Apprentissage non supervisé de dépendances syntaxiques à partir de texte étiqueté, plusieurs variantes de PCFG légères

Marie Arcadias*, Guillaume Cleuziou**,***, Edmond Lassalle*, Christel Vrain**

*Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
prénom.nom@orange.com,
**LIFO, Université d'Orléans, Rue Léonard de Vinci, 45067 Orléans Cedex 2
prénom.nom@univ-orleans.fr
***GREYC, UCBN, Bd. Maréchal Juin, 14032 Caen Cedex 5

Résumé. L'apprentissage de dépendances est une tâche consistant à établir, à partir des phrases d'un texte, un modèle de construction d'arbres traduisant une hiérarchie syntaxique entre les mots. Nous proposons un modèle intermédiaire entre l'analyse syntaxique complète de la phrase et les sacs de mots. Il est basé sur une grammaire stochastique hors-contexte se traduisant par des relations de dépendance entre les catégories grammaticales d'une phrase. Les résultats expérimentaux obtenus sur des benchmarks attestés dépassent pour cinq langues sur dix les scores de l'algorithme de référence DMV, et pour la première fois des scores sont obtenus pour le français. La très grande simplicité de la grammaire permet un apprentissage très rapide, et une analyse presque instantanée.

1 Introduction et état de l'art

Une structure de dépendances d'une phrase traduit une hiérarchie syntaxique des mots, et permet d'en inférer une sémantique. Les applications liées aux structures de dépendance sont multiples, on peut citer entre autres la modélisation de langages, la reconnaissance d'implications textuelles, les moteurs de question/réponse, l'extraction d'information, l'induction d'ontologies lexicales et la traduction automatique.

Une structure de dépendances d'une phrase (cf. Fig. 1 à gauche) est un arbre dont les nœuds sont les mots, ou tokens, de la phrase. Un des mots est désigné comme la racine de l'arbre (en général un verbe), à laquelle sont attachés des sous-arbres couvrant des portions contigües de la phrase. Un arbre de dépendances est constitué de relations orientées entre un mot syntaxiquement plus fort (tête) et un mot plus faible (dépendant). Le modèle de dépendances est un compromis intéressant entre l'analyse syntaxique classique complète et une représentation "sac de mots".

Les modèles d'apprentissage supervisé de dépendances exigent un nombre important d'exemples annotés à la main. Ce travail, très long et fastidieux, demande une expertise de linguiste essentielle et doit être remanié profondément à chaque nouveau type de texte analysé. La quantité de textes annotés en dépendances est faible comparée à la variété des types de textes disponibles sur le Web. Nous proposons une approche non supervisée ne demandant

qu'une connaissance très superficielle de la langue et du type de texte. Nous nous plaçons dans le cadre de l'Apprentissage Non Supervisé de Dépendances (ANSD).

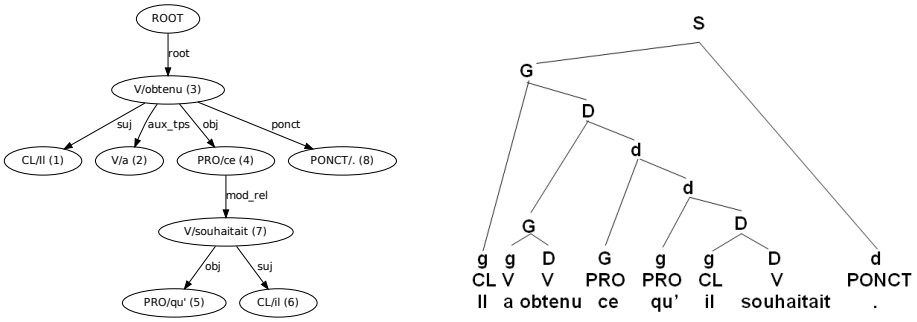

FIG. 1 – *Arbre de dépendances de référence (à gauche) et analyse par grammaire formelle DGdg (à droite).*

Le corpus journalistique américain Penn Treebank propose une version "dépendances" du corpus, donnant les structures de dépendances des phrases. Klein et Manning (2004) furent les premiers à obtenir par ANSD, sur les phrases de moins de 10 mots de ce corpus, des résultats meilleurs que le simple attachement de chaque mot à son voisin de droite, ou que des arbres aléatoires. Ils ont nommé leur modèle *Dependency Model of Valence* (DMV).

L'entrée de cet algorithme est constituée des étiquettes grammaticales des mots des phrases. C'est un modèle génératif dans lequel la racine de la phrase est générée et ensuite, chaque nouvelle tête génère récursivement ses dépendants gauches et droits. La catégorie du dépendant est déterminée en fonction de la tête et de la direction (gauche ou droite). L'apprentissage des probabilités du modèle, dont celles des types de dépendance favorisées, est effectué par une procédure classique de maximisation de l'espérance (EM), basée sur des probabilités a priori calibrées à la main suivant des critères linguistiques.

