Résumés de thèses

Rubrique préparée par Fiammetta Namer

Université Nancy2 de Nancy, UMR "ATILF"
Fiammetta.Namer@univ-nancy2.fr

Julien BOURDAILLET (julien.bourdaillet@lip6.fr)

Titre : Alignement textuel monolingue avec recherche de déplacements : algorithmique pour la critique génétique

Mots-clés : traitement automatique des langues naturelles, algorithmique textuelle, alignement monolingue, critique génétique textuelle, distance d'édition avec déplacements, alignement par fragments, optimisation combinatoire mulitobjectif.

Title : *Monolingual textual alignment with move search : algorithm for genetic criticism.*

Keywords : *Natural Language Processing, Stringology, Monolingual alignment, Textual genetic criticism, Edit distance with moves, Fragment chaining alignment, Multiobjective combinatorial optimization.*

Thèse de doctorat en Informatique, Université de Pierre et Marie Curie, LIP6 (Laboratoire d'Informatique de Paris VI), département d'informatique, sous la direction de Jean Gabriel Ganascia, Pr. Soutenue le 03/12/2007.

Jury : M. Patrice Perny (Pr, Université de Pierre et Marie Curie, président), M. Jean-Gabriel Ganascia (Pr, Université de Pierre et Marie Curie, directeur), M. Maxime Crochemore, (Pr, Université de Marne-La-Vallée, rapporteur), Mme Béatrice Daille (Pr, Université de Nantes, rapporteur), M. Philippe Langlais (MC, Université de Montréal, examinateur), M. Jean-Louis Lebrave (DR, ENS, examinateur).

Résumé *: Ce travail de thèse répond à une problématique suscitée par la critique génétique textuelle. Cette discipline étudie la genèse des œuvres littéraires grâce aux brouillons d'écrivains en recherchant, entre autres, les déplacements entre deux versions d'un même texte. Ceci nous amené à définir la problématique de l'alignement textuel monolingue avec recherche de déplacements.*

D'un point de vue informatique, nous avons mis à jour la nécessité de calculer un

alignement entre deux textes de type distance d'édition avec recherche des déplacements, ce qui est NP-difficile. De plus, notre objectif était l'obtention d'un algorithme efficace permettant le passage à l'échelle, ce qui permet d'envisager la recherche des déplacements dans de longs textes comme des livres. Il devait également permettre l'alignement de textes très différents, tout en identifiant les modifications au caractère près.

Nous proposons une formalisation en un problème d'optimisation combinatoire multiobjectif intégrant un objectif syntaxique. Celle-ci permet de résoudre théoriquement de petites instances du problème avec un solveur SAT, mais ne permet pas le passage à l'échelle.

Un algorithme d'alignement de séquences basé sur l'alignement par fragments est proposé. En bioinformatique, cette technique permet de traiter des génomes de mammifères. Notre algorithme est basé sur la coordination de la résolution des recouvrements entre occurrences d'une répétition. Il résout le problème avec une complexité efficace et passe à l'échelle. De plus, il présente de meilleurs résultats que les méthodes existantes et est maintenant utilisé par les généticiens du texte afin d'étudier les brouillons d'écrivains.

URL où la thèse pourra être téléchargée :

http : //www-poleia.lip6.fr/~bourdaillet/ (bientôt disponible)

Valentina CEAUSU-DRAGOS (valentina.ceausu@math-infouniv-paris5.fr, valentina.dragos@lipn.univ-paris13.fr)

Titre : Définition d'un cadre sémantique pour la catégorisation de données textuelles. Application à l'accidentologie

Mots-Clés : fouilles de textes, ontologie, raisonnement à partir de cas.

Title : *Towards a semantic framework for textual data categorization. Application to accidentology.*

Keywords : *text mining, ontology, case-based reasoning.*

Thèse de doctorat en Informatique, Université de Paris-V René-Descartes, CRIP5 (Centre de Recherche en Informatique de Paris V), UFR de Mathématiques et Informatique, sous la direction de Mme Sylvie Després, MC (HDR). Soutenue le 18/06/2007.

Jury : M. Alain Mille (Pr, Université de Lyon I, président), Mme Sylvie Després (MC HDR, Université de Paris V, directrice), Mme Marie-Christine Jaulent, (DR, INSERM, rapporteur), Mme Adeline Nazarenko (Pr, Université Paris XIII, rapporteur), M. Dominique Fleury (DR, INRETS, examinateur), M. Yannick Toussaint (CR, INRIA-LORIA, examinateur), Mme Sylvie Szulman (MC, Université Paris XIII, examinatrice).

