Laboratoire d'Informatique de Nantes-Atlantique

UMR 6241

École doctorale STIM, n. 503

« Sciences et technologies

de l'information et des mathématiques »

Sujet de thèse pour 2010

Analyseur syntaxique probabiliste contextuel basé sur le modèle des grammaires catégorielles de dépendance

Résumé. L'objectif de cette thèse est d'élaborer un analyseur probabiliste efficace basé sur le modèle les Grammaires Catégorielles de Dépendances à large couverture. Le travail portera, d'une part, sur la recherche d'un modèle d'analyseur probabiliste générique où la notion de dépendances discontinues sera pleinement exploitée, et, d'autre part, sur la création d'un modèle spécifique pour une ou plusieurs langues naturelles (français, russe, anglais, etc.).

Mots clés. Analyse syntaxique, analyse en dépendance, analyseur probabiliste, entraînement, traitement des ellipses.

Ph. D. Proposal for 2010

Context Dependent Stochastic Parser for Categorial Dependency Grammars

Abstract. The objective of this PhD thesis is to develop and to implement an efficent stochastic parser for large scale Categorial Dependency Grammars. The work will consist on the one hand, of the research into a model of such a parser treating the discontinuous dependency, and on the other hand, of implementation of models specific to one of particular languages (French, Russian, English, etc).

Keywords. Parsing, Dependency parsing, Stochastic parser, Learning, Ellipsis treatment.

Directeur de thèse ▷ Nom, Prénom : Dikovsky, Alexandre
Équipe d'accueil : T.A.L.N.
Courriel : alexandre.dikovsky (at) univ-nantes.fr
Téléphone : (+33/0) 2 51 12 58 43

Co-directeur de thèse ▷ Nom, Prénom : Béchet, Denis
Équipe d'accueil : T.A.L.N.
Courriel : Denis.Bechet (at) univ-nantes.fr
Téléphone : (+33/0) 2 51 12 58 52

Introduction

Contexte et problématique

L'apparition d'infrastructures de traitement automatique du langage naturel nécessite des outils de bonne qualité, robustes et efficaces. Pour aller au delà d'une analyse qui resterait en surface, les chaînes de traitement ont besoin d'analyseurs syntaxiques et de grammaires dont les dérivations facilitent l'extraction automatique des structures sémantiques. Dans ce cadre, les structures de dépendance sont très proches d'une forme logique des textes car elles représentent sous une forme explicite les relations entre les têtes nominales et verbales et leurs adjoints subordonnés. La formalisation de grammaires de dépendances et leur utilisation dans des analyseurs syntaxiques qui soient à la fois robuste, rapide et à large couverture sont des objectifs importants dans le cadre de l'accès à l'information dans les textes écrits.

Problèmes et opportunités

Le problème principal des grammaires de dépendances est celui de la définition des dépendances discontinues (dites non-projectives), c'est-à-dire les dépendances entre le mot-tête et le mot subordonné séparés par des mots ne dépendant pas de la tête. Ces dépendances sont dues aux constructions discontinues (telles la négation "ne .. pas", les comparatifs "plus de .. que", etc.) mais aussi à la topicalisation des groupes nominaux qui se manifeste par leur déplacement au début de la phrase. La majorité des grammaires de dépendances ne traitent pas les dépendances non-projectives (cf. [Gai65, Hud84, ST93]). Celles qui les traitent ne sont pas analysables en temps polynomial (cf. [Bla01, DD01]). L'équipe TALN s'intéresse depuis plusieurs années à la notion de grammaires catégorielles de dépendance (GCD) en particulier sur leur formalisation [Dik04, BDF05, DD08]. Les GCD est la seule classe de grammaires de dépendances dans la littérature qui définissent les dépendances non-projectives et, en même temps, sont analysables en temps polynomial [DD08]. Au sein de l'équipe TALN ont été élaborés une GCD à large couverture du français et un analyseur pour ces GCD [Dik09] qui continue d'être amélioré, en particulier, en ce qui concerne la robustesse et la précision. Le problème général de l'analyse syntaxique des grammaires à large couverture (en dépendances ou en constituants) et l'explosion combinatoire des analyses fallacieuses. L'analyseur des GCD donnant actuellement toutes les solutions compatibles avec une GCD, il est souvent nécessaire de trier les solutions suivant leur pertinence.

Travail demandé

Objectifs

L'objectif de cette thèse est d'élaborer un analyseur probabiliste basé sur le modèle des GCD. Le travail portera, d'une part, sur la recherche d'un modèle d'analyseur probabiliste générique où la notion de dépendances discontinues sera pleinement exploitée, et, d'autre part, sur la création d'un modèle spécifique pour une ou plusieurs langues naturelles (français, russe, anglais, etc.) permettant d'atteindre un taux très élevé de précision. L'entraînement de cet analyseur syntaxique pour les grammaires choisies se fera sur les corpus dont disposent l'équipe TALN ou bien sur ceux qu'elle développe actuellement. La mise en oeuvre et l'expérimentation des algorithmes permettra une comparaison avec les résultats obtenus avec les autres méthodes du domaine (voir la littérature citée). La manière totalement lexicalisée de représenter les dépendances discontinues propre aux GCD sera un des atouts de ce travail pour l'analyse des structures discontinues. Il est devenu évident qu'un modèle ne prenant pas en compte la totalité des phrases d'un texte ne peut pas dépasser un taux de précision 90 %. Pour aller au delà de cette limite actuelle de la précision des analyseurs, un enjeu important sera la prise en compte du contexte des phrases à analyser dans l'évaluation des probabilités associées aux différentes analyses possibles d'un paragraphe. Une application particulière de cette notion de contexte concerne la résolution des ellipses dans les textes qui nécessitent, pour une bonne résolution, d'avoir une notion approximative des propositions et des phrases précédentes.

