PRES LUNAM
Ecole Doctorale STIM
Sciences et Technologies de l'Information et Mathématiques

Spécialité : Informatique
Laboratoire : LINA
Equipe : TALN

Pré-étiquetage syntaxique pour analyse en dépendances du français

Lacroix, Ophélie
Mél : ophelie.lacroix@univ-nantes.fr

Résumé : Nous souhaitons connaître l'influence de l'étiquetage syntaxique appliqué en amont d'une analyse syntaxique en dépendances. L'analyse syntaxique en dépendances vise à relier les mots d'une phrase par des dépendances et à leur attribuer une fonction syntaxique (type). Un analyseur syntaxique produit donc des structures de dépendances qui sont des représentations des phrases du français où les mots sont reliés les uns aux autres par des liens syntaxiques. Cet analyseur est coûteux en temps et ne permet pas d'obtenir un bon taux d'analyses abouties en mode autonome. Nous souhaitons alors utiliser un mode qui sélectionne les types des dépendances qui seront utilisés dans l'analyse. Pour cela nous employons une méthode d'étiquetage syntaxique. Cette méthode consiste à trouver les bonnes fonctions syntaxiques (types) des mots d'une phrase donnée. Cet étiquetage pourra ensuite être utilisé pour aider l'analyseur syntaxique à trouver les bonnes dépendances de la phrase. Cela permettra de réduire le temps d'analyse et d'augmenter le taux de réussite de l'analyseur.

Mots clés : *Étiquetage syntaxique, Analyse syntaxique en dépendances, Dépendances discontinues, Grammaire catégorielle de dépendances*

1 Introduction

L'analyse syntaxique consiste à annoter et regrouper les mots d'une phrase par groupe ou à établir des relations binaires entre les mots de la phrase. La première méthode est une méthode d'analyse par constituent tandis que la deuxième est une méthode d'analyse en dépendance. C'est le cas de l'analyse en dépendance permettant de représenter les phrases par des structures de dépendances qui nous intéresse. Les représentations syntaxiques en dépendance ont tout d'abord été mises en avant par Lucien Tesnière [1]. Puis la théorie Sens-Texte de Mel'cuk [2] a permis d'éclaircir l'utilité syntaxique et sémantique de cette représentation. Pour lui, les structures de dépendances sont plus informatives que les structures par constituants. En effet, les structures par constituants indiquent seulement le rôle de chaque mot ou groupe de mots dans une phrase tandis que les structures de dépendances permettent en plus de lier chaque mot dans la phrase aux mots avec lesquels il a une relation particulière. Ces relations sont des relations binaires (dépendances) entre un gouverneur g et un subordonné s où le type de dépendance d est la fonction syntaxique existante entre g et s ($g \xrightarrow{d} s$). Notre travail se situe au niveau de l'analyse syntaxique en dépendances pour le français. Hors cette langue admet des cas de discontinuité. La figure 1 montre un exemple de discontinuité de la langue française : la négation. Dans les structures de dépendances, on retrouve deux sortes de dépendances : les dépendances projectives et les dépendances discontinues. Les dépendances discontinues peuvent croiser les dépendances projectives.

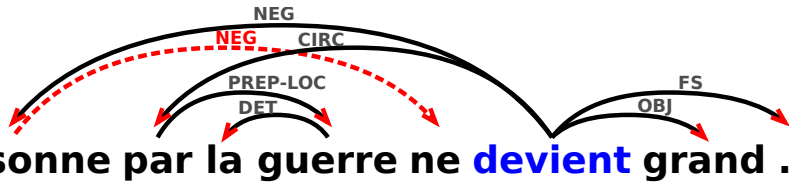

FIGURE 1 – *Structure de dépendances pour la phrase* "Personne par la guerre ne devient grand.". *Les dépendances projectives sont représentées par des lignes pleines noires. Les dépendances discontinues sont représentées par des lignes pointillées rouges. Les types des dépendances sont les noms écrits au dessus de ces dépendances.*

Nous nous intéressons donc particulièrement à l'analyse syntaxique en dépendance gèrant les dépendances discontinues. Pour procéder à une telle analyse nous utilisons un analyseur syntaxique en dépendances que nous présenterons dans la section 3.1. Cependant nous savons que cette méthode d'analyse est coûteuse en temps et ne permet pas d'avoir un bon taux de structures de dépendances en sortie de l'analyseur. Nous souhaitons donc appliquer une méthode d'étiquetage syntaxique en amont de cette analyse pour améliorer la vitesse et le taux d'analyse. Nous expliquons dans la section suivante en quoi consiste l'étiquetage syntaxique.