Résumé *: L'exploitation des connaissances propres à un domaine nécessite des techniques pour l'extraction, la modélisation et la formalisation de ces connaissances.*
Les travaux présentés dans cette thèse portent sur le développement de techniques relatives à la gestion de ressources sémantiques et à leur utilisation comme support à la mise en œuvre d'une application. Les techniques proposées ainsi que leur application définissent un cadre sémantique pour la catégorisation de données textuelles.
Sur le plan théorique, ce cadre sémantique permet : (i) l'extraction et la validation des connaissances à partir de textes ; (ii) la modélisation des connaissances contenues dans les textes sous la forme de ressources sémantiques ; (iii) la mise en correspondance entre une ressource ainsi créée et un corpus du domaine ; (iv) la mise en correspondance entre plusieurs ressources sémantiques modélisant un même domaine.
Sur le plan applicatif, ce cadre propose un mécanisme de raisonnement qui fait appel à ces techniques pour exploiter des connaissances en accidentologie. Il devient ainsi possible de mettre en œuvre une solution à un problème concret qui concerne l'exploitation automatique des scénarios type accidents de la route.
.

URL où la thèse pourra être téléchargée :

Contacter l'auteur

Emmanuel MORIN (emmanuel.morin@univ-nantes.fr)

Titre : Synergie des approches et des ressources déployées pour le traitement de l'écrit.

Mots-clés : fouille terminologique multilingue, modélisation du langage pour la reconnaissance de l'écriture manuscrite.

Mémoire de HDR en Informatique, Université de Nantes, LINA (Laboratoire d'Informatique de Nantes Atlantique), FRE CNRS 2729, sous la direction de Mme Béatrice DAILLE, Pr. Soutenue le 30/11/2007.

Jury : M. Christian Viard-Gaudin (Pr, Université de Nantes, président), Mme Béatrice Daille (Pr, Université de Nantes, directrice), M. Eric Gaussier (Pr, Université Joseph-Fourier, rapporteur), M. Gregory Grefenstette, (IR, CEA, rapporteur), M. Pierre Zweigenbaum (DR, LIMSI CNRS, rapporteur).

Résumé *: Les travaux présentés dans le cadre de cette Habilitation à Diriger des Recherches, qui se situent au carrefour de l'informatique et de la linguistique, s'intéressent au traitement de l'écrit. Ils s'articulent autour de deux axes de recherche, celui de la fouille terminologique multilingue et celui de la reconnaissance de l'écriture manuscrite en ligne.*

Dans un premier temps, notre étude est consacrée à la fouille terminologique multilingue. Nous commençons par rappeler les fondements théoriques en acquisition lexicale multilingue, qui s'inscrivent dans l'héritage de la sémantique distributionnelle de Harris. Nous présentons ensuite les travaux réalisés en acquisition de lexiques bilingues à partir de corpus comparables. Nous décrivons notamment la méthode par similarité interlangue proposée pour l'alignement de termes complexes et la plate-forme informatique associée. À la lumière des nombreux résultats que nous avons engrangés dans ce champ de recherche, nous précisons les apports et limites des différentes approches utilisées.

Dans un deuxième temps, nous présentons les différentes facettes de la reconnaissance de l'écriture manuscrite en ligne auxquelles nous nous sommes intéressés et les modèles développés. Ces travaux, qui se situent au niveau de la modélisation du langage naturel, visent à concevoir des modèles de langage adaptés à la reconnaissance de documents dénotant un « écrit standard » (où un stylo numérique vient remplacer la saisie sur un clavier numérique) ou un « écrit déviant » (où un stylo numérique s'offre comme une nouvelle alternative pour l'écriture de SMS). Nous présentons les modèles développés et les résultats obtenus. Nous revenons aussi sur l'importance et la difficulté de concevoir des ressources adaptées à la prise en compte de ces différents écrits.

Dans un dernier temps, qui constitue le trait d'union entre nos deux axes de recherche, nous indiquons la synergie possible entre les approches et ressources déployées. En particulier, nous montrons que les méthodes probabilistes ne sont plus une alternative aux systèmes à base de règles, mais bien complémentaires et que les ressources exploitées doivent être adaptées à la tâche visée.

URL où l'HDR pourra être téléchargée :

http : //www.sciences.univ-nantes.fr/info/perso/permanents/morin/

Farid NOUIOUA (Farid.Nouioua@lri.fr, Farid.Nouioua@lipn.univ-paris13.fr)

Titre : Extraction et utilisation des normes pour un raisonnement causal dans un corpus textuel

Mots-clés : normes, inférence, sémantique du langage naturel, raisonnement non monotone, raisonnement causal.

Title : *Extracting and using norms for a causal reasoning in a textual corpus.*

Keywords : *norms, inference, natural language semantics, non monotonic reasoning, causal reasoning causal.*

Thèse de doctorat en Informatique, Université de Paris XIII, LIPN (Laboratoire d'Informatique de Paris Nord), UMR 7030, UFR d'Informatique, sous la direction de M. Daniel Kayser, Pr. Soutenue le 25/04/2007.

Jury : M. Jean-Paul Haton (Pr, Université Henri-Poincaré, président), M. Daniel Kayser (Pr, Université Paris XIII, directeur), M. Pierre Marquis (Pr, Université d'Artois, rapporteur), M. Gérard Sabah, (DR, LIMSI, rapporteur), Mme Adeline Nazarenko (Pr, Université Paris XIII, examinatrice), M. Pascal Nicolas (Pr, