Résumés de thèses

Rubrique préparée par Fiammetta Namer

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

Pierre BEUST : (pierre.beust@unicaen.fr)

Titre : Pour une démarche centrée sur l'utilisateur dans les ENT. Apport au Traitement Automatique des Langues

Mots-clés : traitement automatique des langues, sémantique interprétative, énaction, couplage personne-système, environnements d'apprentissage, usages.

Title: *User-Centered Desgin of Digital Workspaces. Contribution to NLP*.

Keywords: *natural language processing, semiotics, e-learning, enaction, user-system interaction, digital workspaces, online education environments, usage.*

Mémoire de HDR en Informatique, GREYC CNRS UMR 6072, UFR de Sciences, Université de Caen, Caen, sous la direction de Nadine Lucas (MC-HDR, Université de Caen). HDR soutenue le 03/04/2013.

Jury : Mme Nadine Lucas (MC-HDR, Université Caen, directrice), M. Ioannis Kanellos (Pr, ENST de Bretagne, président et examinateur), M. Jean-Claude Bertin, (Pr, Université du Havre, rapporteur), M. Pierre De Loor (Pr, École nationale d'Ingénieurs de Brest, rapporteur), M. Luigi Lancieri (Pr, Université de Lille 1, rapporteur), Mme Pascale Sébillot (Pr, IRISA/INSA de Rennes, rapporteur), M. Mathieu Valette (Pr, INaLCO, examinateur).

Résumé : *Notre problématique de recherche est ancrée en traitement automatique des langues (TAL). Au sein du TAL, nous nous intéressons à la conception centrée-utilisateur d'environnements où les ressources et les processus mobilisés sont avant tout construits autour et en fonction des attentes et des capacités interprétatives de l'utilisateur. La conception centrée-utilisateur n'est pas une posture théorique mais c'est déjà une réalité dans des applications utilisées quotidiennement. C'est le cas des architectures Web 2.0 comme c'est également le cas des environnements numériques de travail (ENT). Notre recherche vise à analyser, concevoir et expérimenter des applications centrées-utilisateur dans les ENT où les capacités*

interprétatives s'enrichissent des éléments d'interaction dans l'environnement. Ce faisant nous cherchons à faire enrichir le TAL d'interconnexions avec les interactions homme-machine et les EIAH (environnements informatiques pour l'apprentissage humain).
La problématique de l'interprétation est ici omniprésente et elle nous incite à tirer des ponts entre disciplines : entre l'informatique et la linguistique, plus précisément le courant de la sémantique interprétative, et entre l'informatique et les sciences cognitives, plus précisément le courant de l'énaction. L'interprétation dans un environnement numérique n'est pas dissociable d'un couplage personne-système et de l'action de l'utilisateur dans cet environnement. Il en découle que nos objets d'étude sont principalement des usages et même des contournements d'usages vertueux par sérendipité. Les perspectives de recherche ouvertes s'orientent donc naturellement vers une mise en valeur de « l'agir interprétatif » dans les environnements numériques.

URL où l'HDR pourra être téléchargée : https://beust.users.greyc.fr/hdr/hdr.html

---

Clément DE GROC: (cdegroc@limsi.fr)

Titre : Collecte orientée sur le Web pour la recherche d'information spécialisée

Mots-clés : collecte, recherche d'information, Web, exploration orientée, recherche orientée.

Title: *Focused document gathering on the Web for domain-specific information retrieval*

Keywords: *focused crawling, focused search, domain-specific information retrieval, Web information retrieval.*

Thèse de doctorat en Informatique, LIMSI-CNRS, Université Paris-Sud, Orsay, sous la direction de Pierre Zweigenbaum (DR, LIMSI-CNRS) et de Xavier Tannier (MC, Université Paris-Sud). Thèse soutenue le 05/06/2013.

Jury : M. Pierre Zweigenbaum (DR, LIMSI, codirecteur), M. Xavier Tannier (MC, Université Paris-Sud, codirecteur), Mme Chantal Reynaud (Pr, Université Paris-Sud, présidente et examinatrice), M. Eric Gaussier (Pr, Université Joseph-Fourier - Grenoble, rapporteur), M. Jacques Savoy (Pr, Université de Neuchâtel, rapporteur), M. Mohand Boughanem (Pr, Université Paul Sabatier - Toulouse, examinateur).

