TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Une méthode non supervisée d'apprentissage sur le Web pour la résolution d'ambiguïtés structurelles liées au rattachement prépositionnel

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

Mots-clefs – Keywords

Analyse syntaxique robuste, grammaires de dépendances, apprentissage non supervisé, désambiguïsation du rattachement prépositionnel.
Robust parsing, dependency grammars, unsupervised learning, prepositional phrase attachment resolution.

Résumé - Abstract

Dans cet article, nous proposons une méthode non supervisée d'apprentissage qui permet d'améliorer la désambiguïsation du rattachement prépositionnel dans le cadre d'un analyseur robuste à base de règles pour le français. Les rattachements ambigus d'une première analyse sont transformés en requêtes sur le Web dans le but de créer un grand corpus qui sera analysé et d'où seront extraites automatiquement des informations lexicales et statistiques sur les rattachements. Ces informations seront ensuite utilisées dans une deuxième analyse pour lever les ambiguïtés des rattachements. L'avantage d'une telle méthode est la prise en compte de cooccurrences syntaxiques et non pas des cooccurrences purement textuelles. En effet, les mesures statistiques (poids) sont associées à des mots apparaissant initialement dans une même relation de dépendance, c'est-à-dire, des attachements produits par le parseur lors d'une première analyse.

In this paper we describe an unsupervised method which improves the disambiguation of prepositional attachments in an existing rule-based dependency parser for French. The results obtained after a first analysis (ambiguous attachments) are transformed into queries to the Web in order to obtain a very big corpus. This big corpus being parsed, lexical and statistical information is extracted to create a database that will be used in a second analysis to disambiguate conflictual attachments. The advantage of such a method is to take into account syntactic co-occurrences as opposed to rough textual co-occurrences. That means that statistical measures (weights) are associated to words already co-occurring in a dependency relation, that is, attachments yield by the parser after a first analysis.

1 Problématique

De façon générale, une des difficultés majeures de tout système d'analyse syntaxique automatique est le traitement des ambiguïtés au niveau de la phrase. Les ambiguïtés structurelles les plus importantes concernent des phénomènes comme la coordination, le repérage de termes complexes et le rattachement prépositionnel. Globalement, le traitement de ces phénomènes implique la prise en compte des ambiguïtés de rattachements des constituants. Dans le cas du rattachement prépositionnel, le problème concerne le lien de dépendance entre un syntagme prépositionnel (SP) et une tête syntaxique (un verbe, un nom ou un adjectif). Selon la position du SP (notamment quand il se trouve après un verbe) il n'est pas toujours possible de déterminer quel est le bon rattachement sur la base de la structure morpho-syntaxique. En effet, les phrases (1) *« Les variations soumettent les particules* **à des mouvements** *... »* et (2) *« Les variations soumettent les particules* **du milieu** *... »* ont la même structure syntagmatique (SV SN SP); cependant, dans (1) le SP se rattache au verbe (*soumettre à mouvements*) alors que dans (2) le SP se rattache au nom (*particules du milieu*).

2 Approche proposée

La résolution des ambiguïtés liées au rattachement prépositionnel se fait, dans notre approche, par la combinaison des sorties produites par l'analyseur lors d'une première analyse avec de l'information lexicale et statistique puisée automatiquement sur le Web. En effet, les dépendances en conflit obtenues lors d'une première analyse sont transformées en requêtes pour le Web dans le but de constituer un corpus de grande taille où vérifier la fréquence d'un patron de cooccurrence syntaxique. La notion de « patron de cooccurrence » est, dans ce contexte, plus large que la notion classique de « rection »: nous ne tenons pas compte du caractère optionnel ou obligatoire de l'élément à rattacher[1], nous sommes uniquement intéressée par la résolution du rattachement ambigu, qu'il soit facultatif ou obligatoire.

