L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASY

Romaric Besançon et Gaël de Chalendar
CEA/LIST/LIC2M
BP 6 92265 Fontenay-aux-Roses Cedex France
{Romaric.Besancon,Gael.de-Chalendar}@cea.fr

Mots-clefs : analyse syntaxique, campagne d'évaluation, EASY, grammaires de dépendances, automates

Keywords: syntactic analysis, evaluation campaign, EASY, dependency grammars, automatons

Résumé Le LIC2M, laboratoire du CEA/LIST, a participé à la campagne d'évaluation EASY avec l'analyseur syntaxique de son système LIMA, un analyseur syntaxique robuste qui implémente une grammaire de dépendance. Les résultats obtenus sur le corpus d'exemples sont encourageants et permettent de valider les techniques utilisées. En revanche, le traitement de corpus plus généraux couvrant des phénomènes syntaxiques plus variés nécessiteront sûrement le développement de ressources supplémentaires ou la mise en place de traitements particuliers.

Abstract The LIC2M, a CEA/LIST laboratory, participated in the EASY campaign to test the syntactic parser of the NLP system LIMA, a robust syntactic parser that implements a dependency grammar. The development of this parser is an on-going work, but the results on the test set are promising. Nevertheless, the parsing of more general corpora, containing more varied syntactic phenomena should require additional work on the development of resources.

1 Introduction

Le LIC2M, laboratoire du CEA/LIST, développe depuis trois ans un ensemble d'outils de traitement automatique des langues en vue de leur utilisation dans diverses applications (recherche d'information, question-réponse, résumé automatique, etc.). L'ensemble de ces outils forme le système LIMA[1] (pour "LIc2m Multilingual Analyzer"), un système d'analyse linguistique avancée pensé dans une optique multilingue. Le LIC2M a participé à la campagne d'évaluation EASY (EASY, 2004) pour valider le module d'analyse syntaxique de LIMA, qui était en cours de développement lors du lancement de la campagne. Nous présentons dans la section 2 le module d'analyse syntaxique de l'analyseur multilingue LIMA et les adaptations nécessaires pour la participation à EASY. Puis nous présentons dans la section 3 une évaluation du système sur le corpus d'exemples annoté de la campagne EASY.

[1] LIMA est un travail collégial du LIC2M. En dehors des auteurs du présent article, ont participé à sa réalisation: O. Ferret, C. Fluhr, G. Grefenstette, M. Laib-Boukari, Y. Li, B. Mathieu, O. Mesnard, H. Naets et N. Semmar.

2 L'analyse syntaxique du système LIMA

Le système LIMA a été réalisé dans la lignée des travaux de Christian Fluhr et ses collègues au CEA (Fluhr et al., 1997), et reprend, en les enrichissant, les principes proposés à cette époque : dictionnaires *full-form*, catégories morphosyntaxiques positionnelles, dictionnaires bilingues etc. L'analyse linguistique est réalisée par une chaîne configurable de modules indépendants appliqués successivement sur un texte : segmentation, traitement des expressions figées, analyse morphologique, désambiguïsation syntaxique, reconnaissance des entités nommées, analyse syntaxique, création de termes composés. Actuellement, l'analyseur LIMA fonctionne sur six langues: allemand, anglais, arabe, chinois, espagnol et français. Son extension est en cours pour l'italien et le russe.

L'analyseur syntaxique de LIMA implémente une grammaire de dépendance (Kahane, 2000) en ce sens que les analyses produites sont exclusivement représentées par des relations de dépendance entre deux mots, un recteur et un régi. Nous nous inscrivons aussi dans le cadre des analyseurs robustes (Aït-Mokthar and Chanod, 1997; Grefenstette, 1998) puisque l'analyse est effectuée à l'aide d'automates à états finis dont la stratégie d'application permet d'obtenir une analyse pour toute phrase, même agrammaticale, du moment qu'un sous-ensemble de la phrase est reconnu par un des automates de la grammaire. L'analyse est séparée en deux étapes : la recherche des chaînes nominales et verbales et la recherche des relations de dépendance.

