these models fit within a probabilistic framework. Such a framework allows one to estimate various quantities and to perform inference with incomplete information, two necessary steps for the problem at hand. We present in this paper a model generalizing several models already proposed and integrating several information sources.

We focus here on analysing strings corresponding to a combination of traditional VERB NOUN PREP sequences and noun phrases. More precisely, the NPs we consider are arbitrarily complex noun phrases with no embedded subordinate clause. The PP attachment problems we tackle are those involved in analysing such NPs for languages relying on composition of Romance type (as French or Italian), as well as those present in verbal configurations for languages relying on composition of Germanic type (as German or English). However, our model can easily be extended to deal with other cases as well. The problem raised in analysing such sequences is of the utmost relevance for at least two reasons: 1) NPs constitute the vast majority of terms. Correctly identifying the boundaries and the internal structure of noun phrases is crucial to the automatic discovery of domain-specific terminological databases; 2) PP-attachment ambiguity is one of the main sources of syntactic ambiguity in general.

In the present work, we focus on the French language since we have various lexical information at our disposal for this language that we would like to assess in the given context. Furthermore, French displays interesting propoerties (like gender and number agreement for adjectives) which makes it an interesting language to test on. The remainder of the paper is organized as follows: we first describe how we preprocessed our corpus, and which part we retained for probability estimation. We then present the general model we designed, and show how it compares with previous ones. We then describe the experiments we performed and discuss the results obtained. We also describe, in the experiments section, how we integrated different types of information sources (e.g. prior knowledge encoded as subcategorization frames), and how we weighted multiple sources of evidence according to their reliability.

2 Nuclei and sequences of nuclei

We first take the general view that our problem can be formulated as one of finding dependency relations between nuclei. Without loss of generality, we define a nucleus to be a unit that contains both the syntactic and semantic head and that exhibits only unambiguous internal syntactic structure. For example, the base NP "the white horse" is a nucleus, since the attachments of both the determiner and the adjective to the noun are straightforward. The segmentation into nuclei relies on a manually built chunker, similar to the one described in (Ait-Mokhtar and Chanod, 1997), and resembles the one proposed in (Samuelsson, 2000). The motivation for this assumption is twofold. First, the amount of grammatical information carried by individual words varies greatly across language families. Grammatical information carried by function words in non-agglutinative languages, for instance, is realized morphologically in agglutinative languages. A model manipulating dependencies at the word level only would be constrained to the specific amount of grammatical and lexical information associated with words in a given language. Nuclei, on the other hand, tend to correspond to phrases of the same type across languages, so that relying on the notion of nucleus makes the approach more portable. A second motivation for considering nuclei as elementary unit is that their internal structure is by definition unambiguous, so that there is no point in applying any algorithm whatsoever to disambiguate them.

We view each nucleus as being composed of several linguistic layers with different information, namely a semantic layer comprising the possible semantic classes for the word under consideration, a syntactic layer made of the POS category of the word and its gender and number information, and a lexical layer consisting of the word itself (referred to as the lexeme in the following), and the preposition, for prepositional phrases. For nuclei comprising more than two non-empty words (as "the white horse"), we retain only one lexeme, the one associated with the last word which is considered to be the head word in the sequence. Except for the semantic information, all the necessary information is present in the output of the chunker. The semantic lexicon we used was encoded as a finite-state transducer, which was looked up for injecting semantic classes in each nucleus. When no semantic information is available for a given word, we use its part-of-speech category as its semantic class[1]. For example, starting with the sentence:

[1]The semantic resource we used can be purchased from www.lexiquest.com. This resource contains approximately 90 different sementic classes organized into a hierarchy. We have not made use of this hierarchy in our experiments.

