Grammaire à substitution d'arbre de complexité polynomiale : un cadre efficace pour DOP

Jean-Cédric Chappelier et Martin Rajman
EPFL
DI-LIA, IN (Écublens)
CH-1015 Lausanne, Switzerland
{Jean-Cedric.Chappelier,Martin.Rajman}@epfl.ch

Résumé - Abstract

Trouver l'arbre d'analyse le plus probable dans le cadre du modèle DOP (Data-Oriented Parsing) — une version probabiliste de grammaire à substitution d'arbres développée par R. Bod (1992) — est connu pour être un problème NP-difficile dans le cas le plus général (Sima'an, 1996a). Cependant, si l'on introduit des restrictions *a priori* sur le choix des arbres élémentaires, on peut obtenir des instances particulières de DOP pour lesquelles la recherche de l'arbre d'analyse le plus probable peut être effectuée en un temps polynomial (par rapport à la taille de la phrase à analyser). La présente contribution se propose d'étudier une telle instance polynomiale de DOP, fondée sur le principe de sélection miminale-maximale et d'en évaluer les performances sur deux corpus différents.

Finding the most probable parse tree in the framework of Data-Oriented Parsing (DOP), a Stochastic Tree Substitution Parsing scheme developed by R. Bod (1992), has proven to be NP-hard in the most general case (Sima'an, 1996a). However, introducing some *a priori* restrictions on the choice of the elementary trees leads to interesting DOP instances with polynomial time-complexity. The purpose of this paper is to present such an instance, based on the minimal-maximal selection principle, and to evaluate its performances on two different corpora.

1 Motivations

Introduite par R. Scha (Scha, 1990) puis développée par R. Bod (Bod, 1992; Bod, 1998), l'approche Data-Oriented Parsing (DOP) pour l'analyse syntaxique probabiliste a depuis été largement étudiée par diverses équipes de recherche. La principale limitation de cette approche reste cependant le caractère NP-difficile du problème consistant à trouver l'arbre d'analyse le plus probable (MPP, pour *most probable parse*) (Sima'an, 1996a). Différentes solutions approchées (heuristiques) ont été proposées (Bod, 1992; Goodman, 1996; Chappelier & Rajman, 1998) mais une autre direction de recherche prometteuse consiste à explorer différentes restrictions *a priori* du jeu d'arbres élémentaires utilisé par la grammaire DOP, restrictions pour lesquelles trouver l'arbre d'analyse le plus probable n'est plus un problème NP-difficile. L'objet du présent article est de présenter et d'évaluer un exemple de telle restriction.

Nous commencerons tout d'abord par une brève introduction au Data-Oriented Parsing et définirons quelques notations. Puis nous présenterons dans la seconde partie un principe de sélection d'arbres élémentaires — le principe de sélection minimale-maximale – permettant d'obtenir une restriction de DOP ayant une complexité polynomiale (pour le problème du MPP). Nous terminerons enfin par la présentation de plusieurs expériences sur deux corpus différents.

2 Data-Oriented Parsing

2.1 Le modèle DOP

DOP est un modèle d'analyse syntaxique probabiliste à base de grammaires à substitution d'arbres (STSG pour *Stochastic Tree-Substitution Grammars*). Une STSG est une grammaire pour laquelle les règles sont des arbres élémentaires[1] que l'on peut combiner à l'aide de l'opérateur de substitution[2] pour obtenir les dérivations des arbres d'analyse.

Dans sa définition la plus générale, le modèle DOP utilise comme arbres élémentaires l'ensemble de *tous* les sous-arbres des arbres d'analyse contenus dans un corpus annoté (treebank) disponible pour l'entraînement du modèle. Chaque arbre élémentaire t est associé à une probabilité élémentaire $p(t)$ qui est proportionnelle au nombre d'occurrences de t dans le corpus d'entraînement.

La probabilité $p(d)$ d'une dérivation d est alors définie par[3] $p(d) = \prod_{t \in d} p(t)$ et la DOP-probabilité d'un arbre d'analyse T est donnée par :

$$P_{\text{DOP}}(T) = \sum_{d \Rightarrow T} p(d) = \sum_{d \Rightarrow T} \prod_{t \in d} p(t)$$

où "$d \Rightarrow T$" signifie "pour toute dérivation d produisant l'arbre d'analyse T"[4].