2.1 Chaînes nominales et verbales

Les chaînes nominales et verbales ne représentent pas un objet linguistique standard. En effet, il s'agit surtout d'une aide pour la recherche ultérieure des relations de dépendance et la désambiguisation de l'analyse syntaxique. Les chaînes nominales et verbales peuvent être décrites comme des syntagmes maximaux reliant entre eux l'ensemble des syntagmes minimaux non-récursifs (tels que définis dans le cadre de EASY (EASY, 2004)) susceptibles d'être liés, respectivement, à un même nom ou un même verbe.

Pour l'idenfitication de ces chaînes, une matrice définissant les successions autorisées de catégories et de mots est utilisée ainsi que des listes de catégories et de mots pouvant débuter ou terminer une chaîne, sachant qu'une même catégorie ne peut pas débuter à la fois une chaîne verbale et une chaîne nominale. Toutes les configurations possibles de chaînes sont cherchées pour chaque phrase.

2.2 Recherche des relations de dépendance

La recherche des relations de dépendance utilise des ensembles de règles représentant des grammaires locales et implémentées sous formes d'automates à états finis. Un phénonème syntaxique particulier peut être traité par une ou plusieurs règles (par exemple, la relation adverbe-adjectif est traitée par une seule règle alors que la relation sujet-verbe nécessite 12 règles).

Les règles sont définies par un élément *déclencheur* (qui peut être un mot ou une catégorie morphosyntaxique), et ses *contextes* gauche et droit, représentés par des expressions régulières. Des *contraintes* peuvent être spécifiées sur le déclencheur et/ou les éléments dans ses contextes gauche et droit, permettant de vérifier par exemples des accords en genre et en nombre, ou la

présence de relations existantes entre deux éléments. Des *actions* sont enfin attachées à chaque règle, par exemple pour créer des relations de dépendance supplémentaires. Pour chaque phrase, chaque mot est testé pour savoir s'il est déclencheur d'une règle : si c'est le cas, ses contextes droit et gauche sont testés. Si la règle s'applique, les actions sont effectuées.

Les règles sont regroupées en plusieurs ensembles qui sont appliqués successivement, pour permettre de traiter incrémentalement les relations cherchées en fonction de leur priorité et parce que certaines règles doivent s'appliquer avant d'autres alors que leurs déclencheurs sont situés plus avant dans la phrase. En particulier, les règles des relations homosyntagmatiques sont traitées avant les règles des relations hétérosyntagmatiques : les relations homosyntagmatiques sont les relations internes à une chaîne, donc les relations locales dans les syntagmes non interrompus par des incises, des parenthèses, etc. Elles comptent les relations entre les noms et les adjectifs, celles entre les verbes et les auxiliaires, mais aussi les relations entre les noms ou verbes et les mots dits grammaticaux comme les articles ou les adverbes. Les relations homosyntagmatiques sont traitées pour le français par trois ensembles de règles qui représentent un total de 56 règles. Les relations hétérosyntagmatiques sont les relations entre syntagmes et à longue distance, comme les relations entre le verbe et ses actants (sujet, compléments divers) ou les relations de conjonction ou de subordination. Ces relations sont recherchées après les relations homosyntagmatiques à l'aide d'un seul ensemble contenant 62 règles.

2.3 Adaptations pour EASY

Le corpus EASY étant déjà segmenté, les modules de LIMA concernant la segmentation (avec traitement des mots à tirets et des expressions figées) ont été remplacés par un simple découpage sur les espaces. La liste des formes composées fournie par les organisateurs, avec leurs catégories morphosyntaxiques a été intégrée dans le dictionnaire.

Concernant l'analyse syntaxique, l'analyse en dépendance produite par le système LIMA est fondamentalement équivalente à une analyse en constituants et dépendances. Pour trouver un constituant tel que le GN défini dans EASY, il suffit, à partir d'un noeud donné, par exemple un article, de suivre récursivement certains types de relations (ici, entre autres, les relations déterminant-substantif et adjectif prénominal-substantif) et de collecter les noeuds. L'ensemble de noeuds collectés forme le GN. En ordonnant les tests sur les noeuds et les ensembles de relations à suivre, on parvient à obtenir des syntagmes cohérents avec ceux définis dans le guide d'annotation EASY. Pour ce qui est des relations définies dans EASY, à un renommage près, il s'agit des relations extraites par notre système qui ne sont pas utilisées dans la construction des groupes. Leur extraction ne présente donc pas de difficulté particulière.