Il doit rencontrer le président de la fédération française.
(He has to meet the president of the French federation.)

we obtain the following sequence of nuclei:

il
CAT="PRON", GN="Masc-Sg", PREP="", SEM="PRON"
rencontrer
CAT="VERB", GN="", PREP="", SEM="VERB"
président
CAT="NOUN", GN="Masc-Sg", PREP="", SEM="FONCTION"
fédération
CAT="NOUN", GN="Fem-Sg", PREP="de", SEM="HUMAIN"
française
CAT="ADJ", GN="Fem-Sg", PREP="", SEM="GEO"

As we see in this example, the semantic resource we use is incomplete and partly questionable. The attribute HUMAN for *federation* can be understood if one views a federation as a collection of human beings, which we believe is the rationale behind this annotation. However, a federation also is an institution, a sense which is missing in the resource we use.

In the preceding example, the preposition *de* can be attached to the verb *rencontrer* or to the noun *président*. It cannot be attached to the pronoun *il*. As far as terminology extraction is our final objective, *président de la fédération française* can be deemed a good candidate term. However, in order to accurately identify this unit, a high confidence in the fact that the preposition *de* attaches to the noun *président* must be achieved. Sentences can be conveniently segmented into smaller self-contained units according to some heuristics to reduce the combinatorics of attachments ambiguities. We define **safe chains** as being sequences of nuclei in which all the items but the first are attached to other nuclei within the chain itself. In the preceding example, for instance, only the nucleus associated with *rencontrer* is not attached to a nucleus within the chain *rencontrer ... française*. This chain is thus a safe chain. To keep the number of alternative (combinations of) attachments as low as possible, we are interested in isolating as short safe chains as possible given the information available at this point, i.e. words and their parts-of-speech (the knowledge of semantic classes is of little help in this task).

In French, and except for few cases involving embedded clauses and coordination, the following heuristics can be used to identify "minimal" safe chains: extract the longest sequences beginning with a nominal, verbal, prepositional or adjectival nucleus, containing only nominal, prepositional, adjectival, adverbial or verbal nuclei in indefinite moods.

There is a tension in parameter estimation of probabilistic models between relying on accurate information and relying on enough data. In an unsupervised approach to PP-attachment resolution and NP analysis, accurate information in the form of dependency relations between words is not directly accessible. However, specific configurations can be identified from which accurate information can be extracted. Safe chains provide such configurations. Indeed if there is only one possible attachment site to the left of a nucleus, then its attachment is unambiguous. Due to the possible ambiguities the French language displays (e.g. a preposition can be attached to a noun, a verb or an adjective), only the first two nuclei of a safe chain provide reliable information (we skip adverbs, the attachment of which obeys specific and simple rules). From the preceding example, for instance, we can infer a direct relation between *rencontrer* and *président*, but this is the only attachment we can be sure of. The use of less reliable information sources for model parameters whose estimation would otherwise require manual supervision is the object of an experiment described in Section 6.

3 Attachment Model

Let us denote the i_{th} nucleus in a chain by (i), and the the nucleus to which it is attached by $a(i)$ (for each chain, we introduce an additional empty nucleus to which the head of the chain is attached). Given a chain of nuclei C, we denote by $\mathcal{A}_i$ the set of dependency relations covering the chain of nuclei $C_k, 1 \leq k \leq i$. We are interested in the set $\mathcal{A}_n$ such that $p(\mathcal{A}_n)$ is maximal. Assuming that the dependencies are built by processing the chain in linear order, We have:

$$p(\mathcal{A}_n|C) = \sum_{n-1} p(\mathcal{A}_{n-1}|C)p(\mathcal{A}_n|\mathcal{A}_{n-1}, C)$$
$$= \sum_{1} \dots \sum_{n-1} \prod_{j=1}^{n} p(\mathcal{A}_j|\mathcal{A}_{j-1}, C) \quad (1)$$