2 Étiquetage syntaxique

L'étiquetage de phrases du langage naturel consiste à attribuer une étiquette à un mot (ou groupe de mots). Dans le cas de l'étiquetage syntaxique, il s'agit de donner une fonction syntaxique c'est à dire d'attribuer le rôle d'un mot dans une phrase. Ces étiquettes sont en réalité les noms des dépendances utilisées dans les structures de dépendances (c'est à dire les types). Pour la phrase de la figure 1 on obtiendra donc les étiquettes suivantes :

Personne	par	la	guerre	ne	devient	grand	.
NEG	**CIRC**	**DET**	**PREP-LOC**	**NEG**	**SENT**	**OBJ**	**FS**

Notre objectif est ici de trouver automatiquement ces étiquettes syntaxiques à partir d'un apprentissage des étiquettes. Nous disposons d'un ensemble de 2778 phrases (corpus) sur lequel nous souhaitons procéder à un apprentissage puis à un étiquetage. L'apprentissage se fait à l'aide de la méthode des CRF [1] [3]. Il s'agit d'une méthode stochastique qui, sur un ensemble d'exemples, apprend des probabilités sur les caractéristiques (suffixe des mots, existance d'une majuscule, classe grammaticale) des mots du corpus. À partir de ces informations il est possible de retrouver automatiquement quelle est l'étiquette la plus probable pour un mot ou quelle est la séquence d'étiquettes la plus probable pour une phrase donnée. Nous avons donc entraîné l'étiqueteur sur 90% du corpus et choisi de produire automatiquement les 10 meilleures séquences d'étiquettes pour les 10% restants. Nous obtenons donc ainsi, pour chaque mot de chaque phrase, 10 étiquettes au maximum (elles peuvent être redondantes). Pour évaluer ces résultats, nous voulons savoir si, pour chaque phrase, parmi les 1, 1 à 2, 1 à 5 et 1 à 10 meilleures étiquettes attribuées on retrouve la bonne étiquette. Les résultats sont présentés dans la table 1. Ces résultats signifient que 91.6%

Top 1	Top 2	Top 5	Top 10
91.6%	93.7%	96.0%	97.1%

TABLE 1 – *Résultats de l'étiquetage syntaxique sur 1, 2, 5 ou 10 meilleures étiquettes.*

des étiquettes attribuées sont correctes dans la première meilleure séquence d'étiquettes sur l'ensemble des mots du corpus. Lorsque l'on regarde le top 10 on obtient un taux intéressant de 97.1% d'étiquettes correctes. Ce qui signifie que l'on obtient très souvent la bonne étiquette parmi les 10 étiquettes attribuées à chaque mot du corpus. Nous allons donc utiliser ces résultats d'étiquetage pour l'analyse syntaxique en dépendance.

3 Analyse en dépendances

3.1 Grammaire catégorielle de dépendances

Une grammaire est un ensemble de règles permettant de définir la syntaxe d'un langage. Les grammaires peuvent décrire des langages de programmation ou des langages naturels. Ici nous nous intéressons au français et utiliserons une grammaire catégorielle de dépendances du français [4]. C'est une grammaire catégorielle de dépendances car elle utilise des catégories pour décrire ses règles et ces catégories sont les noms des dépendances utilisées dans la représentation des phrases en structures de dépendances. En outre cette grammaire est capable de gérer les dépendances discontinues. Ci-dessous, la figure 2 présente l'arbre de dérivation découlant de l'analyse syntaxique permettant d'obtenir la structure de dépendance de la figure 1. Les catégories des mots de cette phrase sont les suivantes : **Personne**→

1. *Conditional Random Fields* ou champs markovien conditionnels

NEG$^{\nearrow NEG}$, **par**→ CIRC/PREP-LOC, **la**→ DET, **guerre**→DET\PREP-LOC, **ne**→#NEG$^{\searrow NEG}$, **devient**→#NEG\CIRC\NEG\SENT/PT/OBJ, **grand**→OBJ, **.**→FS. On peut voir que dans chaque catégorie on retrouve l'étiquette syntaxique attribuée aux mots de cette phrase. Nous verrons ensuite en quoi cette étiquette est importante. Un arbre de dérivation correct pour une phrase donnée indique que cette phrase est acceptée par la grammaire. Dans notre cas, elle est donc acceptée par la grammaire catégorielle de dépendance du françcais.