Résumé : *Les moteurs de recherche verticaux, qui se concentrent sur des segments spécifiques du Web, deviennent aujourd'hui de plus en plus présents dans le*

paysage d'Internet. Les moteurs de recherche thématiques, notamment, peuvent obtenir de très bonnes performances en limitant le corpus indexé à un thème connu. Les ambiguïtés de la langue sont alors d'autant plus contrôlables que le domaine est bien ciblé. De plus, la connaissance des objets et de leurs propriétés rend possible le développement de techniques d'analyse spécifiques afin d'extraire des informations pertinentes.

Dans le cadre de cette thèse, nous nous intéressons plus précisément à la procédure de collecte de documents thématiques à partir du Web pour alimenter un moteur de recherche thématique. La procédure de collecte peut être réalisée en s'appuyant sur un moteur de recherche généraliste existant (recherche orientée) ou en parcourant les hyperliens entre les pages Web (exploration orientée).

Nous étudions tout d'abord la recherche orientée. Dans ce contexte, l'approche classique consiste à combiner des mots-clés du domaine d'intérêt, à les soumettre à un moteur de recherche et à télécharger les meilleurs résultats retournés par ce dernier.

Après avoir évalué empiriquement cette approche sur 340 thèmes issus de l'OpenDirectory, nous proposons de l'améliorer en deux points. En amont du moteur de recherche, nous proposons de formuler des requêtes thématiques plus pertinentes pour le thème afin d'augmenter la précision de la collecte. Nous définissons une métrique fondée sur un graphe de cooccurrences et un algorithme de marche aléatoire, dans le but de prédire la pertinence d'une requête thématique. En aval du moteur de recherche, nous proposons de filtrer les documents téléchargés afin d'améliorer la qualité du corpus produit. Pour ce faire, nous modélisons la procédure de collecte sous la forme d'un graphe triparti et appliquons un algorithme de marche aléatoire biaisé afin d'ordonner par pertinence les documents et termes apparaissant dans ces derniers.

Dans la seconde partie de cette thèse, nous nous focalisons sur l'exploration orientée du Web. Au cœur de tout robot d'exploration orientée se trouve une stratégie de crawl qui lui permet de maximiser le rapatriement de pages pertinentes pour un thème, tout en minimisant le nombre de pages visitées qui ne sont pas en rapport avec le thème. En pratique, cette stratégie définit l'ordre de visite des pages. Nous proposons d'apprendre automatiquement une fonction d'ordonnancement indépendante du thème à partir de données existantes annotées automatiquement.

URL où la thèse pourra être téléchargée : http://www.theses.fr (À venir)

Karen FORT : (karen.fort@loria.fr)

Titre : Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus

Mots-clés : annotation manuelle, méthodologie, évaluation, accords inter-annotateurs.

Title: *Annotated resources, the cornerstone of content analysis: towards a methodology for manual corpus annotation*

Keywords: *manual annotation, methodology, evaluation, inter-annotator agreements.*

Thèse de doctorat en Informatique, LIPN - UMR CNRS 7030, UFR d'Informatique, Paris 13 - Sorbonne Paris Cité, Villetaneuse, sous la direction de Adeline Nazarenko (Pr, Université de Paris 13). Thèse soutenue le 07/12/2012.

Jury : Mme Adeline Nazarenko, (Pr, Université de Paris 13, directrice), Mme Lori Lamel, (DR, LIMSI-CNRS, présidente), M. Frédéric Béchet (Pr, Université de la Méditerranée, rapporteur), M. Benoît Habert (Pr, ENS-Lyon, rapporteur), Mme Claire François (IR, INIST-CNRS, examinatrice), M. François Lévy (PE, Université Paris 13, examinateur), M. Eric de la Clergerie (CR, ALPAGE-INRIA, examinateur).

Résumé : *L'annotation manuelle de corpus est devenue un enjeu fondamental pour le traitement automatique des langues (TAL). En effet, les corpus annotés sont utilisés aussi bien pour créer que pour évaluer des outils de TAL. Or, le processus d'annotation manuelle est encore mal connu et les outils proposés pour supporter ce processus souvent mal utilisés, ce qui ne permet pas de garantir le niveau de qualité de ces annotations. Nous proposons dans cette thèse une vision unifiée de l'annotation manuelle de corpus pour le TAL. Ce travail est le fruit de diverses expériences de gestion et de participation à des campagnes d'annotation, mais également de collaborations avec différents chercheurs. Nous proposons dans un premier temps une méthodologie globale pour la gestion de campagnes d'annotation manuelle de corpus qui repose sur deux piliers majeurs : une organisation des campagnes d'annotation qui met l'évaluation au cœur du processus et une grille d'analyse des dimensions de complexité d'une campagne d'annotation. Un second volet de notre travail a concerné les outils du gestionnaire de campagne. Nous avons pu évaluer l'influence exacte de la pré-annotation automatique sur la qualité et la rapidité de la correction humaine, grâce à une série d'expériences menées sur l'annotation morpho-syntaxique de l'anglais. Nous avons également apporté des solutions pratiques concernant l'évaluation de l'annotation manuelle, en donnant au gestionnaire les moyens de sélectionner les mesures les plus appropriées. Enfin, nous avons mis au jour les processus en œuvre et les outils nécessaires pour une campagne d'annotation et instancié ainsi la méthodologie que nous avons décrite.*