2.1 Analyseur à la base

Le modèle d'analyseur que nous avons utilisé est fondé sur la plate-forme d'analyse XIP (Aït-Mokhtar *et al.*, 2002), les propriétés de ce système s'adaptant parfaitement à nos objectifs: ouverture (règles de grammaire facilement enrichissables), souplesse (possibilité d'articuler plusieurs ressources différentes), modularité (facilité d'ajout ou de suppression de grammaires). Pour notre approche, nous avons utilisé la plate-forme XIP et nous avons construit un ensemble de grammaires dans une perspective d'analyse en deux étapes (Gala Pavia, 2003), selon les caractéristiques linguistiques et/ou structurelles des phrases en entrée.

Du point de vue de l'analyse linguistique, les grammaires réalisent un marquage en constituants (syntagmes noyau, *chunks*) et une extraction de dépendances syntaxiques. Une dépendance est ici une relation entre les têtes de deux syntagmes, dans le cas du rattachement prépositionnel, entre deux têtes liées par une préposition. Ce dernier type de dépendance est représenté par `A(X,Prép,N)`, où `X` est la tête syntaxique (indistinctement un verbe, un nom ou un adjectif), `Prép` la préposition et `N` le nom à rattacher. Par exemple, pour la phrase *« Les variations de*

[1] Nous ne faisons pas la différence entre « argument », « modifieur » ou « adjoint ».

cette pression soumettent les particules du milieu à des mouvements vibratoires. » l'analyseur produit la sortie suivante[2]:

```
SUBJ(soumettent,variations)
OBJ(soumettent,particules)
NADJ(mouvements,vibratoires)
NADJ(milieu,vibratoires)
A(variation,de,pression)
A(particule,de,milieu)
A(milieu,à,mouvements)
A(particule,à,mouvements)
A(soumettent,à,mouvements)

240>MAX{NP{Les variations} PP{de NP{cette pression}}
FV{soumettent} NP{les particules} PP{du NP{milieu}}
PP{à NP{des mouvements}} AP{vibratoires} .}
```

Les heuristiques implémentées pour le rattachement prépositionnel permettent d'extraire tous les rattachements possibles lors d'une première analyse (la stratégie est non déterministe uniquement pour ce type de dépendance), d'où l'ambiguïté de rattachement pour *mouvements* dans l'exemple.

2.2 Grammaire de dépendances de l'analyseur

La grammaire de dépendances implémentée pour l'extraction du rattachement prépositionnel contient deux types de règles en ce qui concerne la qualité des relations de dépendance produites. D'une part, il y a des règles (cinq en total) qui décrivent des contextes très sûrs, par exemple un rattachement produit dans SN SP SV. L'évaluation individuelle de ces règles dans nos corpus donne un taux de précision supérieur à 93 %. D'autre part, les autres règles (seize en total) extraient des dépendances dans des contextes moins sûrs et souvent ambigus. Elles ne s'appliquent que si une tête n'a pas été rattachée par le biais d'une règle du premier type.

Globalement, cette grammaire à un taux de précision de 71,34 %, un rappel de 92,10 % et une moyenne F1[3] de 80,40 %. La distinction initiale des règles est marquée par le trait MF1 ou MF2 et permet d'identifier le degré de fiabilité des dépendances produites par l'analyseur lors de la première analyse. Pour l'exemple précédent:

```
A_MF1(variation,de,pression)
A_MF1(particule,de,milieu)
A_MF2(soumettent,à,mouvements)
A_MF2(particule,à,mouvements)
A_MF2(milieu,à,mouvements)
```

[2]Pour les dépendances, SUBJ est le sujet, OBJ l'objet directe, COMP un complément verbal, NADJ un complément nominal. Pour les constituants, MAX est le regroupement maximal, NP un syntagme nominal, AP un syntagme adjectival, PP un syntagme prépositionnel, FV un syntagme verbal conjugué, GV un syntagme verbal participe présent.

[3]F1 = 2(P*R)/P+R

Les dépendances produites par une règle MF2 (« moins fiable ») sont transformées en requêtes et utilisées pour construire la base lexicale à partir du Web.