3 Evaluation

Les résultats officiels de la campagne EASY n'étant pas disponibles lors de l'écriture de cet article, l'évaluation proposée dans cette section a été faite sur le corpus d'exemples annoté fourni lors de la campagne. Les résultats obtenus sur ce corpus (en précision et rappel, pour chaque type de groupe et de relation) sont présentés dans la table 1. Ces résultats sont assez bons, mais il faut noter que ce corpus est essentiellement un corpus d'illustration des phénomènes syntaxiques de la langue française et n'est pas représentatif des phrases rencontrées dans un corpus réel. De plus, ce corpus ayant servi de corpus de référence lors de l'écriture des règles, ces

groupes	prec	rappel	*relations*	prec	rappel		prec	rappel
GA	81.2	97.5	ATB-SO	100.0	30.2	MOD-A	60.0	13.0
GN	85.4	95.0	AUX-V	94.4	87.2	MOD-N	61.4	50.0
GP	85.9	93.4	COD-V	60.2	60.2	MOD-R	87.5	87.5
GR	83.9	100.0	COMP	72.7	34.8	MOD-V	86.7	35.1
NV	96.5	100.0	COORD	71.4	20.8	SUJ-V	93.0	81.2
PV	81.8	90.0	CPL-V	52.9	42.2			
TOT	88.9	90.6				TOT	75.8	54.9

Table 1: Résultats obtenus sur le corpus d'entraînement EASY

résultats sont biaisés, et les résultats attendus sur les corpus de test plus généraux seront certainement moins bons. En particulier, il est difficile d'écrire des règles génériques (qui restent par essence assez locales) pour capter des relations lointaines dans des phrases complexes.

4 Conclusion et perspectives

L'analyseur syntaxique robuste du système LIMA donne des résultats encourageants sur le corpus d'exemples annoté fourni dans la campagne EASY. Néanmoins, les résultats attendus sur les corpus de test seront sans doute moins bons. Il est en effet difficile de développer manuellement (mais aussi d'apprendre automatiquement) des ensembles de règles complets et cohérents permettant d'analyser correctement du texte tout venant, pouvant contenir des structures arbitrairement complexes (relatives, incises, etc). Le type d'analyse que nous effectuons fonctionne bien sur des phrases simples. Par conséquent, nous développons actuellement des algorithmes et des ressources permettant de détecter les éléments complexes dans une phrase, de les supprimer temporairement, de les remettre en place après analyse de la phrase simple obtenue et enfin de les rattacher aux restes de l'analyse, ce qui accentuera l'aspect incrémental de l'analyse.

Par ailleurs, une des utilisations de l'analyse syntaxique dans LIMA est la construction de termes composés (utilisés pour la recherche d'information). Les utilisateurs du moteur de recherche se disent très satisfaits de ces mots composés mais une évaluation plus quantitative reste à faire. Une partie de cette évaluation pourra être faite dans le cadre de la campagne CESART, cousine de EASY dédiée à l'extraction de ressources terminologiques.

Références

Aït-Mokthar, S. and Chanod, J.-P. (1997). Incremental finite-state parsing. In *Proceedings of the 8th conference on Applied Natural Language Processing ANLP-97*, pages 72–79, Washington.

EASY (2004). Campagne d'évaluation des analyseurs syntaxiques. http://www.technolangue.net/article64.html http://www.elda.org/easy http://www.limsi.fr/corval/easy.

Fluhr, C., Schmit, D., Ortet, P., Elkateb, F., and Gurtner, K. (1997). Spirit-w3, a distributed crosslingual indexing and retrieval engine. In *INET'97*.

Grefenstette, G. (1998). Light parsing as finite-state filtering. In Kornai, A., editor, *Extended Finite State Models of Language*. Cambridge University Press.

Kahane, S., editor (2000). *Les grammaires de dépendance*, volume 41 of *Traitement automatique des langues*. Hermès.