Résumés de thèses

Rubrique préparée par Fiammetta Namer

Université de Lorraine, UMR « ATILF »
Fiammetta.Namer@univ-lorraine.fr

Cyril GROUIN : cyril.grouin@limsi.fr

Titre : Anonymisation automatique de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique

Mots-clés : anonymisation, comptes-rendus médicaux, guide d'annotation, méthodes symboliques, apprentissage statistique, traitement automatique des langues.

Title: *Clinical Records De-Identification: Performances and Limits of Rule-based and Machine-Learning based Approaches*.

Keywords: *De-identification, clinical records, guidelines, rule-based methods, machine-learning based approach, natural language processing*.

Thèse de doctorat en Informatique biomédicale, INSERM U872 Eq20 et LIMSI-CNRS, UFR des Sciences de la vie, UPCM, Orsay, sous la direction de Marie-Christine Jaulent (DR, INSERM U872) et Pierre Zweigenbaum (DR, LIMSI-CNRS). Thèse soutenue le 26/06/2013.

Jury : Marie-Christine Jaulent (DR, INSERM U872, codirectrice) et M. Pierre Zweigenbaum (DR, LIMSI-CNRS, codirecteur), Mme Anita Burgun (Pr, hôpital européen Georges-Pompidou/INSERM U872 Eq22, présidente), M. Stefan J. Darmoni (Pr, CHU de Rouen, rapporteur), Pr Pascal Staccini (Pr, CHU de Nice, rapporteur), M. Thierry Artières (Pr, LIP6-UPMC, examinateur).

Résumé **:** *Ce travail porte sur l'anonymisation automatique de comptes-rendus cliniques. L'anonymisation consiste à masquer les informations personnelles présentes dans les documents tout en préservant les informations cliniques. Cette étape est obligatoire pour utiliser des documents cliniques en dehors du parcours de soins, qu'il s'agisse de la publication de cas d'étude ou d'utilisation en recherche scientifique (mise au point d'outils informatiques de traitement du contenu des dossiers, recherche de cas similaires, etc.). Nous avons défini douze*

catégories d'informations à traiter : nominatives (noms, prénoms, etc.) et numériques (âge, dates, codes postaux, etc.). Deux approches ont été utilisées pour anonymiser les documents. L'une, dite « symbolique », est à base de connaissances d'experts formalisées par des expressions régulières et la projection de lexiques. L'autre fonctionne par apprentissage statistique au moyen de CRF de chaînes linéaires. Plusieurs expériences ont été menées parmi lesquelles l'utilisation simple ou enchaînée de chacune des deux approches. Nous obtenons nos meilleurs résultats (F-mesure globale = 0,922) en enchaînant les deux méthodes avec rassemblement des noms et prénoms en une seule catégorie (pour cette catégorie : rappel = 0,953 et F-mesure = 0,931). Ce travail de thèse s'accompagne de la production de plusieurs ressources : un guide d'annotation, un corpus de référence de 562 documents dont 100 annotés en double avec adjudication et calculs de taux d'accord interannotateurs (κ = 0,807 avant fusion) et un corpus anonymisé de 17 000 comptes-rendus cliniques.

URL où la thèse pourra être téléchargée : http://tel.archives-ouvertes.fr/tel-00848672

Enrique HENESTROZA ANGUIANO : ehenestroza@gmail.com

Titre : Analyse syntaxique probabiliste en dépendances : approches efficaces à large contexte avec ressources lexicales distributionnelles

Mots-clés : linguistique informatique, analyse syntaxique, ressources lexicales, machines à vecteurs supports, analyse à base de transitions, grammaires de dépendance, apprentissage semi-supervisé, adaptation de domaines.

Title: *Efficient large-context dependency parsing and correction with distributional lexical resources*

Keywords: *computational linguistics, syntactic parsing, lexical resources, support vector machines, transition-based parsing, dependency grammar, semi-supervised learning, domain adaptation.*

.

Thèse de doctorat en Informatique, UMR ALPAGE-INRIA & Paris-Diderot, UFR d'Informatique, Université Paris-Diderot, sous la direction de Laurence Danlos (Pr, Université Paris-Diderot), Marie Candito (MC, Université Paris-Diderot) et Alexis Nasr (Pr, Université d'Aix-Marseille). Thèse soutenue le 27/06/2013.

Jury : Mme Laurence Danlos (Pr, Université Paris-Diderot, codirectrice), Mme Marie Candito (MC, Université Paris-Diderot, codirectrice), M. Alexis Nasr

(Pr, Université d'Aix-Marseille, codirecteur), M. Matthieu Constant (MC, Université Paris-Est Marne-la-Vallée, président & rapporteur), M. Joachim Nivre (Pr, Uppsala University, rapporteur), M. Bernd Bohnet (MC, University of Birmingham, examinateur).

