« Simple comme EASy :-) »

Pierre Boullier, Lionel Clément, Benoît Sagot, Éric Villemonte de La Clergerie
INRIA - Projet Atoll
Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay (France)
{Benoit.Sagot,Eric.De_La_Clergerie}@inria.fr
Lionel.Clement@lefff.net

Mots-clefs : Analyse syntaxique, évaluation

Keywords: Parsing, Evaluation

Résumé Cet article présente les deux systèmes d'analyse déployés par le projet ATOLL (INRIA) lors de la campagne EASy d'Évaluation d'Analyseurs Syntaxiques (décembre 2004). Nous donnons quelques résultats quantitatifs en termes de couverture et de temps d'analyse. Cette expérience permettra la comparaison à grande échelle du résultat de différents analyseurs, mais montre aussi que les techniques d'analyse syntaxique profonde sont désormais à même de traiter des corpus volumineux tout en conservant leur puissance d'expression linguistique.

Abstract This paper presents both parsing systems used by project ATOLL (INRIA) for the EASy parsing evaluation campaign (December, 2004). We give a few quantitative results in terms of coverage and parsing time. These experiments will allow the comparison of parsing results on huge corpus, but also show that deep parsing techniques can cope with large corpus while preserving their linguistic expressive power.

1 Introduction

Pour la campagne EASy d'Évaluation des Analyseurs Syntaxiques, l'équipe ATOLL de l'INRIA a déployé deux chaînes de traitement syntaxiques (Boullier *et al.*, 2005a). L'analyse d'environ 35000 phrases fournies par les organisateurs nous permet de présenter ici quelques résultats préliminaires en attendant les résultats définitifs.

Les phrases se répartissaient en un ensemble de corpus non retravaillés couvrant divers styles de langage et présentant toutes sortes de problèmes à régler, bien en amont de l'analyse syntaxique proprement dite. En particulier, la segmentation en phrases et en tokens n'était pas toujours justifiée linguistiquement, et par conséquent pas toujours compatible avec nos outils. Il a fallu adapter nos programmes pour leur permettre de re-segmenter les corpus tout en préservant au mieux la segmentation fournie qui devait être celle des résultats.

En sortie d'analyse syntaxique s'est posé le problème de convertir nos sorties au format attendu par les organisateurs (Gendner & Vilnat, 2004), donnant des informations sur des constituants

simples non récursifs et sur un jeu de dépendances. Malgré le fait que nos analyseurs produisent des analyses ambiguës, nous avons essayé de rendre des résultats non-ambigus sur les constituants et les dépendances en nous appuyant sur quelques heuristiques.

2 Les systèmes FRMG et SxLFG

Nos deux systèmes s'appuient sur deux formalismes syntaxiques « profonds » différents :
- le système FRMG[1] (Thomasset & de la Clergerie, 2005) repose sur une grammaire TAG compacte constituée de quasi-arbres sous-spécifiés générés automatiquement à partir d'une méta-grammaire ; cette grammaire est compilée par le constructeur d'analyseurs syntaxiques DyALog pour produire un analyseur,
- le système SxLFG (Boullier *et al.*, 2005b) repose sur une grammaire LFG qui est une évolution de celle citée par (Clément & Kinyon, 2001) ; cette grammaire est utilisée par le générateur SxLFG d'analyseurs LFG qui est lui-même fondé sur l'environnement SYNTAX de génération d'analyseurs non contextuels.

Les deux systèmes utilisent le même lexique *Lefff* (Sagot *et al.*, 2005) et la même chaîne SxPipe de traitement pré-syntaxique (Sagot & Boullier, 2005), ces deux composants étant d'une importance considérable. En particulier, SxPipe s'est révélé essentiel pour traiter au mieux les corpus fournis, avec leurs fautes, bizarreries typographique et artefacts divers. Notons que SxPipe produit une sortie ambiguë sous forme de treillis de mots.

Enfin, chaque système a dû mettre en place des procédures distinctes de désambiguïsation et de conversion des sorties des analyseurs vers le format demandé par le Guide d'annotation EASy.