Ce modèle est riche et intéressant, mais l'initialisation des paramètres du modèle est un problème complexe et essentiel. Elle demande à la fois de l'innovation technique de la part d'un expert en apprentissage et des connaissances linguistiques poussées de la part d'un expert dans la construction syntaxique de la langue étudiée.

2 Apprentissage de grammaire hors contexte

Nous proposons une démarche incluant un apprentissage de grammaire hors contexte probabiliste (PCFG) par l'algorithme Inside-Outside (Lari et Young, 1990), puis une analyse des phrases par une version probabiliste de CYK (Jurafsky et Martin, 2009) appliquée à cette PCFG, suivie d'une phase de « traduction » de l'arbre formel en arbre de dépendances. Nous ne rappelons pas ici les définitions des grammaires formelles. On peut se référer par exemple à Jurafsky et Martin (2009).

Inside-Outside est un algorithme génératif, que l'on peut considérer comme une extension des modèles de Markov cachés (HMM) permettant d'apprendre des grammaires hors contexte probabilistes. Alors que les HMM apprennent les probabilités des règles de dérivation par des calculs sur les sous-séquences précédant et suivant une position t, Inside-Outside se base sur des calculs de sous-séquences à l'intérieur et à l'extérieur de deux positions t_1 et t_2.

CYK probabiliste est un algorithme d'analyse qui choisit parmi toutes les analyses autorisées par les règles de la grammaire, celle qui est la plus probable.

Le formalisme. L'originalité de notre méthode réside dans le choix d'une grammaire simple permettant d'exprimer des dépendances entre les mots d'une phrase. Par exemple, dans la phrase "il a obtenu ce qu'il souhaitait", "obtenu" est un dominant dont dépendent à gauche "il" et "a" et à droite "ce", dominant lui-même "qu'il souhaitait". L'arbre de dépendances de référence est représenté en Figure 1 (gauche). Notre modèle associe à chaque mot (représenté par son étiquette grammaticale) sa qualité de dominant ou dominé par rapport à ses voisins. Pour analyser une phrase, le modèle combine ensuite, par un jeu de symboles, les groupes de mots jusqu'à ce que chaque mot trouve sa place dans l'arbre de dépendances (cf. Figure 1 droite).

À ces fins, nous considérons 5 symboles non terminaux (nt) : le symbole de départ S, G et D représentant des dominants respectivement à gauche et à droite et enfin g et d représentant des dominés. Les terminaux représentent les différentes catégories grammaticales ; elles peuvent différer en fonction de la langue et de l'outil d'étiquetage. Nous présentons ici la grammaire avec des étiquettes universelles :

$$\Sigma = \{ADJ, ADP, ADV, CONJ, DET, NOUN, NUM, PRON, PRT, PUNC, VERB, X\}$$

Les règles de production suivent la forme normale de Chomsky (imposée dans Inside-Outside et CYK). Les contraintes que nous imposons sont les suivantes :

- Un non terminal en majuscule va dominer le non terminal en minuscule auquel il est associé dans une règle de dérivation. Par exemple $G \rightarrow G\ d$ exprime qu'un dominant gauche peut se décomposer en un dominant gauche et un dominé droit.
- Un non terminal gauche g (resp. G) est associé par la gauche à D (resp. d). $nt \rightarrow G\ d$ ou $nt \rightarrow g\ D$.

Le sens que l'on donne aux non terminaux interdit de nombreuses règles, et permet de limiter la taille de la grammaire tout en gardant une signification de grammaire de dépendances latéralisée (la position relative, gauche ou droite, des mots importe). Les règles de Chomsky du premier type ($nt \rightarrow nt\ nt$) sont des règles exprimant la construction interne des phrases, nous parlons de règles de structure. Les règles de Chomsky de second type ($nt \rightarrow terminal$) sont celles par lesquelles on transmet l'information qu'une catégorie peut (ou non) dominer ses voisins de gauche ou de droite. Par exemple, nous interdirons systématiquement pour le français les règles $nt \rightarrow DET$ pour tout $nt \neq g$ car un déterminant est toujours dominé par un nom dont il placé à la gauche.

Les variantes. En fonction de la structure de la langue considérée, les règles de structure de la phrase peuvent ne pas correspondre à la forme intrinsèque des phrases. La grammaire présentée est nommée 4bin car elle contient, en plus du symbole de départ, 4 non terminaux (D, G, d et g) et les règles de structure de la phrase sont écrites de façon binaire, suivant la forme normale de Chomsky.