2.2 Analyse syntaxique la plus probable

L'analyse syntaxique d'une phrase s'effectue généralement en deux étapes distinctes :

l'analyse proprement dite, dont le but est de produire une représentation compacte de l'ensemble des arbres d'analyse possibles pour la phrase analysée ;

l'exploitation des résultats à partir de la structure compacte précédemment construite. Cette étape peut par exemple consister à présenter l'ensemble de la forêt d'analyse, à extraire l'arbre d'analyse le plus probable, à extraire la dérivation la plus probable[5], etc...

Dans le cas des STSGs, l'étape d'analyse peut, comme dans le cas des grammaires hors-contexte, être réalisée en un temps polynomial (cubique) par rapport à la taille de la phrase à analyser. Par contre, l'extraction de l'arbre d'analyse le plus probable (MPP) constitue, dans le cas général, un problème NP-difficile (Sima'an, 1996a) et ne peut donc pas s'effectuer en

[1] comme pour les TAG (Tree Adjoining Grammars)

[2] mais pas de l'opérateur d'adjonction, contrairement aux TAG

[3] $t \in d$ représente le fait que le sous-arbre t participe à la dérivation d.

[4] Un même arbre d'analyse peut en effet posséder plusieurs dérivations différentes (et ce malgré la convention de réécriture de la feuille non-terminale la plus à gauche en premier).

[5] Dans le cas de DOP, et à la différence des grammaires hors-contexte probabilistes classiques, l'arbre d'analyse le plus probable et l'arbre associé à la dérivation la plus probable ne sont pas nécessairement identiques.

général de façon efficace (polynomiale). Notons toutefois que la **dérivation** la plus probable peut par contre être trouvée en un temps cubique en utilisant les algorithmes usuels développés pour les grammaires hors-contexte probabilistes.

Pour contourner le caractère NP-difficile de la recherche du MPP dans le modèle DOP, diverses heuristiques correspondant à des solutions approchées ont été proposées : Monte-Carlo Sampling (Bod, 1992), General Recall (Goodman, 1998), échantillonage contrôlé (Chappelier & Rajman, 1998; Chappelier & Rajman, 2000).

Une autre approche possible, développée dans cette contribution, consiste à explorer les restrictions du modèle DOP pour lesquelles la recherche de l'arbre d'analyse le plus probable peut se faire de façon exacte en un temps polynomial (cubique) par rapport à la taille de la phrase à analyser. L'idée générale d'une telle approche consiste à limiter *a priori* l'ensemble des arbres élémentaires constituant la grammaire du modèle DOP de telle sorte que l'on puisse exhiber une grammaire hors-contexte probabiliste qui soit équivalente à la STSG utilisée, c.-à-d. telle que pour tout arbre d'analyse de DOP-probabilité maximale dans le modèle DOP, il existe au moins une dérivation dans la grammaire hors-contexte équivalente dont la probabilité

1. est égale à la DOP-probabilité de cet arbre d'analyse ;
2. est maximale (parmi les probabilités de toutes les dérivations produites pour la phrase analysée par la grammaire hors-contexte équivalente).

Une STSG pour laquelle une grammaire hors-contexte probabiliste équivalente peut être construite sera appelée « *grammaire probabiliste à substitution d'arbres polynomiale* » (PSTSG)[6].

Notons que, dans ce cas, la complexité algorithmique liée à la sommation des probabilités sur les diverses dérivations d'un même arbre d'analyse peut être efficacement factorisée[7].

Tout l'enjeu de l'approche proposée ici est donc de trouver diverses restrictions du modèle DOP correspondant à des PSTSG. Un exemple trivial d'une telle restriction est donné par le modèle DOP où les arbres élémentaires sont limités aux sous-arbres de profondeur 1. Ce modèle DOP est strictement équivalent à une grammaire hors-contexte probabiliste (SCFG) et ne présente donc aucun apport par rapport aux SCFG usuelles[8].

Dans la suite de cet article, nous nous proposons d'examiner une autre PSTSG, moins triviale et linguistiquement plus intéressante. Elle correspond à la restriction du modèle DOP dans laquelle les arbres élémentaires sont limités aux sous-arbres de profondeur 1 **et** aux sous-arbres totalement ancrés, i.e. les sous-arbres dont les feuilles sont toutes des terminaux de la grammaire (i.e. des mots). Ce choix pour les arbres élémentaires sera appelé *principe de sélection minimale-maximale* [9] et la restriction du modèle DOP correspondante sera appelée sa *restriction minimale-maximale*.