3 Résultats et comparaisons préliminaires

Les résultats en temps d'analyse et en couverture[2] pour nos deux systèmes sont détaillés dans les tableaux 1, 2 et 3. Les taux de couverture pour des analyses complètes[3] sont comparables entre les systèmes mais varient significativement selon les types de corpus. Une analyse fine des temps d'analyse permet de montrer que SxLFG est beaucoup plus rapide pour les phrases courtes (moins de 15 tokens environ)[4], mais que la polynomialité du formalisme TAG permet à FRMG d'être plus efficace pour les longues phrases. Ces conclusions sont toutefois à nuancer par le fait que les deux systèmes d'analyse sont très récents, et continuent à être améliorés dans des proportions importantes, permettant d'envisager une forte diminution des temps d'analyse.

Puisque nous disposons de deux jeux de résultats au format EASy, nous avons mené quelques expériences préliminaires pour les comparer, en se focalisant sur les constituants (les dépen-

[1] utilisable en ligne sur `http://atoll.inria.fr/parserdemo`.

[2] Pour la définition précise de ce que signifient les expressions *couverture de la CFG support*, *couverture avec/sans vérification de cohérence*, le lecteur est invité à se reporter à (Boullier *et al.*, 2005a).

[3] Les 2 systèmes sont robustes et fournissent également des résultats partiels quand une analyse complète n'est pas trouvée.

[4] On notera aussi l'extrême efficacité de la partie CFG de l'analyse effectuée par SxLFG. Cette partie, qui repose sur le système SYNTAX, permet par exemple de trouver en moins de 6 secondes les 5.10^{52} analyses CFG d'une des phrases du corpus de courrier électronique, et ne met que 2 minutes environ à construire une forêt partagée représentant 3.10^{73} analyses CFG après rattrapage d'erreur pour une autre phrase du même corpus qui n'a pas d'analyse CFG correcte.

« Simple comme EASy :-) »

Corpus	#phrases	% couv.	temps d'analyse					amb.
			moy.	méd.	$\geq$ 1s	$\geq$ 10s	Timeout	
general	6160	41.01%	10.31s	2.44s	71.19%	18.01%	5.27%	0.65
litteraire	7960	38.93%	5.59s	2.10s	71.75%	9.64%	1.80%	0.53
mail	7962	32.83%	4.37s	1.46s	59.65%	7.08%	1.27%	0.70
medical	2225	44.00%	5.47s	1.46s	63.26%	8.21%	2.40%	0.70
oral	6892	44.39%	5.58s	1.19s	54.58%	6.39%	0.78%	0.53
questions	3509	66.28%	3.47s	1.32s	66.04%	4.53%	1.08%	0.58
Total	**34438**	**42.45%**	**5.55s**	**1.61s**	**64.41%**	**9.32%**	**2.07%**	**0.60**

TAB. 1 – Résultats pour FRMG (*time-out*=100s)

corpus	#phrases	couv. de la CFG support	couverture sans vérif. de cohérence	couverture avec vérif. de cohérence	temps médian	*timeout* (t $\geq$ 15s)
general	6952	89.24%	57.03%	32.42%	0.54s	22.64%
litteraire	11408	89.92%	69.07%	40.52%	0.07s	13.03%
mail	9308	83.02%	66.08%	40.18%	0.01s	9.53%
medical	2553	87.34%	60.95%	39.99%	0.06s	12.10%
oral	7075	85.24%	67.94%	46.47%	0.01s	8.30%
questions	3563	92.73%	80.33%	62.19%	0.01s	5.22%
Total	**40859**	**87.51%**	**66.62%**	**41.95%**	**0.03s**	**12.31%**

TAB. 2 – Résultats pour SXLFG (*time-out*=15s)

dances étant plus délicates à comparer). Nos deux systèmes produisent exactement la même analyse en constituants pour 7714 phrases. La phrase la plus longue sur laquelle nous soyons d'accord a 42 tokens EASy, la moyenne étant d'environ 7 tokens. Le tableau 4 indique les distributions des différents constituants pour chaque système et celle des constituants communs aux deux (même type et même couverture de tokens).