Résumé **:** *Cette thèse présente des méthodes pour améliorer l'analyse syntaxique probabiliste en dépendances. Nous employons l'analyse à base de transitions avec une modélisation effectuée par des machines à vecteurs supports (Cortes and Vapnik, 1995), et nos expériences sont réalisées sur le français. L'analyse à base de transitions est rapide, en raison de la faible complexité des algorithmes sous-jacents, eux-mêmes fondés sur une optimisation locale des décisions d'attachement. Ainsi notre premier fil directeur est l'élargissement du contexte syntaxique utilisé. Partant du système de transitions arc-eager (Nivre, 2008), nous proposons une variante qui considère simultanément plusieurs gouverneurs candidats pour les attachements à droite. Nous testons aussi la correction des analyses, inspirée par Hall and Novák (2005), qui révise chaque attachement opérant un choix parmi plusieurs gouverneurs alternatifs dans le voisinage syntaxique. Nos approches améliorent légèrement la précision globale, celle de l'attachement des groupes prépositionnels, ainsi que celle de l'attachement de la coordination. Notre deuxième fil explore des approches semi-supervisées. Nous testons l'auto-entraînement avec un analyseur en deux étapes, fondé sur McClosky* et al. *(2006), adapté au domaine journalistique puis au domaine médical. Nous passons ensuite à la modélisation lexicale à base de corpus, avec des classes lexicales généralisées pour réduire la dispersion des données, et des préférences lexicales de l'attachement des groupes prépositionnels pour aider à la désambiguïsation. Nos approches améliorent, dans certains cas, la précision et la couverture de l'analyseur, sans en augmenter la complexité théorique.*

URL où la thèse pourra être téléchargée : http://tel.archives-ouvertes.fr/tel-00860720

Frédéric LANDRAGIN : frederic.landragin@ens.fr

Titre : Dialogue homme-machine multimodal : de la pragmatique linguistique à la conception de systèmes

Mots-clés : dialogue naturel en langage naturel, pragmatique, référence, acte de langage, architecture logicielle, évaluation de systèmes.

Title: *Multimodal Human-Machine Dialogue: From Linguistic Pragmatics to System Design*

Keywords: *Natural dialogue in natural language, pragmatics, referring, speech act, software architecture, system evaluation.*

HDR en Informatique, Lattice (UMR 8094), UFR d'Informatique, Université Paris-Sud, Paris-Orsay, sous la direction de Anne Vilnat (Pr, Université de Paris-Sud). HDR soutenue le 28/06/2013.

Jury : Mme Anne Vilnat (Pr, Université de Paris-Sud, directrice), Mme Sophie Rosset (DR, LIMSI-CNRS, présidente), M. Harry Bunt (Pr, Université de Tilburg, rapporteur), Mme Catherine Schnedecker (Pr, Université de Strasbourg, rapporteur), Mme Mariët Theune (assistante Pr, Université de Twente, rapporteur), Mme Isabelle Tellier (Pr, Université Paris3, examinatrice).

Résumé : *Un des objectifs fondamentaux du dialogue homme-machine est de se rapprocher du dialogue naturel en langage naturel, c'est-à-dire de permettre une interaction entre la machine et son utilisateur humain dans la langue de celui-ci (langage naturel), avec une structure d'échanges similaire à un dialogue humain (dialogue naturel). Les recherches impliquées se nourrissent de travaux linguistiques qui analysent la langue et de travaux pragmatiques qui analysent l'usage du langage en contexte. Deux facettes importantes de la pragmatique linguistique portent ainsi sur les phénomènes de référence, par exemple les désignations des objets accessibles dans le contexte situationnel, et sur les actes de langage, ou actes de dialogue, c'est-à-dire les actions communicatives effectuées par les énoncés constituant les tours de parole. Nous présentons nos travaux de modélisation et de formalisation de ces deux facettes, avec leur application au dialogue avec support visuel et au dialogue associant parole et gestes coverbaux (dialogue multimodal). Un autre objectif du dialogue homme-machine est de mettre en œuvre des méthodologies et des moyens, par exemple des architectures logicielles réutilisables, pour faciliter le développement de systèmes. Nous présentons nos réflexions et nos réalisations dans ce sens, à travers notamment notre participation à un ensemble de projets européens. Nous proposons enfin des perspectives de recherche qui visent à mieux intégrer au dialogue homme-machine des phénomènes linguistiques et pragmatiques tels que la saillance et l'ambiguïté.*

URL où l'HDR pourra être téléchargée : http://tel.archives-ouvertes.fr/tel-00848533/