Nous allons maintenant montrer qu'une telle STSG est effectivement polynomiale, c.-à-d. qu'il existe une grammaire hors-contexte qui lui soit équivalente au sens défini ci-dessus.

[6] pour *Polynomial Stochastic Tree Substitution Grammar*. « polynomiale » fait ici référence à la complexité du problème MPP pour une telle grammaire.

[7] C'est précisément cette sommation nécessaire qui est la source du caractère NP-difficile de la recherche du MPP dans le cas général.

[8] En particulier la probabilisation des analyses est strictement la même.

[9] Ce principe de sélection a été introduit pour la première fois par Jacques Han dans son travail doctoral non encore publié.

3 Restriction minimale-maximale de DOP

3.1 Définition de la grammaire hors-contexte équivalente

Pour toute grammaire DOP $\mathcal{G}_{\text{DOP}}$, soit $\mathcal{G}_{\text{equiv}}$ la grammaire hors-contexte contenant toutes les règles racine-feuilles[10] associées à tous les arbres élémentaires de $\mathcal{G}_{\text{DOP}}$.[11] D'un point de vue purement structurel (i.e. sans les probabilités) $\mathcal{G}_{\text{DOP}}$ est équivalente à $\mathcal{G}_{\text{equiv}}$.[12] Cette propriété est toujours vraie, même dans le cas général de DOP, mais dans le cas particulier des restrictions minimales-maximales, cette équivalence peut être étendue aux probabilités. Plus précisément, il devient alors possible de probabiliser $\mathcal{G}_{\text{equiv}}$ de sorte que trouver la dérivation la plus probable au sens de $\mathcal{G}_{\text{equiv}}$ soit équivalent à trouver l'**analyse** la plus probable au sens de $\mathcal{G}_{\text{DOP}}$.

La solution permettant d'obtenir une telle équivalence consiste à associer à chaque règle de $\mathcal{G}_{\text{equiv}}$ un coefficient stochastique correspondant à la DOP-probabilité de l'arbre élémentaire correspondant. Remarquez bien la différence avec la démarche usuelle qui consiste à associer à la règle la probabilité élémentaire p de l'arbre élémentaire et non pas sa DOP-probabilité.

Nous montrerons dans la section 3.3 qu'une telle probabilisation de $\mathcal{G}_{\text{equiv}}$ est facile à réaliser lors de la construction de $\mathcal{G}_{\text{DOP}}$ à partir du corpus d'entraînement. Mais montrons tout d'abord qu'avec une telle probabilisation, trouver la dérivation la plus probable avec $\mathcal{G}_{\text{equiv}}$ est effectivement équivalent à trouver l'analyse la plus probable au sens de $\mathcal{G}_{\text{DOP}}$.

3.2 Démonstration de l'équivalence probabiliste

L'équivalence à établir repose sur la propriété générale suivante : la probabilité d'une dérivation dans $\mathcal{G}_{\text{equiv}}$ est toujours inférieure ou égale à la DOP-probabilité de l'arbre d'analyse de $\mathcal{G}_{\text{DOP}}$ équivalent. La démonstration de cette propriété, trop longue pour être donnée ici, peut être trouvée dans (Chappelier & Rajman, 2001). À l'aide de cette propriété, pour montrer qu'il est équivalent de trouver la dérivation la plus probable au sens de $\mathcal{G}_{\text{equiv}}$ et l'arbre d'analyse le plus probable au sens de $\mathcal{G}_{\text{DOP}}$, il est donc suffisant de montrer que, pour chaque arbre d'analyse T au sens de $\mathcal{G}_{\text{DOP}}$, il *existe au moins une* dérivation équivalente dans $\mathcal{G}_{\text{equiv}}$ dont la probabilité *est égale à* la DOP-probabilité de T.

Cette dernière propriété peut se démontrer par récurrence sur la profondeur de T :
1) La propriété est trivialement vraie pour tout arbre de profondeur 1, puisque dans ce cas il n'y a qu'une dérivation possible de l'arbre d'analyse et la DOP-probabilité de cet arbre est alors, par construction, égale à la probabilité de cette dérivation.
2) Supposons maintenant que pour tout arbre d'analyse (au sens de $\mathcal{G}_{\text{DOP}}$) T de profondeur au plus n, il existe dans $\mathcal{G}_{\text{equiv}}$ une dérivation d équivalente à T dont la probabilité est la DOP-probabilité de T. Nous devons alors établir que cette proposition est également vraie pour tout arbre d'analyse de profondeur $n+1$.