4 Conclusion

Naturellement, cette brève présentation n'est pas complète car il manque les informations sur le taux de précision des analyses fournies. Nous avons déjà repéré de nombreux problèmes, certains liés à la segmentation et à la difficulté des corpus, de nombreux autres liés à la couverture des grammaires et aux processus de désambiguïsation et de conversion au format EASy. Néanmoins, nous tirons déjà trois principaux enseignements de cette campagne :

		Corpus complet	Phrases valides pour la CFG support	
		Analyse CFG	Analyse CFG	Analyse complète
Corpus	#phrases	40859	35756	
	n_{moy} - n_{max}	20.95 - 541	19.06 - 173	
	UW_{moy} - UW_{max}	0.79 - 97	0.75 - 65	
Temps d'analyse	med	0.00s	0.00s	0.03s
	$\geq$ 0.1s	1.79%	1.20%	42.2%
	$\geq$ 1s	0.24%	0.09%	29.0%
Nombre d'analyses	med - max	32 028 - 3.10^{73}	29 582 - 5.10^{52}	1 - 1
	$\geq 10^6$	36.13%	35.28%	0%
	$\geq 10^{12}$	8.86%	7.84%	0%

TAB. 3 – Détail des temps d'analyse pour SXLFG (*time-out*=15s)

Groupes	GA	GN	GP	GR	NV	PV
FRMG	27843	96880	71332	25833	90166	6953
SXLFG	28128	118321	61392	34229	85564	6823
communs	16182	56281	42656	15728	57460	1061

TAB. 4 – Comparaison par types de constituants entre FRMG et SXLFG

– elle n'évalue pas uniquement la phase d'analyse syntaxique, mais également le lexique, la chaîne de traitement pré-syntaxique, et la phase d'extraction des annotations EASy à partir des sorties des analyseurs ;
– elle permettra des travaux intéressants de comparaison des résultats d'analyseurs différents, notamment pour la constitution de corpus annotés mais aussi pour la détection d'erreurs et de manques dans les grammaires, les lexiques et la chaîne de traitement pré-syntaxique ;
– elle montre que les technologies d'analyse syntaxique profonde retenues par ATOLL sont suffisamment efficaces pour permettre le traitement de gros corpus.

Les deux derniers points nous confortent dans nos choix de recherche. En effet, la plupart des traitements linguistiques actuels sur gros corpus reposent sur des technologies de surface (probabilités, chunks, automates finis) choisies pour leur efficacité algorithmique, mais ne permettant pas une modélisation linguistiquement satisfaisante des mécanismes de la langue. À l'inverse, les technologies d'analyse profonde permettent une telle modélisation, seule à même de permettre à moyen terme d'effectuer avec un haut degré de précision des tâches linguistiquement complexes telles que la traduction automatique. Nos expériences montrent que leurs points faibles traditionnels sont de moins en moins pertinents : les analyseurs vont de plus en plus vite et les grammaires peuvent être développés de plus en plus rapidement, en particulier grâce aux développements récents autour du concept de méta-grammaires. De plus, le développement des lexiques riches en information qui sont nécessaires peut être facilité par l'examen statistique des sorties d'analyses profondes de corpus suffisamment importants.

Références

BOULLIER P., CLÉMENT L., SAGOT B. & VILLEMONTE DE LA CLERGERIE E. (2005a). Chaînes de traitement syntaxique. In *Proceedings of TALN'05*, Dourdan, France.

BOULLIER P., SAGOT B. & CLÉMENT L. (2005b). Un analyseur LFG efficace : SXLFG. In *Actes de TALN'05*, Dourdan, France.

CLÉMENT L. & KINYON A. (2001). XLFG-an LFG parsing scheme for French. In *Proc. of LFG'01*.

GENDNER V. & VILNAT A. (2004). Les annotations syntaxiques de référence PEAS. En ligne sur `www.limsi.fr/Recherche/CORVAL/easy/PEAS_reference_annotations_v1.6.html`.

SAGOT B. & BOULLIER P. (2005). From raw corpus to word lattices : robust pre-parsing processing. In *Actes de L&TC 2005*, Poznań, Pologne.

SAGOT B., CLÉMENT L., ÉRIC VILLEMONTE DE LA CLERGERIE & BOULLIER P. (2005). Vers un méta-lexique pour le français : architecture, acquisition, utilisation. In *Journée ATALA sur l'interface lexique-grammaire*. `http ://www.atala.org/article.php3?id_article=240`.

THOMASSET F. & DE LA CLERGERIE E. V. (2005). Comment obtenir plus des méta-grammaires. In *Actes de TALN'05*, Dourdan, France.