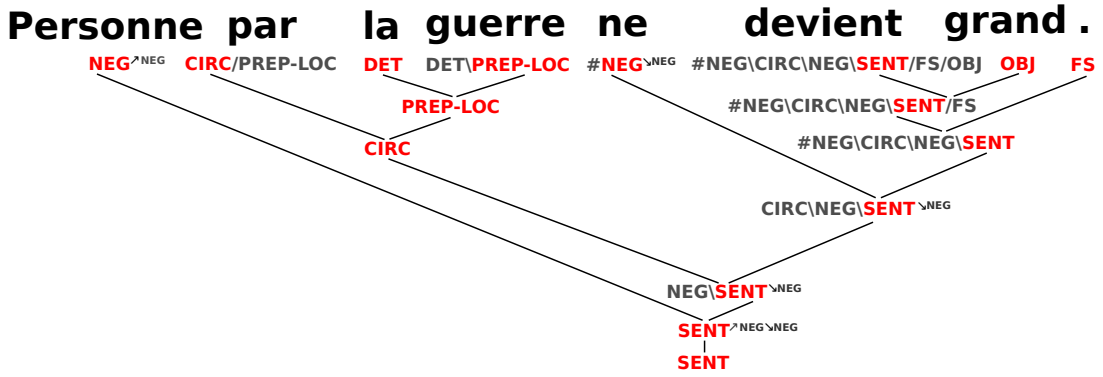

FIGURE 2 – *Arbre de dérivation correct pour la phrase* "Personne par la guerre ne devient grand." *basé sur les catégories de la grammaire catégorielle de dépendances du français. Les noms en rouge sont les étiquettes syntaxiques des mots. Les noms en gris sont les étiquettes des mots qu'ils attendent à gauche ou à droite. La dérivation procède par élimination de ces noms. En bas de l'arbre on obtient l'étiquette SENT qui représente la racine de la phrase et qui signifie que la dérivation est correcte.*

3.2 Analyseur en dépendances

Nous avons vu que l'analyse en dépendances consiste à trouver les dépendances existantes entre les mots d'une phrase, c'est à dire à trouver une bonne structure de dépendances pour cette phrase. Dans la section précédente nous avons noté que trouver un arbre de dérivation pour une phrase signifie que cette phrase est correcte vis-à-vis de la grammaire et permet d'obtenir la structure de dépendances associée. Donc l'analyseur [5] que nous utilisons cherche les bons arbres de dérivations permettant de construire les bonnes structures de dépendances pour les phrases données. Cet analyseur en dépendances est composé de 3 modes d'analyse différents. Nous allons présenter les 2 modes qui nous intéressent ici :

- *l'analyse autonome* est un mode permettant de lancer l'analyse à partir d'une phrase du français brute (sans indiquer d'informations complémentaires). C'est à dire que l'analyseur doit trouver lui-même les catégories des mots pour trouver un arbre syntaxique correct pour cette phrase. Par conséquent l'analyseur produit un nombre important de structures de dépendances à partir des diverses catégories possibles pour une seule phrase.
- *l'analyse par sélection* est un mode semi-automatique. Avant de procéder à l'analyse, l'utilisateur à la possibilité de choisir les types de dépendances qui seront utilisés pour l'analyse. Ainsi l'analyseur connait les fonctions syntaxiques des mots de la phrase et peut en déduire un nombre plus restreint de catégories correctes pour les mots de cette phrase. De la sorte, le nombre de structures de dépendances produites est moindre. Cela permet de réduire le temps d'analyse.

Nous nous intéressons donc au mode d'analyse par sélection. Nous avons vu que dans les catégories associées aux mots de la phrase on retrouve toujours l'étiquette syntaxique de ce mot. Il est le coeur de la catégorie. Connaître cette étiquette pour chaque mot permet de retrouver plus facilement les catégories attribuables à ces mots et donc de trouver plus facilement un bon arbre syntaxique pour la phrase. Nous avons vu dans la section 2 une méthode pour attribuer automatiquement ces étiquettes aux mots d'une phrase. Nous allons donc utiliser ces résultats pour aider l'analyse en dépendances.