Plan de travail prévisionnel de l'étude

Dans un premier temps, il faudra étudier les modèles d'analyse stochastique actuels. Ensuite, il faudra trouver de nouveaux modèles qui prennent en compte la spécificité des dépendances discontinues des GCD. Le travail portera à la fois sur la définition d'un nouvel analyseur stochastique et sur une modification de l'analyseur complet déterministe actuel avec une fonction de classification probabiliste des analyses entraînée sur des corpus arborés. Les analyseurs seront testés et comparés avec les meilleur analyseurs en dépendances.

Candidats

Compétences

La base de la thèse se trouve dans le domaine du Traitement Automatique du Langage Naturel. Le candidat devra en particulier connaître la notion de grammaire appliquée au langage naturel et des méthodes classiques d'analyse syntaxique déterministe et stochastique. Une première expérience de l'apprentissage automatique serait appréciée. Les

réalisations nécessitent une certaine maitrise des langages informatiques tels que Common Lisp, Java, C++ et pour les interfaces utilisateur, PHP, Javascript, Perl et Bash.

Bibliography

[A. 08] A. Cahill et al. Wide-coverage deep statistical parsing using automatic dependency structure annotation. *Comput. Ling.*, 34(1):81–124, 2008.

[Abe03] A. Abeillé. *Building a treebank for French in Treebanks*, pages 165–188. Kluwer, Dordrecht, 2003.

[BDF05] Denis Béchet, Alexander Dikovsky, and Annie Foret. Dependency structure grammar. In *Proceedings of the 5th International Conference on Logical Aspects of Computational Linguistics, Bordeaux, France, April 2005*, volume 3492 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 18–34. Springer-Verlag, 2005.

[Bla01] P. Blache. *Les Grammaires de Propriétés : Des contraintes pour le traitement automatique des langues naturelles*. Hermès, 2001.

[Cha00] E. Charniak. A maximum entropy inspired parser. In *Proc. of the First Annual Meeting of the North American Chapter of the ACL (NAACL 2000), Seattle, WA.*, pages 132–139, 2000.

[Col99] M. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, Penn State Univ., Philadelphia, PA, 1999.

[DD01] D. Duchier and R. Debusmann. Topological dependency trees: A constraint-based account of linear precedence. In *Proc. of the ACL'2001*, pages 180–187, 2001.

[DD08] M. Dekhtyar and A. Dikovsky. Generalized categorial dependency grammars. In Arnon Avron, Nachum Dershowitz, and Alexander Rabinovich, editors, *Pillars of Computer Science, Essays Dedicated to Boris (Boaz) Trakhtenbrot on the Occasion of His 85th Birthday*, volume 4800 of *Lecture Notes in Computer Science (LNCS)*, pages 230–255. 2008.

[Dik04] A. Dikovsky. Dependencies as categories. In *Proc. of the COLING'04 Workshop "Recent Advances in Dependency Grammars". Vienna : Austria*, 2004.

[Dik09] A. Dikovsky. Towards wide coverage categorial dependency grammars. In *Proceedings of the ESSLLI'2009 Workshop Parsing with Categorial Grammars - Parsing with Categorial Grammars Workshop ESSLLI 2009 Book of Abstracts, Bordeaux : France*, 2009.

[Gai65] H. Gaifman. Dependency systems and phrase structure systems. *Information and Control*, 8(3):304–337, 1965.

[HS07] J. Hockenmaier and M. Steedman. Ccgbank: A corpus of ccg derivations and dependency structures extracted from the penn treebank. *Comput. Ling.*, 33(3):355–396, 2007.

[Hud84] R.A. Hudson. *Word Grammar*. Basil Blackwell, 1984.

[KM03] D. Klein and C. Manning. Accurate unlexicalized parsing. In *Proc. of the First Annual Meeting of the ACL (ACL-41), Sapporo, Japan.*, pages 423–430, 2003.

[Mag94] D. Magerman. *Natural Language Parsing as Statistical Pattern Recognition*. PhD thesis, Dept of CS, Stanford University, CA, 1994.

[Niv07] Nivre et al. Multiparser: A language-independent system for data-driven dependency parsing. *Nat.Lang. Engineering*, 13(2):95–135, 2007.

[Pre03] J. Preiss. Using grammatical relations to compare parsers. In *Proc.of the Tenth Conference of the European Chapterof the ACL (EACL'03), Budapest, Hungary*, pages 291–298, 2003.

[ST93] D. Sleator and D. Temperly. Parsing english with a link grammar. In *Proc. of the Int. Workshop on Parsing Techn.'93*, pages 277–291, 1993.