[10] Pour tout arbre T, la règle racine-feuilles associée $R(T)$ est la règle hors-contexte dont la partie gauche est la racine de T et la partie droite la séquence des feuilles de T.

[11] Notez qu'une règle hors-contexte différente est créée pour chacun des arbres élémentaires. Les règles associées à des arbres élémentaires différents mais ayant même racine et mêmes feuilles sont différenciées par leur indices. Il y a donc bijection entre les règles de $\mathcal{G}_{\text{equiv}}$ et les arbres élémentaires de $\mathcal{G}_{\text{DOP}}$.

[12] Il y a en effet bijection entre les dérivations de ces deux grammaire. La reconstruction de l'arbre d'analyse complet à partir d'une dérivation dans $\mathcal{G}_{\text{equiv}}$ peut être effectuée de façon triviale à l'aide des indices des règles de $\mathcal{G}_{\text{equiv}}$ qui référencent également les arbres élémentaires de $\mathcal{G}_{\text{DOP}}$.

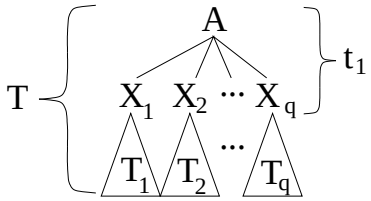


Figure 1: Dérivation avec une grammaire suivant le principe de restriction minimale-maximale de DOP : toute dérivation d'un arbre d'analyse donné (T ici), qui n'est pas lui-même un arbre élémentaire, commence par un arbre élémentaire de profondeur 1 (t_1 ici), lequel est partagé par toutes les dérivations de l'arbre d'analyse considéré.

Soit maintenant $T \Rightarrow^* W_1^p$ un arbre d'analyse (de $\mathcal{G}_{\text{DOP}}$) de profondeur $n+1$ de la chaîne W_1^p. Si T est lui-même un arbre élémentaire de la grammaire $\mathcal{G}_{\text{DOP}}$, alors la règle hors-contexte racine-feuilles $R(T)$ qui lui est associée est une règle de $\mathcal{G}_{\text{equiv}}$ dont la probabilité est par construction la DOP-probabilité de T. Et puisque $R(T)$ est une dérivation de W_1^p, il existe au moins une dérivation représentant T dans $\mathcal{G}_{\text{equiv}}$ (et dont la probabilité est la DOP-probabilité de T).

Si au contraire T n'est pas un arbre élémentaire de $\mathcal{G}_{\text{DOP}}$, alors T possède au moins une dérivation. Notons-la $d = t_1 \circ ... \circ t_k$ ($k > 1$).

Il est alors important de remarquer qu'en raison du choix des arbres élémentaires imposé par la sélection minimale-maximale, t_1 est nécessairement un arbre de profondeur 1. Il est de plus partagé par toutes les dérivations de T. La DOP-probabilité de T est donc :

$$\begin{aligned} P_{\text{DOP}}(T) &= \sum_{d' \Rightarrow T} \prod_{t \in d'} p(t) = p(t_1) \sum_{d_1 \Rightarrow T_1} \cdots \sum_{d_q \Rightarrow T_q} \prod_{i=1}^{q} p(d_i) \\ &= p(t_1) \cdot \prod_{i=1}^{q} \left(\sum_{d_i \Rightarrow T_i} p(d_i) \right) = p(t_1) \cdot \prod_{i=1}^{q} P_{\text{DOP}}(T_i) \end{aligned}$$

en notant T_1, ..., T_q les sous-arbres totalement ancrés de T qui sont fils de t_1 dans d (cf fig 1).

Or, puisque T_1, ..., T_q sont de profondeur au plus n, il existe par hypothèse de récurrence pour chaque T_i une dérivation $d_{\text{equiv}}(T_i)$ dans $\mathcal{G}_{\text{equiv}}$ dont la probabilité (dans $\mathcal{G}_{\text{equiv}}$) est égale à $P_{\text{DOP}}(T_i)$.

La dérivation $(t_1, d_{\text{equiv}}(T_1), ..., d_{\text{equiv}}(T_q))$ est une dérivation de T au sens de $\mathcal{G}_{\text{equiv}}$ dont la probabilité est[13] $p(t_1) \cdot P_{\text{DOP}}(T_1) \cdot ... \cdot P_{\text{DOP}}(T_q)$, c.-à-d. la DOP-probabilité de T.