$\mathcal{A}_j$ differs from $\mathcal{A}_{j-1}$ only in that it additionally specifies a particular attachment site (i) for (j) such that no cycle nor crossing dependencies are produced. In order to avoid sparse data problems, we make the simplifying assumption (similar to the one presented in (Eisner, 1996)) that the attachment of nucleus (j) to nucleus (i) depends only on the set of indices of the preceding dependency relations (in order to avoid cycles and crossing dependencies) and on the three nuclei (j), (i) and (k_j), where (k_j) denotes the last nucleus being attached to (i). (k_j) is thus the closest sibling of (j). Conditioning attachment on it the attachment of (j) allows capturing the fact that the object of a verb may depend on its subject, that the indirect object may depend on the direct object, and other similar indirect dependencies. In order to focus on the probabilities of interest, we use the following simplified notation:

$$p(\mathcal{A}_j|\mathcal{A}_{j-1}, C) \approx$$
$$p(\mathcal{M}(\mathcal{A}_j)) \times p((j)|(i), (k_j), \mathcal{M}(\mathcal{A}_j)) \quad (2)$$

where $\mathcal{M}(\mathcal{A}_j)$ represents the graph produced by the dependencies generated so far. If this graph contains cycles or crossing links, the associated probability is 0. Making explicit the different elements of a nucleus, we obtain:

$$p((j)|(i), (k_j), \mathcal{M}(\mathcal{A}_j)) =$$
$$\sum_{S \in S(j)} p(S|(i), (k_j)) \quad (3)$$
$$\times p(pr_j|S, (i), (k_j)) \quad (4)$$
$$\times p(cat_j|pr_j, S, (i), (k_j)) \quad (5)$$
$$\times p(gn_j|cat_j, pr_j, S, (i), (k_j)) \quad (6)$$
$$\times p(lex_j|gn_j, cat_j, pr_j, S, (i), (k_j)) \quad (7)$$

since the graph $\mathcal{M}(\mathcal{A}_j)$ provides the index of the nucleus (i) to which (j) is attached to. Obviously, most of the above probabilities cannot be directly estimated. A number of simplifying assumptions preserving significant conditional dependencies were adopted.

Assumption 1: except for graphs with cycles and crossing links, for which the associated probability is 0, we assume a uniform distribution on the set of possible graphs.

A prior probability $p(\mathcal{M}(\mathcal{A}_j))$ could be used to model certain corpus-specific preferences such as privileging attachments to the immediatly preceding nucleus (in French or English for example). However, we decided not to make use of this possibility for the moment.

Assumption 2: the semantic class of a nucleus depends only on the semantic class of its regent.

This assumption, also used in (Lauer and Dras, 1994), amounts to considering a 1st-order Markov chain on the semantic classes of nuclei, and represents a good trade-off between model accuracy and practical estimation of the probabilities in (3). It leads to:

$$p(S|(i), (k_j)) = p(S|S(i)) \quad (8)$$

Assumption 3: the preposition of a nucleus depends only on its semantic class and on the lexeme and POS category of its regent, thus leading to:

$$p(pr_j|(i), (k_j)) = p(pr_j|S, cat_i, lex_i) \quad (9)$$

The nucleus (k_j) does not provide any information on the generation of the preposition, and is thus not retained. As far as the regent nucleus (i) is concerned, the dependence on the POS category controls the fact that adjectives are less likely to subcategorize prepositions than verbs. For arguments, the preposition is controlled by subcategorization frames, which directly depend on the lexeme under consideration, and to a less extent to its semantic class (even though this dependence does exist, as for movement verbs which tend to subcategorize prepositions associated with location and motion). In the absence of subcategorization frame information, the conditioning is placed on the lexeme, which also controls prepositional phrases corresponding to adjuncts. Lastly, the semantic class of the nucleus under consideration may also play a role in the selection of the preposition, and is thus retained in our model.

Assumption 4: the POS category of a nucleus depends only on its semantic class.