La signification de ces quatre non terminaux témoigne d'une différenciation essentielle entre les rôles des catégories grammaticales qui domineraient à gauche ou à droite. Pour le français, dans de nombreux cas cette différenciation est pertinente. On peut cependant rencontrer des situations où cette différenciation n'est pas pertinente. Par exemple, deux adjectifs qualifiant le même nom, l'un à sa gauche, l'autre à sa droite, sont dominés par le nom. Nous avons donc envisagé une version 3bin ne comportant que trois non terminaux (en plus de S). Nous avons conservé dans cette variante la latéralisation des dominés g et d, mais n'avons gardé qu'un non terminal dominant N non latéralisé.

	4bin	3bin	4ter	5ter	5ter+
Les différences importantes entre les variantes	Grammaire de base	Le dominant n'est pas latéralisé	On autorise les règles ternaires récursives en D,G,d et g	4ter + non terminal dominant centré N	5ter + règles récursives pour N
Les règles de structure	$nt \rightarrow G\ d$, $nt \rightarrow g\ D$,	$nt \rightarrow N\ d$, $nt \rightarrow g\ N$,	$nt \rightarrow G\ d$, $nt \rightarrow g\ D$, $nt \rightarrow g$ MAJ d,	$nt \rightarrow G\ d$, $nt \rightarrow g\ D$, nt (sauf N) $\rightarrow g$ MAJ d,	$nt \rightarrow G\ d$, $nt \rightarrow g\ D$, $nt \rightarrow g$ MAJ d,
Non terminaux	S, D, G, d, g	S, N, d, g	S, D, G, d, g	S, D, N, G, d, g	S, D, N, G, d, g

TAB. 1 – *Différentes variantes de la grammaire formelle DGdg (MAJ désigne les non terminaux majuscules (G ou D)).*

Le choix de ces deux formes (4bin et 3bin) implique une vision "binaire" des découpages de la phrase en groupes de mots, imposée par la forme normale de Chomsky. Nous pouvons cependant nous affranchir de cette contrainte par la traduction des règles ternaires en règles binaires et envisager ainsi une structure ternaire (4ter) suggérée par les phrases du type : *sujet* (à gauche), *verbe* (au centre) et *complément* (à droite).

Dans la continuité de cette idée, en conservant les rôles latéralisés des dominants D et G, mais en permettant aussi une domination centrale non latéralisée, nous avons introduit un nouveau symbole N (pour neutre) dans les variantes 5ter et 5ter+ qui se distinguent par le fait que l'on interdit (5ter) ou que l'on autorise (5ter+) l'utilisation récursive du symbole N, conduisant à des structures plus complexes dans le dernier cas. Le Tableau 1 résume les différentes variantes proposées.

Le calibrage. Tous les modèles d'ANSD sont calibrés en fonction de la langue du corpus. Ce calibrage consiste à ne sélectionner que les règles du type $nt \rightarrow terminal$ qui ont linguistiquement un sens. Par exemple, en français, un déterminant dépendra toujours d'un nom situé à sa droite ; c'est pourquoi parmi les règles $nt \rightarrow DET$ seule la règle $g \rightarrow DET$ sera conservée. Dans les expérimentations à suivre, le calibrage a été réalisé en observant pour chaque langue quelques arbres donnés comme référence dans les treebanks de dépendances.

3 Expérimentations et résultats

Le French Treebank (Abeillé et Barrier, 2004) donne la structure en constituants (groupes nominaux, verbaux…) ainsi que les fonctions syntaxiques (sujet, objet…) de nombreuses phrases issues d'articles du journal Le Monde. Depuis 2009, ce treebank a été converti en

arbres de dépendances (Candito et al., 2010). Nous avons comparé les arbres appris par notre modèle à ceux donnés en référence dans le treebank de Candito et al. (2010). Pour la comparaison, nous avons utilisé le score UAS (*Unlabeled Attachment Score*) qui calcule, pour un ensemble de phrases, le nombre de dépendances correctes (sans les ponctuations).

Les scores sont différents en fonction des variantes de la grammaire DGdg. Nous obtenons pour 3bin : 29.8%, pour 4bin : 32.9%, pour 4ter : 42.1%, pour 5ter : 37.4% et 5ter+ : 42.2 % ; à titre de comparaison nous obtenons un score de 14.2% pour des arbres générés aléatoirement. On observe que les deux variantes (4ter et 5ter+), qui autorisent des règles ternaires récursives, avec un groupe de mots central dominant et deux groupes latéraux dominés, engendrent des scores pratiquement identiques, nettement supérieurs à ceux obtenus pour les autres variantes. Cela suggèrerait que la structure sous-jacente de ces phrases journalistiques, assez sophistiquées, est mieux capturée par un modèle plus complexe. À notre connaissance, nous sommes les premiers à traiter cette tâche d'ANSD pour le français.