3.3 Expérimentations

Pour procéder aux expérimentations, nous étiquetons automatiquement, par la méthode des CRF, les 2778 phrases du corpus. Nous choisissons de garder les 10 meilleures étiquettes de chaque mot car nous

savons qu'il y a plus de chance d'avoir les bonnes étiquettes parmi un choix plus large[2]. Nous exécutons ensuite l'analyse en dépendances sur ces phrases pré-étiquetées. L'analyseur peut produire plusieurs structures de dépendances pour chaque phrase. Nous cherchons alors parmi ces structures celle étant la plus proche d'une structure correcte pour cette phrase et nous évaluons les résultats sur ces structures. Les résultats sont présentés dans la table 2. Tout d'abord, on constate que le nombre de phrases qui ont été

	Nombre de phrases analysées	(%)	Précision (sur les dépendances)	Rappel	Temps d'analyse	(par phrase)
Analyse autonome	1150	(41.4)	98.0%	19.2%	5h13	(6.77s)
Analyse par sélection	2505	(90.2)	96.1%	78.7%	9min24	(0.22s)

TABLE 2 – *Tableau regroupant diverses informations sur les résultats d'une analyse autonome et sur les résultats d'une analyse par sélection basée sur les étiquettes fournies par l'étiquetage syntaxique préalablement réalisé. Il présente le nombre de phrases ayant été analysées (ainsi que le pourcentage sur les 2778 phrases), la précision, le rappel et le temps d'analyse total (et par phrase). La précision est le nombre de dépendances correctes sur le nombre de dépendances assignées (pour les phrases analysées). Le rappel est le nombre de dépendances correctes sur le nombre de dépendances totales (pour toutes les phrases).*

analysées est beaucoup plus important dans le cas de l'analyse par sélection. En effet, lors d'une analyse autonome, un nombre important de phrases ne sont pas analysées par manque de temps car les phrases sont trop longues et/ou l'analyse génère trop de structures de dépendances possibles. Lors de l'analyse par sélection, le nombre de structures de dépendances possibles est réduit et permet à plus de phrases d'être analysées. On obtient donc 90.2% de phrases analysées face à seulement 41.4% au préalable. C'est aussi la raison qui nous permet d'avoir un meilleur taux de rappel. Par ailleurs la précision nous indique que parmi les structures de dépendances produites on trouve très souvent la bonne pour chaque phrase analysée. La dernière remarque importante concerne le temps d'analyse puisque celui-ci est considérablement réduit grâce à la pré-sélection. L'analyseur ayant moins de catégories à prendre en compte et donc moins de structures de dépendances à produire, l'analyse converge plus rapidement vers une solution.

4 Conclusion

Les résultats de l'analyse syntaxique en dépendance contrainte par la sélection des types de dépendances reflète d'une réelle utilité de cette sélection automatique. Dans un premier temps, cette pré-sélection en amont de l'analyse en dépendance permet de réduire de manière significative le temps d'analyse. De nombreuses phrases, d'une longueur conséquente, ne permettant pas d'aboutir à une analyse autonome peuvent être finalement analysées grâce à cette sélection. Ce facteur est très important pour atteindre des taux de rappel intéressant et donc des résultats réellement exploitables. Par ailleurs, en étiquetant syntaxiquement les mots des phrases de 1 à 10 étiquettes différentes, on obtient un bon score en précision. Il nous indique que parmi les structures de dépendances produites par l'analyseur, on obtient très souvent la bonne structure de dépendances pour une phrase donnée.

Références

[1] Lucien Tesnière. *Éléments de syntaxe structurale*. Klincksieck, 1959.

[2] Igor Mel'cuk. *Dependency syntax : Theory and Practice*. State University of New York Press, 1988.

[3] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *ICML 2001*, Williamstown, July 2001.

[4] Alexander Dikovsky. Categorial dependency grammars : from theory to large scale grammars. In *DEPLING 2011*, Barcelona, Spain, September 2011.

[5] Ramadan Alfared, Denis Béchet, and Alexander Dikovsky. "CDG Lab" : a Toolbox for Dependency Grammars and Dependency Treebanks Development. In *Proceedings of DEPLING 2011*, pages 272–281, Barcelona, Spain, September 2011.

2. Les résultats étaient de 97.1% (voir section 2).