This assumption reflects the fact that our lexical resources assign semantic classes from disjoint sets for nouns, adjectives and adverbs (except for the TOP class, identical for adjectives and adverbs). This assumption leads to:

$$p(cat_j|pr_j, S, (i), (k_j)) = p(cat_j|S) \tag{10}$$

Since any dependence on (i) and (k_j) is lost, this factor has no impact on the choice of the most probable attachment for (j). However, it is important to note that this assumption relies on the specific semantic resource we have at our disposal, and could be replaced, in other situations, with a 1st-order Markov assumption.

Assumption 5: the gender and number of a nucleus depend on its POS category, the POS category of its regent, and the gender and number of its regent.

In French, the language under study, gender and number agreements take place between the subject and the verb, and between adjectives, or past participles, and the noun they modify/qualify. All, and only, these dependencies are captured in assumption 5 which leads to:

$$\begin{aligned} p(gn_j|cat_j, pr_j, S, (i), (k_j)) = \\ p(gn_j|catj, cat_i, gn_i) \end{aligned} \tag{11}$$

Assumption 6: the lexeme of a nucleus depends only on the POS category and the semantic class of the nucleus itself, the lexeme, POS category and semantic class of its regent lexeme, and the lexeme and POS category of its closest preceding sibling.

This assumption allows us to take bigram frequencies for lexemes into account, as well as the dependencies a given lexeme may have on its closest sibling. In fact, it accounts for more than just bigram frequencies since it leads to:

$$\begin{aligned} p(lex_j|gn_j, cat_j, pr_j, S, (i), (k_j)) = \\ p(lex_j|cat_j, S, cat_i, lex_i, S(i), cat_{k_j}, lex_{k_j}) \end{aligned} \tag{12}$$

Assumptions 1 to 6 lead to a set of probabilities which, except for the last one, can be confidently estimated from training data. However, we still need to simplify equation (12) if we want to derive practical estimations of lexical affinities. This is the aim of the following assumption.

Assumption 7: (i) and (k_j) are independent given (j).

Let us first see with an example what this assumption amounts to. Consider the sequence *eat a fish with a fork*. Assumption 7 says that given *with a fork*, *eat* and *a fish* are independent, that is, once we know *with a fork*, the additional observation of *a fish* doesn't change our expectation of observing *eat* as well, and vice-versa. This does not entail that *with a fork* and *eat* are independent given *a fish*, nor that *a fish* and *with a fork* are independent given *eat*, this last dependence being the one we try to account for. However, this independence assumption is violated as soon as nucleus (k_j) brings more or different constraints on the distribution of nucleus (i) than nucleus (j) does, i.e. when *with a fork* imposes constraints on the possible forms the verb of nucleus (i) (*eat* in our example) can take, and so does *a fish*. With assumption 7, we claim that the constraints imposed by *with a fork* suffice to determine *eat*, and that *a fish* brings no additional information.

Assumption 7 allows us to rewrite equation (12) as:

$$\begin{aligned} p(lex_j|gn_j, cat_j, pr_j, S, (i), (k_j)) = \\ \sum_{S' \in S(i)} \frac{p(S'|lex_i, cat_i)}{p(lex_j|cat_j, S)} \\ \times p(lex_j|cat_j, S, cat_i, lex_i, S') \\ \times p(lex_j|cat_{k_j}, lex_{k_j}) \end{aligned} \tag{13}$$

4 Comparison with other models

It is interesting to compare the proposed models to others previously studied. The probabilistic model described in (Lauer and Dras, 1994), addresses the problem of parsing English nominal compounds. A comparison with this model is of interest to us since the sequences we are interested in contain both verbal and nominal phrases in French. A second model relevant to our discussion is the one proposed in (Ratnaparkhi, 1998), addressing the problem of unsupervised learning for PP attachment resolution in *VERB NOUN PP* sequences. Lastly, the third model, even though used in a supervised setting, addresses the more complex problem of probabilistic dependency parsing on complete sentences [2].