URL où la thèse pourra être téléchargée : http://tel.archives-ouvertes.fr/tel-00797760/

Gaël PATIN : (gpatin@gmail.com)

Titre : Extraction interactive et non supervisée de lexique en chinois contemporain appliquée à la constitution de ressources linguistiques dans un domaine spécialisé

Mots-clés : unité lexicale, chinois contemporain, lexique, extraction non supervisée.

Title: *Interactive and unsupervised chinese lexicon extraction for linguistic ressources extraction on a domain-specific corpus*

Keywords: *lexical unit, contemporary chinese, lexicon, unsupervised extraction.*

Thèse de doctorat en Traitement Automatique des Langues, INALCO, laboratoire EA2520, Paris, sous la direction de Pierre Zweigenbaum (DR, LIMSI-CNRS). Thèse soutenue le 31/01/2013.

Jury : M. Pierre Zweigenbaum (DR, LIMSI-CNRS, directeur), M. Drocourt-Yang Zhitang (MC, INALCO, président/examinateur), Mme Béatrice Daille (Pr, Université Nantes, rapporteur), Mme Pascale Fung (Ass. Pr, Hong Kong University of Science and Technology, rapporteur), M. Nicolas Dessaigne (chef d'entreprise, Aloglia, examinateur), M. Alain Polguère (Pr, Université de Lorraine, examinateur).

Résumé **:** *Cette thèse traite de l'extraction d'unités lexicales en chinois contemporain à partir d'un corpus de textes de spécialité. Elle aborde la tâche d'extraction de lexique en chinois en utilisant des techniques se basant sur des caractéristiques linguistiques de la langue chinoise. La thèse traite également de la manière d'évaluer l'extraction de lexique dans un environnement industriel. La première partie de la thèse est consacrée à la description du contexte de l'étude. Nous nous attachons dans un premier temps à décrire les concepts linguistiques de lexique et d'unité lexicale, et nous donnons une description du processus de construction des unités lexicales en chinois contemporain. Nous faisons ensuite un inventaire des différentes techniques utilisées par la communauté scientifique pour traiter la tâche de l'extraction de lexique en chinois contemporain. Nous concluons cette partie par une description des pratiques d'extraction de lexique en milieu industriel, et nous proposons une formalisation des critères utilisés par les terminographes d'entreprise pour sélectionner les unités lexicales pertinentes. La deuxième partie du mémoire porte sur la description d'une méthode d'extraction de lexique en chinois contemporain et sur son évaluation. Nous introduisons une nouvelle méthode numérique non supervisée s'appuyant sur des caractéristiques structurelles de l'unité lexicale en chinois et sur des particularités syntaxiques de cette langue. La méthode comporte un module optionnel permettant une interaction avec un opérateur (*i.e. *semi-automatique). Dans la section consacrée à*

l'évaluation, nous évaluons d'abord le potentiel de la méthode en comparant les résultats de l'extraction avec un standard de référence et une méthode de référence. Nous mettons ensuite en œuvre une évaluation plus pragmatique de la méthode en mesurant les gains apportés par l'usage de la méthode en comparaison avec l'extraction manuelle de lexique par des terminographes. Les résultats obtenus par notre méthode sont de bonne qualité et sont meilleurs que ceux produits par la méthode de référence sur le standard de référence. Ces résultats sont encourageants, mais ils doivent être confirmés par une évaluation plus complète. L'évaluation pragmatique montre que la méthode n'améliore pas significativement la productivité des terminographes, mais permet d'extraire des unités lexicales différentes de celles obtenues manuellement.

URL où la thèse pourra être téléchargée :
https://docs.google.com/file/d/0B2922E-EiRcdSzY0TjIyUzE3Z3c/edit?usp=sharing