3 Apprentissage avec le Web

L'utilisation du World Wide Web comme grande base d'exemples pour différentes tâches liées au traitement automatique est une idée exploitée depuis peu: pour la traduction de noms composés (Grefenstette, 1999), pour l'acquisition d'entités nommées (Jacquemin & Bush, 2000), pour la désambiguïsation de relations liées au rattachement prépositionnel (Volk, 2001) et (Lebarbé, 2002).

Dans notre approche, cette ressource nous permet d'obtenir une base de documents de grande taille à partir de laquelle nous construisons une base de patrons de cooccurrence. Pour ce faire, les dépendances « moins fiables » produites par notre grammaire après l'analyse d'un nouveau corpus sont transformées en requêtes pour le Web. Nous faisons l'hypothèse que les occurrences des rattachements à valider sont présentes dans le Web (ou alors elles n'y sont pas si elles sont erronées).

3.1 Requêtes

Chaque requête reprend les trois formes de base (*tokens*) d'une dépendance MF2[4]. Le moteur de recherche choisi a été Altavista (`www.Altavista.com`) avec l'option de recherche avancée qui permet l'utilisation de l'opérateur booléen `NEAR` (une distance de dix mots est possible entre deux mots donnés). Certains des documents obtenus ne sont pas pertinents pour notre tâche, mais nous faisons l'hypothèse que ce problème de pertinence sera pallié par le nombre de bonnes occurrences obtenues dans la grande collection de documents.

3.2 Constitution d'une base lexicale

Pour chaque requête nous avons collecté 20 URLs. Pour ceci nous avons utilisé un ensemble de scripts `perl` combinés avec la commande d'UNIX `wget`. À la fin de cette collecte nous avons obtenu environ 17.000 documents (de tailles variées). Après le nettoyage de marques HTML, nous avons obtenu un nouveau corpus (38.242.073 mots, 1.368.903 phrases) que nous avons analysé avec la même grammaire.

L'analyse de ce grand corpus a donné environ 4 millions de dépendances de type `A(X,Prép,N)`. Nous avons considéré indifféremment les dépendances MF1 et MF2 car nous étions ici intéressées en la *quantité* de dépendances et non en leur *qualité*. Ces dépendances ont alors été transformées en patrons de cooccurrence (`X Prép`) et nous avons calculé des fréquences pour `X` ainsi qu'une mesure d'*estimation de la probabilité de rattachement* (EPR). Cette mesure exprime la fréquence d'apparition du mot `X` par rapport à la cooccurrence de `X Prép`:

$$\texttt{EPR(X,Prép)} = \texttt{fréq(X,Prép)/fréq(X)}$$

[4]Nous avons utilisé les formes et non pas les lemmes car aucun moteur de recherche ne permet d'obtenir toutes les variantes morphologiques à partir d'un lemme.

La base contient environ 100.000 patrons avec leurs poids. L'ensemble de ces informations s'avère crucial lors de la levée des ambiguïtés de rattachement.

4 Levée d'ambiguïtés de rattachement prépositionnel

Nous avons appliqué un algorithme de désambiguïsation pour toutes les dépendances de type MF2 produites lors d'une première analyse. L'algorithme prend cet ensemble de dépendances à valider et compare les patrons de cooccurrence en conflit pour un même élément à rattacher, avec les patrons et leurs poids encodés dans la base d'informations lexicales. Le résultat de cette comparaison aboutit à la proposition d'un seul rattachement et à la suppression des dépendances qui ne correspondent pas au patron choisi.

4.1 Algorithme de désambiguïsation

Lorsque deux patrons `(X1,Prép)` et `(X2,Prép)` (provenant de deux dépendances ayant le même élément `N` à rattacher) existent dans la base de patrons de cooccurrence, leurs poids sont comparés. C'est le patron avec le poids supérieur qui va être gardé.

Pour la phrase en exemple, le poids de `soumettre à` est supérieur à celui de `particule à` (respectivment, 0.7140 et 0.0032). La dépendance `A_MF2(soumettent,à,mouvements)` est donc gardée. Dans l'éventualité où une égalité de poids des patrons se produise, étant donné que l'algorithme ne dispose pas d'autres informations pour lever l'ambiguïté, les deux configurations sont alors gardées.