In the model proposed in (Lauer and Dras, 1994), that we will refer to as model L, the quantity denoted as $P(s_j \rightarrow s_i | \exists z : z \rightarrow s_i)$ is the same as the quantity defined by our equation (8). The quantity $p(m)$ in model L is the same as our quantity $p(\mathcal{M}(\mathcal{A}_j))$. There is no equivalent for probabilities involved in equations (9) to (11) in model L, since there is no need for them in analysing English nominal compounds. Lastly, our probability to generate lex_j depends only on S in model L (the dependency on the POS category is obvious since only nouns are considered). For the rest, i.e. the way these core quantities are combined to produce a probability for a parse as well as the decision rule (selection of the most probable parse), there is no difference between the two models. We thus can view our model as a generalization of model L since we can handle PP attachment and take into account indirect independencies.

The model proposed in (Ratnaparkhi, 1998) is similar to a version of our model based solely on equation (9), with no semantic information. This is not surprising since the goal of this work is to disambiguate between prepositional attachment to the noun or to the verb in V N P sequences. In fact, by adding to the set of prepositions an empty preposition, $\mathcal{P}$, the counts of which are estimated from unsafe configurations (that is $c(verb) = \sum_{prep} c(verb, prep) + c(verb, \mathcal{P})$), equation (9) captures both the contribution from the random variable used in (Ratnaparkhi, 1998) to denote the presence or absence of any preposition that is unambiguously attached to the noun or the verb in question, and the contribution from the conditional probability that a particular preposition will occur as unambiguous attachment to the verb or to the noun. We present below the results we obtained with this model.

From the models proposed in (Eisner, 1996), we retain only the model referred to as model C in this work, since the best results were obtained with it. Model C does not make use of semantic information, nor does it rely on nuclei. So the sequence *with a fork*, which corresponds to only one nucleus is treated as a three word sequence in model C. Apart from this difference, model C directly relies on a combination of equations (10) and (12), namely conditioning by lex_i, cat_i and cat_{k_j}, both the probability of generating cat_j and the one of generating lex_j. Thus, model C uses a reduced version of equation (12) and an extended version of equation (10). This extension could be used in our case too, but, since the input to our processing chain consists of tagged words (unless the input of the stochastic dependency parser of (Eisner, 1996)), we do not think it necessary.

Furthermore, by marginalizing the counts for the estimates of our general model, we can derive the probabilities used in other models. We thus view our model as a generalization of the previous ones.

5 Estimation of probabilities

We followed a maximum likelihood approach to estimate the different probabilities our model relies on, by directly computing relative frequencies from our training data. We then used Laplace smoothing to smooth the obtained probabilities and deal with unobserved events.

As mentioned before, we focus on safe configurations to extract counts for probability estimation, which implies that, except for particular configurations involving adverbs, we use only the first nuclei of the chains we arrived at. In most cases, only the first two nuclei of each chain are not ambiguous with respect to attachment. However, since equation (12) relies on (k_j) in addition to (i), we consider the first three nuclei of each chain (but we skip adverbs since their attachment quite often obeys precise and simple rules), but treat the third nucleus as being ambiguous with respect to which nucleus it should be attached to, the two possibilities being *a priori* equi-probable. Thus, from the sequence:

[*implantée*, VERB] (a)
département, NOUN, Masc-Sg, PREP = *dans*(b)
Hérault, NOUN, Masc-Sg, PREP = *de*(c)
(located in the county of Hérault) (En.)

we increment the counts between nuclei (a) and (b) by 1, then consider that nucleus (c) is attached to nucleus (a) and increment the respective counts (in particular the counts associated with equation 12) by 0.5, and finally consider that nucleus (c) is attached to nucleus (b) (which is wrong in this case) and increment the corresponding counts by 0.5.

6 Experiments

We made two series of experiments, the first one to assess whether relying on a subset of our training corpus to derive probability estimates was a good strategy, and the second one to assess the different information sources and probabilities our general model is based on. For all our experiments, we used articles from the French newspaper *Le Monde* consisting of 300000 sentences, split into training and test data.

