Validation des calculs de relations de dépendance
Une expérience sur le corpus "Internet"

Thomas LEBARBÉ
LIDILEM - EA 609
Université Stendhal - Grenoble 3
thomas.lebarbe@u-grenoble3.fr

Abstract

Nous relatons ici une expérience effectuée dans le cadre de l'analyse syntaxique afin de calculer, au sein d'un empan graphique interne à la phrase (le segment), les relations de dépendance entre les chunks qui s'y trouvent, notamment dans le cas du rattachement prépositionnel complexe (un chunk nominal suivi de plusieurs chunks prépositionnels). En quête d'un corpus d'envergure et d'outils d'extraction performants, nous avons choisi Internet et un moteur de recherche populaire comme ressource langagière, tout en acceptant *a priori* les contraintes que l'un et l'autre peuvent apporter d'un point de vue aussi bien linguistique qu'informatique.

1 Contexte

Nous avons montré l'existence du segment (Lebarbé, 2002), intermédiaire de calcul syntaxique entre le chunk (Abney, 1991) et la phrase, et dont les principales propriétés, observées sur corpus, sont :

1. un chunk dans un segment obéit principalement à une règle de dépendance directe au chunk qui le précéde directement
2. les relations de dépendance d'un segment à un autre sont mises en évidence par des patrons syntaxiques parallèles et similaires d'un segment à l'autre.

Ces règles allègent les calculs de relation de dépendance, notamment la première qui se réduit à une décision par défaut. Néanmoins, cette première règle n'est valide que dans 85 à 90% des cas suivant les corpus analysés. Nous avons donc cherché un moyen pratique, léger (nécessitant peu de ressources langagières) et rapide de valider ou d'invalider ce calcul par défaut de branche unaire de chunks prépositionnels.

2 Hypothèse

Nous avons posé l'hypothèse suivante :

> Un chunk prépositionnel `A` est plus probablement rattaché syntaxiquement à un autre chunk nominal ou prépositionnel `B` qui le précède plutôt qu'à un autre `C` si la paire `A-B` est plus fortement coprésente que la paire `A-C` dans un corpus de grande taille.

Hypothèse qui possède une variante :

> Un chunk prépositionnel `A` est plus probablement rattaché syntaxiquement à un autre chunk nominal ou prépositionnel `B` qui le précède plutôt qu'à un autre `C` si la **chaîne exacte** `AB` est plus fortement coprésente que la **chaîne exacte** `AC` dans un corpus de grande taille.

Afin d'illustrer cette hypothèse, nous proposons l'exemple pédagogique suivant, extrait du corpus d'évaluation de la campagne GRACE, sous-ensemble *Le Monde* :

> *"... la demande d'arrestation du juge ..."*

Intuitivement, il semble raisonnable de concevoir qu'il est question de la demande du juge et non de son arrestation. Cette intuition fait toutefois appel à des connaissances pragmatiques dont le coût de développement et d'intégration serait bien trop élevé en comparaison avec le gain éventuel (et non garanti) en qualité d'analyse syntaxique.

En revanche, il semble tout aussi raisonnable de supposer qu'un corpus suffisamment conséquent présentera plus de co-occurrences de *demande* et *arrestation* que de *juge* et *arrestation*.

3 Méthode, outils et expériences

Nous ne nous apesentirons pas sur les détails de l'intégration logicielle de l'approche que nous présentons ici. Dans notre modèle informatique d'analyse syntaxique par système multi-agent, un agent observe la construction de la structure syntaxique et intervient lorsqu'un segment a été correctement délimité. Toute autre forme d'implantation logicielle est envisageable pour ce type de calcul.

Pour chaque séquence `N pN pN`$^+$, la combinatoire des arborescences possibles est calculée. L'on notera que cette combinatoire n'est pas constituée de l'ensemble des appariements possibles entre chaque paire de chunks de la séquence `N pN pN`$^+$, mais est contrainte par le fait que deux relations de dépendance peuvent se chevaucher mais jamais se croiser (Hays, 1964).

Chaque paire de chunks de cette arborescence est alors utilisée au sein d'une requête dans un moteur de recherche sur Internet (*Google* en l'occurrence), afin d'obtenir le nombre de textes en ligne contenant ladite paire de chunks.

Ces valeurs sont alors utilisées comme poids qualifiant les arbres dans lesquels la paire de chunks apparaît. L'arbre le plus lourd est alors considéré comme le plus probable et est donc retenu dans l'analyse.

Variante 1: Deux types de requêtes ont été envisagés :

- requête stricte par chaîne exacte (pour reprendre notre exemple *"demande du juge"*, *"demande d'arrestation"* et *"arrestation du juge"*)
- requête par substantifs (*demande juge*, *demande arrestation* et *arrestation juge*)

La figure 1 montre les différentes valeurs obtenue. Dans ce cas particulier, le choix du mode de requête ne fait en rien varier la décision. Cependant, l'expérience montre que dans de nombreux cas, certaines co-occurrences strictes sont inexistantes pour une même séquence `N pN pN+`. Il est donc souvent judicieux de choisir la requête par substantif, même si celle-ci ne garantit pas la présence des deux substantifs au sein de la même phrase ou du même contexte.