Autrement, lors de l'existence d'un seul patron de sous-catégorisation, on suppose que le patron inexistant n'a pas été trouvé dans les corpus lors de la construction de la table. Son poids équivaut alors à zéro. Dans ce cas, la comparaison n'a pas lieu car c'est le patron existant dans la table qui est proposé. Pour l'exemple, la dépendance `A_MF2(soumettent,à,mouvements)` est gardée car le patron `A_MF2(milieu,à,mouvements)` n'a pas été trouvé. Quand aucun des patrons n'existe dans la table, l'algorithme se retrouve devant d'un cas similaire à celui évoqué plus haut: à défaut d'autres informations pour lever l'ambiguïté, les deux patrons sont alors gardés.

La liste de dépendances liées au rattachement prépositionnel pour la phrase en exemple est finalement la suivante:

```
A(variation,de,pression)
A(particule,de,milieu)
A(soumettent,à,mouvements)
```

Les deux premières relations ont été produites lors de la première analyse par le biais de règles « fiables » alors que la troisième a été validée grâce à l'algorithme qui prend comme ressource la base d'informations lexicales et statistiques obtenues par apprentissage sur le Web.

4.2 Évaluation

Les résultats de l'application de cette méthode (évaluation des dépendances de type `A(X,Prép,N)`) après la levée d'ambiguïtés donnent un taux de précision de 83,21 %, un rappel de 85,12 % et une moyenne F1 de 84,16 %. Ces résultats obtenus confirment nos hypothèses initiales, à savoir, l'apprentissage sur le Web à partir de cooccurrences syntaxiques améliore les sorties des dépendances liées au rattachement prépositionnel produites uniquement avec des grammaires à base des règles (incrément du taux de précision de 11,84 %). Il est important de signaler que, à la différence de la plupart des méthodes existantes, ces résultats sont obtenus pour tout type d'ambiguïté de rattachement, c'est-à-dire, quelle que soit la configuration des constituants et non seulement pour la configuration classique `SV SN SP`. Des configurations avec plusieurs syntagmes prépositionnels (rattachement multiple) sont ici prises en compte (`SV (SN)+ (SA)+ SP1 SP2 ... SPn`).

5 Conclusion

Dans cet article, nous avons présenté une méthode non supervisée d'apprentissage d'informations sur le Web dans le but d'améliorer le résultat de l'extraction de dépendances liées au rattachement prépositionnel. L'analyseur syntaxique initial est enrichi avec une base de données lexicales (patrons de cooccurrence) et statistiques (poids de cooccurrence syntaxique), puisées automatiquement sur le Web. Au regard des résultats obtenus, la combinaison d'une grammaire de règles avec des techniques d'apprentissage se révèle utile dans le cas de la résolution d'ambiguïtés liées au rattachement prépositionnel.

Références

AÏT-MOKHTAR, S. CHANOD J. P. & ROUX C. (2002). Robustness beyond shallowness: Incremental deep parsing. *Special Issue of the Natural Language Engineering Journal on Robust Methods in Analysis of Natural Language Data*, p. 121–144.

BOURIGAULT D. & FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire*, **25**, 139–151.

GALA PAVIA N. (2003). Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires. Thèse de doctorat, Université de Paris-Sud, UFR Scientifique d'Orsay.

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

JACQUEMIN C. & BUSH C. (2000). Combining lexical and formatting cues for named entity acquisition from the Web. In H. SCHUTZE, Ed., *Proceedings of Joint Sigdat Conference On Empirical Methods In Natural Language Processing And Very Large Corpora (EMNLP/VLC-2000)*, Hong Kong.

LEBARBÉ T. (2002). Hiérarchie inclusive des unités linguistiques en analyse syntaxique coopérative. Thèse de doctorat, Université de Caen.

VOLK M. (2001). Exploiting the WWW as a corpus to resolve PP attachment. In *Proceedings of Conference on Corpus Linguistics*, p. 601–606, Lancaster.