[2] Other models, as (Collins and Brooks, 1995; Merlo et al., 1998) for PP-attachment resolution, or (Collins, 1997; Samuelsson, 2000) for probabilistic parsing, are somewhat related, but their supervised nature makes any direct comparison impossible.

6.1 Accurate vs. less accurate information

We conducted a first experiment to check whether the accurate information extracted from safe chains was sufficient to estimate probabilities. We focused, for this purpose, on the task of preposition attachment on 200 *VERB NP PP* sequences randomly extracted and manually annotated. Furthermore, we restricted ourselves to a reduced version of the model, based on a reduced version of equation (9), so as to have a comparison point with previous models for PP-attachment. In addition to the accurate information, we used a windowing approach in order to extract less accurate information and assess the estimates derived from accurate information only. Each time a preposition is encountered with a verb or a noun in a window of k (k=3 in our experiment) words, the corresponding counts are incremented.

The French lexicons we used for tagging, lemmatization and chunking contain subcategorization information for verbs and nouns. This information was encoded by several linguists over several years. Here's for example two entries, one for a verb and one for a noun, containing subcategorization information:

quêter - en faveur de, pour
to raise funds - in favor of, for
constance - dans, en, de
- constancy - in, of

Subcategorization frames only contain part of the information we try to acquire from our training data, since they are designed to capture possible arguments, and not adjuncts, of a verb or a noun. In our approach, like in other ones, we do not make such a distinction and try to learn parameters for attaching prepositional phrases independently of their status, adjuncts or arguments. We used the following decision rule to test a method solely based on subcategorization information:

if the noun subcategorizes the preposition,
then attachment to the noun
else if the verb subcategorizes the preposition,
then attachment to the verb
else attachment according to the default rule

and two default rules, one for attachment to nouns, the other to verbs, in order to which of these two alternatives is the best. Furthermore, since subcategorization frames aim at capturing information for specific prepositional phrases (namely the ones that might constitute arguments of a given word), we also evaluated the above decision rule on a subset of our test examples in which either the noun or the verb subcategorizes the preposition. The results we obtained are summarized in table 1.

	Precision
default: noun	0.68
default: verb	0.56
subset	0.75

Table 1: Using subcategorization frames

We then mixed the accurate and less accurate information with a weighting factor α to estimate the probability we are interested in, and let α vary from 0 to 1 in order to see what are the respective impacts of accurate and less accurate information. By using $c1_j$ (resp. $c2_j$) to denote the number of times pr_j occurs with (lex_i, cat_i) in accurate (resp. less accurate) configurations, and by using c_i to denote the number of occurrences of (lex_i, cat_i), the estimation we used is summarized in the following formula:

$$p(pr_j|cat_i, lex_i) = \frac{c1_j + \alpha c2_j + 1}{c_i + n(prep)} \quad (14)$$

where $n(prep)$ is the number of different prepositions introduced by our smoothing procedure. The results obtained are summarized in table 2, where an increment step of 0.2 is used.

α	0	0.2	0.4	0.6	0.8	1
precision	0.83	0.85	0.83	0.81	0.8	0.78

Table 2: Influence of α

These results first show that the accurate information is sufficient to derive good estimates. Furthermore, discounting part of the less accurate information seems to be essential, since the worst results are obtained when $\alpha = 1$. We can also notice that the best results are well above the baseline obtained by relying only on information present in our lexicon, thus justifying a machine learning approach to the problem of PP attachment resolution. Lastly, the results we obtained are similar to the ones obtained by different authors on a similar task, as (Ratnaparkhi, 1998; Hindle and Rooth, 1993; Brill and Resnik, 1994) for example.

6.2 Evaluation of our general model

The model described in Section 3 was tested against 900 manually annotated sequences of nuclei from the newspaper "Le Monde", randomly selected from a portion of the corpus which was held out from training. The average length of sequences was of 3.33 nuclei. The trivial method consisting in linking every nucleus to the preceding one achieves an accuracy of 72.08%.