Il existe donc au moins une dérivation de T au sens de $\mathcal{G}_{\text{equiv}}$ dont la probabilité est égale à la DOP-probabilité de T. □

Rechercher la dérivation la plus probable au sens de $\mathcal{G}_{\text{equiv}}$ est donc bien équivalent à rechercher l'arbre d'analyse le plus probable au sens de $\mathcal{G}_{\text{DOP}}$.

3.3 Construction pratique de la grammaire hors-contexte équivalente

Il nous faut maintenant expliquer comment la grammaire hors-contexte équivalente $\mathcal{G}_{\text{equiv}}$ peut être construite à partir du corpus d'entraînement utilisé pour le modèle DOP considéré.

[13] par construction (dérivation hors-contexte)

Les arbres de profondeur 1 sont tout d'abord extraits, de façon identique à la constitution d'une grammaire hors-contexte à partir d'un corpus annoté. Puis, pour chaque nœud de chaque arbre du corpus, le sous-arbre totalement ancré correspondant est extrait et la règle racine-feuilles associée est ajoutée aux règles de $\mathcal{G}_{\text{equiv}}$ (en regroupant les occurrences multiples correspondant au **même** arbre élémentaire mais en différenciant, par leurs indices, les règles racine-feuilles identiques provenant d'arbres élémentaires différents).

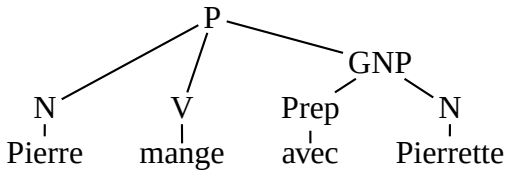
La DOP-probabilité de chaque sous-arbre totalement ancré est ensuite calculée par ordre de profondeur d'arbre croissante. En effet, si la DOP-probabilité de tout arbre élémentaire de profondeur n a déjà été calculée, alors la DOP-probabilité de tout arbre élémentaire T de profondeur $n+1$ peut facilement être calculée par la formule suivante :

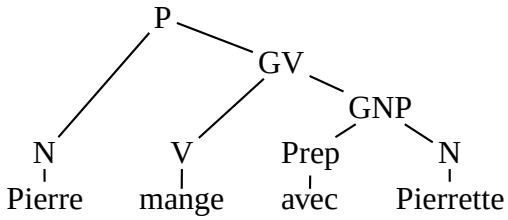
$$P_{\text{DOP}}(T) = p(T) + p(t_1) \cdot \prod_{i=1}^{q} P_{\text{DOP}}(T_i)$$

4 Un exemple

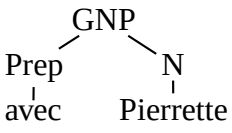
Considérons maintenant un exemple jouet illustrant l'approche précédemment détaillée et montrant en quoi notre approche diffère du modèle DOP général.

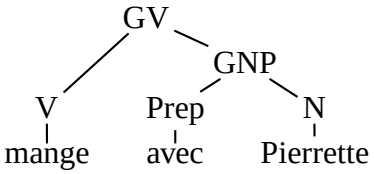
Considérons le corpus d'entraînement trivial contenant 2 arbres suivants :

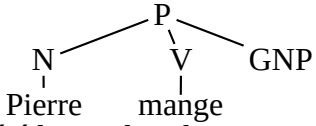
$t_1 =$ 

$t_2 =$ 

Pour construire la STSG $\mathcal{G}_{\text{DOP}}$ suivant le principe de sélection minimale-maximale, les sous-arbres de profondeur 1 ainsi que tous les sous-arbres totalement ancrés doivent être extraits ; ce qui donne les 12 arbres suivants : 8 arbres de profondeur 1, t_1 et t_2 eux-mêmes, et les deux sous-arbre totalement ancrés suivants :





Remarquons qu'un arbre comme par exemple 

n'appartient pas à $\mathcal{G}_{\text{DOP}}$, alors que cela aurait été le cas dans la version non restreinte de DOP.

Comme $\mathcal{G}_{\text{DOP}}$ contient 12 arbres élémentaires, la grammaire hors-contexte équivalente $\mathcal{G}_{\text{equiv}}$ contient 12 règles, lesquelles sont données dans la table 1. Remarquez bien que la règle `P -> Pierre mange avec Pierrette` apparaît deux fois dans la grammaire. Ceci est dû au fait que chacune de ces deux règles correspond à un arbre élémentaire de $\mathcal{G}_{\text{DOP}}$ différent.