Variante 2: Il apparaît dans de nombreux travaux que plus la distance entre deux unités linguistiques est longue, moins il est probable que ces deux unités dépendent l'une de l'autre, ne serait-ce que pour des raisons de coût intellectuel aussi bien pour l'éméteur (le locuteur, l'auteur de l'écrit) que pour le récepteur (l'auditeur, le lecteur). Nous avons donc intégré cette notion en pondérant par division les valeurs retournées par le moteur de recherche (cf. la combinatoire et la pondération dans la figure 2).

Les résultats obtenus en utilisant cette méthode ont été évalués sur une partie du corpus de la campagne GRACE ont permis une augmentation de 7% du score de reconnaissance des relations de dépendance.

Toutefois, d'autres expériences ont été menées sur d'autres corpus de domaines moins enclins à être représentés sur l'Internet et ont montré les limites d'une telle démarche. En effet, il s'avère que certaines séquences `N pN pN`$^+$ génèrent des résultats aberrants dont les raisons se rassemblent sous deux rubriques majoristaires :

1. les termes de langue de spécialité peu présents en ligne
2. les noms propres

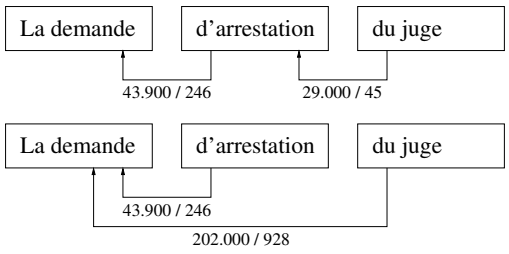

Figure 1: Combinatoire, co-occurrences et relations de dépendance

4 Le Corpus Internet

L'hypothèse présentée dans cette proposition de communication pourrait être considérée comme une assimilation du web à une forme de connaissance pragmatique. Un tel raccourci ne saurait être envisageable. Le terme "corpus Internet" choisi dans le sous-titre de cette proposition de communication est délibérément provocateur.

Le Web ne peut être considéré comme une ressource linguistique à tous les niveaux d'analyse : orthographique, syntaxique, sémantique, pragmatique, terminologique. Du moins ne peut-il pas l'être sans un minimum de "nettoyage" préalable, sans une approche résolument critique. Il n'en est pas moins une source d'écrits multilingue offrant de nombreux avantages : accessibilié, outils d'extraction et moteurs de recherche, relative tolérance des auteurs relativement à leurs droits (qui doivent néanmoins être négociés).

"Internet se joue des obstacles éditoriaux! Les délires les plus dingues étant aussi les plus vifs, le réseau fait depuis sa naissance la part belle aux plus énormes d'entre eux, auxquels des sites sont consacrés, mais qui se propagent également par courrier électronique ou dans les groupes de discussion. C'est ainsi qu'on peut apprendre que Bill Gates, le patron de Microsoft, est l'Antéchrist, qu'Elvis est bien vivant, et que George W. Bush s'emploie à dissimuler les liens du gouvernement américain avec les extraterrestes…"
(G. Dasquié, J. Guisnel, L'effroyable mensonge, thèses et foutaises sur les attentats du 11 septembre, Editions La Découverte, 2002)

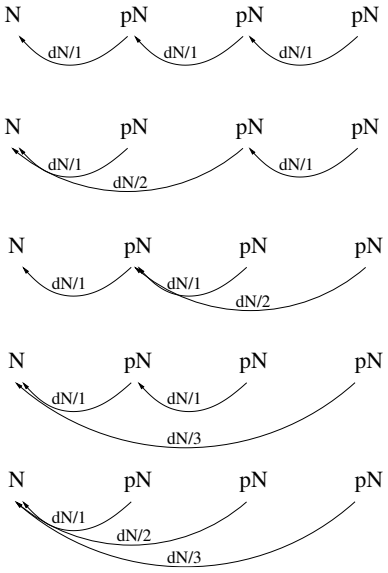

Figure 2: Combinatoire et pondération des longueurs de dépendance

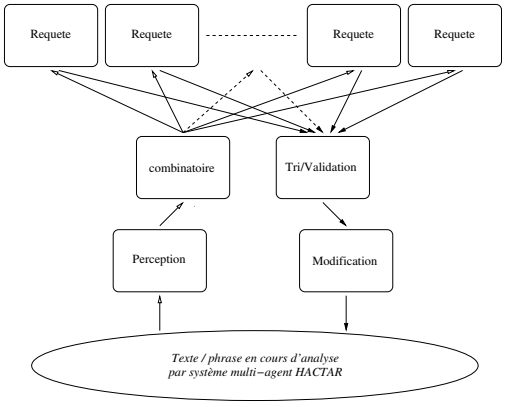

Figure 3: Combinatoire et pondération des longueurs de dépendance

References

ABNEY S. (1991), Parsing by Chunks. *In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, Principle-Based Parsing: Computation and Psycholinguistics*, pages 257–278. Kluwer Academic Publishers, Boston, 1991.

HAYS D. G. (1964), Dependency Theory. in *Language n. 40*, pp. 511-525

LEBARBÉ T. (2002), Hiérarchie Inclusive des Unités Linguistiques en Analyse Syntaxique Coopérative, *Thèse de doctorat, Université de Caen*.