The proposed model was used to assign probability estimates to dependency links between nuclei in our own implementation of the parser described in (Eisner, 1996). The latter is a "bare-bones" dependency parser which operates in a way very similar to the CKY parser for context-free grammars, in which the notion of a subtree is replaced by that of a **span**. A span consists of two or more adjacent nuclei together with the dependency links among them. No cycles, multiple parents, or crossing dependencies are allowed, and each nucleus not on the edge of the span must have a parent (i.e.: a regent) within the span itself. The parser proceeds by creating spans of increasing size by combining together smaller spans. Spans are combined using the "covered concatenation" operator, which connects two spans sharing a nucleus and possibly adds a dependency link between the leftmost and the rightmost nucleus, or vice-versa. The probability of a span is the product of the probabilities of the dependency links it contains. A span is pruned from the parse table every time that there is another span covering the same nuclei and having the same **signature** but a higher probability. The signature of a span consists of three things:

- A flag indicating whether the span is *minimal* or not. A span is minimal if it is not the simple concatenation of other legal spans;
- A flag indicating whether the leftmost nucleus in the span already has a regent within the span;
- A flag indicating whether the rightmost nucleus in the span already has a regent within the span.

Two spans covering the same nuclei and with the same signiture are interchangeable in terms of the complete parses they can appear in, and so the one with the lower probability can be dropped, assuming that we are only interested in the analysis having the overall highest probability. For more details concerning the parser, see (Eisner, 1996).

A number of tests using different variants of the proposed models were done. For some of those tests, we decided to make use of the subcategorization frame information contained in our lexicon, by extending Laplace smoothing for the probability involved in equation (9) by considering Dirichlet priors over multinomial distributions of the observed data.

We use three different variables to describe the different experiments we made: **se**, being 1 or 0 depending on whether or not we used semantic information, **sb**, which indicates the equivalent sample size for priors that we used in our smoothing procedure for equation (9) (when $sb = 1$, the subcategorization information contained in our lexicon is not used), **Kj**, which is 1 if the variables associated with the closest sister are used in equation (12), and 0 if not. The results obtained with the different experiments we conducted were evaluated in terms of accurracy of attachments against the manually annotated reference. We did not take into account the attachment of the second nucleus in a chain to the first one (since this attachment is obvious). Results are summarized in the following table:

Exp. name	se	sb	kj	Accuracy
base	-	-	-	0.72
exp1	0	1	0	0.662
exp2	0	1	1	0.701
exp3	0	100	1	0.706
exp4	0	200	1	0.705
exp5	1	1	1	0.731
exp6	1	100	1	0.731
exp7	1	200	1	0.735
exp8	1	500	1	0.737
exp9	1	1000	1	0.735

Table 3: General results

7 Discussion

There are two main conclusions we can draw from the preceding results. The first one is that the results are disappointing, in so far as we were not able to really outperform our baseline. The second one is that the best results are achieved with the complete model integrating subcategorization information.

With respect to our model, the difference between experiment 1 and experiment 2 shows that the closest sister brings valuable information to establish the best parse of the chain of nuclei. Even though this information was derived from ambiguous configurations, the extraction heuristics we used does capture actual dependencies, which validates our assumptions 6 and 7. The integration of subcategorization frame information in experiments 3 and 4 does not improve the results, indicating that most of information is already carried in the corresponding version of the general model by bigram lexical statistics. Furthermore, the results obtained with sucategorization information only for parsing V N P sequences do not compare well with an approach solely based on bigram statistics, thus validating the hypothesis behind most work in probabilistic parsing that world knowledge can be approximated, up to a certain extent, by bigram statistics.