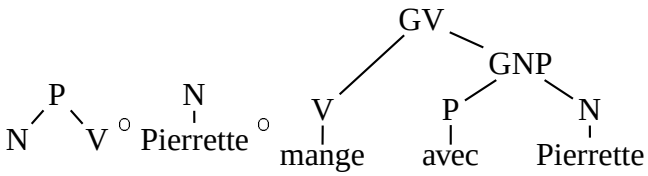
Notez aussi que $\sum_\alpha P(X \to \alpha)$ **n'**est **pas** égale à 1 dans le cas général, c.-à-d. que $\mathcal{G}_{\text{equiv}}$ n'est

règle	P_{DOP}	p
r_1: `P -> N V GNP`	0.25	0.25
r_2: `P -> N GV`	0.25	0.25
r_3: `P -> Pierre mange avec Pierrette`	0.344	0.25
r_4: `P -> Pierre mange avec Pierrette`	0.359	0.25
r_5: `N -> Pierre`	0.5	0.5
r_6: `N -> Pierrette`	0.5	0.5
r_7: `V -> mange`	1.0	1.0
r_8: `Prep -> avec`	1.0	1.0
r_9: `GNP -> Prep N`	0.5	0.5
r_{10}: `GNP -> avec Pierrette`	0.75	0.5
r_{11}: `GV -> V GNP`	0.5	0.5
r_{12}: `GV -> mange avec Pierrette`	0.875	0.5

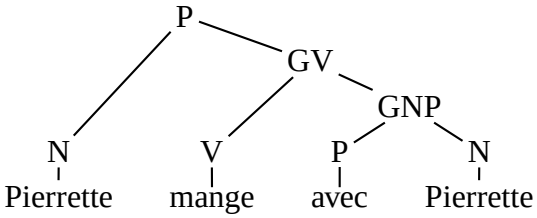
Table 1: Les 12 règles de la grammaire hors-contexte équivalente. Le coefficient stochastique est donné par P_{DOP}. p correspond à la probabilité élémentaire de l'arbre élémentaire dans la grammaire DOP d'origine.

pas une grammaire probabiliste « *propre*[14] ». $\mathcal{G}_{\text{equiv}}$ doit simplement être vue comme un moyen efficace d'implémenter la recherche de l'arbre d'analyse le plus probable dans le modèle DOP (restreint) et non pas comme une grammaire stochastique en tant que telle. En particulier, le modèle probabiliste sous-jacent reste bien celui associé au modèle DOP et non pas celui d'une grammaire hors-contexte probabiliste (propre).

Si l'on considère maintenant la phrase s="*Pierrette mange avec Pierrette*", sa dérivation la plus probable dans $\mathcal{G}_{\text{equiv}}$ est $d = r_2 \circ r_6 \circ r_{12}$ de probabilité $p(d) = 0.25 \cdot 0.5 \cdot 0.875 = 0.109$. Cette dérivation correspond dans $\mathcal{G}_{\text{DOP}}$ à la dérivation suivante :



et l'arbre d'analyse le plus probable (pour s dans $\mathcal{G}_{\text{DOP}}$) est donc :



de DOP-probabilité égale à 0.109.

[14]Traduction libre du terme *proper* introduit dans (Booth & Thompson, 1973)

corpus	nombre de phrases annotées	nombre de règles hors-contexte	nombre de non-terminaux	nombre de terminaux	nombre de PoS tags	longueur moyenne de phrase	nb moyen de règles par phrase
ATIS	1'381	1'027	40	1'167	38	12.5	23.3
Susanne	6'728	20'302	767	17'863	130	20.4	36.0
Susanne réduit	4'000	8'882	469	10'284	122	12.9	23.8

Table 2: Différentes caractéristiques des 2 corpus utilisés pour les expériences.

5 Expériences

Pour évaluer de façon expérimentale les performances du modèle DOP restreint par le principe de sélection minimale-maximale, nous avons utilisé deux corpus différents : le corpus ATIS (Hemphill *et al.*, 1990) et le corpus Susanne3 (Sampson, 1994). Pour des raisons de taille de la grammaire complète du modèle restreint nous avons cependant dû utiliser une version réduite du corpus Susanne3 : parmi les 6803 phrases d'origine, nous n'avons finalement utilisé que 4000 d'entre elles (ce corpus est appelé « Susanne réduit » ci-après). Les caractéristiques des ces corpus sont données en table 2.