Celle-ci ayant été largement abordée pour l'anglais, puis dans d'autres langues à partir de la conférence CONLL 2006 (Buchholz et Marsi, 2006), nous avons confronté notre modèle à la référence DMV. Le Tableau 2 résume les résultats obtenus, ainsi que la variante ayant engendré le meilleur score. Les résultats pouvent différer beaucoup d'une variante à une autre, celle-ci doit être judicieusement choisie en fonction de la langue et du type de texte.

CONLL 2006 + FTB	Aléatoire	DMV soft-EM	DGdg	Variante utilisée	Temps d'apprentissage	Nb. Mots du corpus	Nb. Catégories distinctes
Allemand	13.1%	**33.3%**	33.1%	4bin	196 min	699 331	52
Anglais	13.4%	38.1%	**39.0%**	5ter+	99 min	937 545	23
Bulgare	16.1%	**39.1%**	23.9%	4bin	23 min	190 217	12
Danois	14.7%	**43.5%**	26.7%	3bin	35 min	94 386	10
Espagnol	13.3%	33.3%	**40.2%**	5ter+	35 min	89 334	15
Français	14.2%	pas de réf.	42.2%	5ter+	320 min	278 083	15
Hollandais	14.8%	21.3%	**34.6%**	4bin	12 min	195 069	13
Japonais	20.7%	56.6%	**64.7%**	5ter+	19 min	151 461	21
Portugais	15.3%	37.9%	**54.0%**	4bin	46 min	206 490	16
Slovène	13.7%	**30.8%**	23.2%	4ter	16 min	28 750	12
Suédois	14.8%	**41.8%**	21.65%	4ter	80 min	191 467	15

TAB. 2 – *Les meilleurs scores UAS obtenus comparés aux références soft-EM données dans Spitkovsky et al. (2011). Les treebank de dépendances des différentes langues proviennent : pour l'Anglais : Marcus et al. (1993), pour le français : Candito et al. (2010); Abeillé et Barrier (2004), les autres langues étant celles de CONLL 2006, Buchholz et Marsi (2006).*

4 Discussion et conclusion

Les temps d'apprentissage dépendent fortement du volume des données et faiblement du nombre de catégories. Le petit nombre de règles de structures de la grammaire permet cependant un apprentissage raisonnable en temps, voire très rapide sur de petits corpus. Une fois la grammaire apprise, l'analyse est quasiment instantanée (quelques secondes pour des milliers de phrases). Ceci argue de la souplesse de notre modèle et de la rapidité de sa mise en œuvre. Celui-ci est donc portable et performant, relativement à DMV. D'autres tests ont révélé que les

résultats peuvent encore être améliorés en considérant des catégories grammaticales plus fines (morpho-syntaxe). Le temps d'apprentissage s'en ressent nécessairement.

Pour améliorer encore notre modèle, nous envisageons d'y intégrer des informations lexicales pour que deux séquences de catégories identiques puissent, en fonction du vocabulaire, être interprétées en arbres de dépendances différents.

Remerciements

Nous remercions Mark Johnson pour ses codes d'Inside-Outside et CYK et Marie Candito pour la mise à disposition de son treebank de dépendance ainsi que pour ses conseils.

Références

Abeillé, A. et N. Barrier (2004). Enriching a french treebank. In *Proc. of LREC 2004, Lisbon*.

Buchholz, S. et E. Marsi (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

Candito, M., B. Crabbé, et P. Denis (2010). Statistical french dependency parsing : treebank conversion and first results. In *LREC 2010*, pp. 1840 à 1847.

Jurafsky, D. et J. H. Martin (2009). *Speech and Language Processing*. Prentice Hall.

Klein, D. et C. D. Manning (2004). Corpus-based induction of syntactic structure : Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on ACL*, pp. 478.

Lari, K. et S. Young (1990). The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech & Language 4*(1), 35 à 56.

Marcus, M. P., M. A. Marcinkiewicz, et B. Santorini (1993). Building a large annotated corpus of english : The penn treebank. *Computational linguistics 19*(2), 313 à 330.

Spitkovsky, V. I., H. Alshawi, et D. Jurafsky (2011). Lateen EM : unsupervised training with multiple objectives, applied to dependency grammar induction. In *EMNLP*, pp. 1269 à 1280.

Summary

Dependency learning is about building a model that allows transforming textual sentences into trees representing a syntactical hierarchy between the words of the sentence. We present an intermediate model between full syntactic parsing of a sentence and bags of words. It is based on a very light probabilistic context free grammar, allowing to express dependencies between the words of a sentence. Our model can be tuned a little in respect of the learned language. Experimentally, we could surpass the scores of the DMV reference on attested benchmarks for five over ten languages, such as English, Portuguese or Japanese. We give the first results on French corpora. Learning is very fast and parsing is almost instantaneous.