The main jump in performance is achieved with the use of semantic classes. All the experiments involving semantic classes yield results over the baseline, thus indicating the well-foundedness of models making use of them. Even though our semantic resource is incomplete (out of 70000 different tokens our corpus comprises[3], only 20000 have an entry in our semantic lexicon), its coverage is still sufficient to constrain word distributions and partly solve the data sparseness problem. The results obtained in previous works relying on semantic classes are above ours (around 0.82

[3]This huge number of tokens can be explained by the fact that the lexicon used for tokenization and tagging integrates many multi-word expressions which are not part of the semantic lexicon

for (Brill and Resnik, 1994) and 0.77 for (Lauer and Dras, 1994)), but a direct comparison is difficult inasmuch as only three-word sequences (V N P, for (Brill and Resnik, 1994) and N N N for (Lauer and Dras, 1994)) were used for evaluation in those works, and the language studied is English. However, it may well be the case that the semantic resource we use does not compare well, in terms of coverage and homogeneity, with WordNet, the semantic resource usually used.

Several choices we made in the course of developing our model and estimating its parameters need now to be more carefully assessed in light of these first results. First of all, our choice to stick with (almost) accurate information, if it leads to good results for the estimation of the probability of generating the preposition of a nucleus given its parent nucleus, may well lead us to rely too often on the smoothing parameters only when estimating other probabilities. This may well be the case for the probability in (12) where bigram statistics extracted with a windowing approach may prove to be more suited to the task. Furthermore, the Laplace smoothing, even though appealing from a theoretical point of view since it can be formalized as assuming a prior over our distributions, may not be fully adequate in the case where the denominator is always low compared to the normalizing contraint, a situation we encounter for equation (12). This may result in over-smoothing and thus prevent our model from accurately discriminating between alternate parses. Lastly, (Lauer and Dras, 1994) uses a prior over the graphs defined by parse trees to score the different parses. We have assumed a uniform prior over graphs, but the results obtained with our baseline clearly indicate that we should weigh them differently.

8 Conclusion

We have presented a general model for PP attachment resolution and NP analysis in French, which generalizes previously proposed models. We have shown how to integrate several different information sources in our model, and how we could use it in an incremental way, starting with simple versions to more complex and accurate ones. We have also presented a series of experiments conducted on a corpus of newspaper articles, and tried to assess the various components of the model, as well as the different information sources used. Our results show that the complete model, making use of all the available information yields the best results. However, these results are still low, and we still need to precisely identify how to improve them.

References

S. Ait-Mokhtar and J.-P. Chanod. 1997. Incremental Finite-State Parsing. *Proceedings of the International Conference on Applied Natural Language Processing*

E. Brill and P. Resnik. 1994. A Rule Based Approach to Prepositional Phrase Attachment Disambiguation. *Proceedings of the 15th International Conference on Computational Linguistics*.

M. Collins and J. Brooks. 1995. Prepositional Phrase Attachment through a Backed-Off Model. *Proceedings of the Third Workshop on Very Large Corpora*.

M. Collins. 1997. Three Generative, Lexicalised Models for Statistical Parsing. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*.

J. Eisner. 1996. Three New Probabilistic Models for Dependency Parsing: An Exploration. *Proceedings of the 16th International Conference on Computational Linguistics*.

D. Hindle and M. Rooth. 1993. Structural Ambiguity and Lexical Relations. *Journal of the Association for Computational Linguistics*. 19(1).

M. Lauer and M. Dras. 1994. A Probabilistic Model of Compound nouns. *Proceedings of the Seventh Joint Australian Conference on Artificial Intelligence*.

P. Merlo, M.W. Croker and C. Berthouzoz. 1997. Attaching Multiple Prepositional Phrases: Generalized Backed-off Estimation. *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*.

A. Ratnaparkhi. 1998. Statistical Models for Unsupervised Prepositional Phrase Attachment. *Proceedings of the joint COLING-ACL conference*.

C. Samuelsson. 2000. A Statistical Theory ofDependency Syntax. *Proceedings of the 18th International Conference on Computational Linguistics*.

A. Yeh and M. Vilain. 1998. Some Properties of Preposition and Subordinate Conjunction Attachment. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics"*.