Contrairement à la plupart des expériences faites sur DOP jusqu'à présent (Bod, 1998; Sima'an, 1996b; Goodman, 1996), nous n'avons pas mis les corpus sous forme normale de Chomsky (arbres binaires), mais sommes plutôt restés le plus proche possible des données d'origine[15]. Dans le même ordre d'idée, nous n'avons pas tronqué les arbres au niveau des étiquettes morpho-syntaxiques mais avons travaillé sur les phrases elles-mêmes (i.e. au niveaux des mots).

La méthodologie d'évaluation utilisée pour la production de ces résultats a été la même pour toutes les expériences :

- partitionner (de façon aléatoire) le corpus en un corpus d'entraînement (90%) et un corpus de test (les 10% restants) ;
- extraire la grammaire et les probabilités à partir du corpus d'entraînement. Le lexique est par contre toujours extrait du corpus complet : nous n'avons pas cherché ici à étudier le comportement du modèle sur les mots inconnus.
- évaluer les performances sur le corpus de test ; la mesure utilisée est le nombre d'arbres d'analyse complets corrects obtenus.

Pour chacun des deux corpus ATIS et « Susanne réduit » décrits précédemment, les résultats fournis sont des moyennes sur au moins 10 partionnements aléatoires indépendants.

Pour disposer d'une référence, nous avons également extrait et évalué la grammaire hors-contexte probabiliste standard. Il ne nous a par contre pas encore été possible d'effectuer des expériences sur le modèle DOP complet en raison du nombre gigantesque d'arbres élémentaires générés dans les conditions expérimentales choisies (arbres n-aires et prise en compte des mots).

Les résultats obtenus sont résumés dans la table 3. Pour chaque modèle, « *% parsed* » représente la couverture du modèle, c.-à-d. le pourcentage de phrases du corpus de test qui ont obtenu au moins une analyse[16], « *% correct on parsed* » indique sa précision, c.-à-d. le pourcentage de

[15] Seules les productions vides (« traces ») et quelques erreurs manifestes ont été supprimées.

[16] Notez bien que ce nombre est par construction toujours le même pour le modèle hors-contexte et pour le

		Susanne réduit			ATIS		
		% parsed	% correct on parsed	% correct overall	% parsed	% correct on parsed	% correct overall
test (10%)	hors-contexte	45.5	23.0	10.5	99.6	25.4	25.3
	min-max DOP	45.5	24.4	11.1	99.6	21.0	20.9
entr.	hors-contexte	100	61.2	61.2	100	33.8	33.8
	min-max DOP	100	87.9	87.9	100	76.1	76.1

Table 3: Résultats expérimentaux obtenus dans les conditions de test (séparation aléatoire du corpus en 90%–10%) et sur le corpus d'entraînement complet (100%).

phrases analysées pour lesquelles l'analyse la plus probable est correcte, et « *% correct overall* » indique la performance globale du modèle, c.-à-d. le pourcentage de phrases (parmi toutes celles du corpus de test) pour lesquelles l'analyse la plus probable est correcte.

Les conclusions générales que nous pouvons tirer de ces résultats sont les suivantes :

1. La restriction minimale-maximale du modèle DOP fournit de meilleurs résultats que les grammaires hors-contexte probabilistes standard sur le corpus Susanne réduit. La différence en performance globale observée est effectivement statistiquement significative à un niveau de confiance de 95%.
2. La performance et la précision sur le corpus ATIS corpus sont assez mauvaises pour les deux modèles : cette mauvaise performance est liée à la nature trop bruitée de ce corpus[17].
3. La faible couverture obtenue sur le corpus Susanne réduit est liée au très fort taux (77 %) de règles hors-contexte n'apparaissant qu'une fois dans tout le corpus (hapax).

6 Conclusion

Cette contribution présente une nouvelle approche du Data-Oriented Parsing : sa restriction à des grammaires à substitution d'arbres polynomiales (PSTSG), c.-à-d. à des STSG pour lesquelles la recherche de l'arbre d'analyse le plus probable peut s'effectuer de façon exacte en un temps polynomial (par opposition au caractère NP-difficile de cette recherche dans le cas général).

Le cas particulier de la restriction des arbres élémentaires aux seuls arbres de profondeur 1 et aux arbres totalement ancrés (principe de sélection minimale-maximale) a été analysé. On a pu dans ce cas exhiber une grammaire hors-contexte probabiliste équivalente au modèle DOP considéré, c.-à-d. une grammaire hors-contexte pour laquelle la recherche de la dérivation la plus probable est équivalente à la recherche de l'analyse la plus probable au sein du modèle DOP considéré. Cette construction est rendue possible grâce au fait qu'avec la restriction appliquée, il existe toujours au moins une dérivation pouvant porter la totalité de la DOP-probabilité de l'arbre d'analyse correspondant.

Le modèle présenté constitue donc un compromis intéressant entre le modèle DOP général et les grammaires hors-contexte : il est aussi « simple » à analyser qu'une grammaire hors-contexte

modèle DOP

[17]Ce qui a déjà été évoqué dans la littérature : voir par exemple (Goodman, 1998) p. 179.

(complexité cubique) tout en permettant une probabilisation plus riche que celle des modèles hors-contexte usuels.

Une question encore ouverte sur laquelle nous travaillons actuellement est de savoir s'il existe d'autres PSTSG, c.-à-d. d'autres restrictions de DOP linguistiquement intéressantes, associées à des mécanismes de sélection des arbres élémentaires différents de celui présenté ici, mais possédant les mêmes propriétés d'efficacité (analyse polynomiale) et de facilité de construction.

Une comparaison avec d'autres extensions des SCFGs, comme par exemple les « Stochastic Lexicalized CFG » (Schabes & Waters, 1993), est également envisagée.

Références

BOD R. (1992). Applying Monte Carlo techniques to Data Oriented Parsing. In *Proceedings Computational Linguistics in the Netherlands*, Tilburg (The Netherlands).

BOD R. (1998). *Beyond Grammar, An Experience-Based Theory of Language*. Number 88 in CSLI Lecture Notes. Standford (CA): CSLI Publications.

BOOTH T. L. & THOMPSON R. A. (1973). Applying probability measures to abstract languages. *IEEE Transactions on Computers*, **C-22**(5), 442–450.

CHAPPELIER J.-C. & RAJMAN M. (1998). Extraction stochastique d'arbres d'analyse pour le modèle DOP. In *Proc. of 5ème conférence sur le Traitement Automatique du Langage Naturel (TALN98)*, p. 52–61, Paris (France).

CHAPPELIER J.-C. & RAJMAN M. (2000). Monte-Carlo sampling for NP-hard maximization problems in the framework of weighted parsing. In D. CHRISTODOULAKIS, Ed., *Natural Language Processing – NLP 2000*, number 1835 in Lecture Notes in Artificial Intelligence, p. 106–117. Springer.

CHAPPELIER J.-C. & RAJMAN M. (2001). *Polynomial Tree Substitution Grammars: an efficient framework for Data-Oriented Parsing*. Rapport interne 01/tocome, Département Informatique, EPFL, Lausanne (Switzerland).

GOODMAN J. (1996). Efficient algorithms for parsing the DOP model. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, p. 143–152.

GOODMAN J. (1998). *Parsing Inside-Out*. PhD thesis, Harvard University. cmp-lg/9805007.

HEMPHILL C. T., GODFREY J. J. & DODDINGTON G. R. (1990). The ATIS spoken language systems pilot corpus. In M. KAUFMANN, Ed., *DARPA Speech and Natural Language Workshop*.

SAMPSON G. (1994). The Susanne corpus, release 3. In *School of Cognitive & Computing Sciences*, Brighton (England): University of Sussex Falmer.

SCHA R. (1990). Language theory and language technology: competence and performance. In DE KORT & LEERDAM, Eds., *Computertoepassingen in de Neerlandistiek*. Almere (The Netherlands): LVVN-jaarboek. in Dutch.

SCHABES Y. & WATERS R. C. (1993). *Stochastic Lexicalized Context-Free Grammars*. Rapport interne 93–12, Mitshbishi Electric Research Labs.

SIMA'AN K. (1996a). Computational complexity of probabilistic disambiguation by means of tree grammars. In *Proceedings of COLING'96*, Copenhagen (Denmark). cmp-lg/9606019.

SIMA'AN K. (1996b). Efficient disambiguation by means of stochastic tree substitution grammars. In R. MITKOV & N. NICOLOV, Eds., *Recent Advances in NLP*, volume 136 of *Current Issues in Linguistic Theory*. Amsterdam (The Netherlands): Benjamins.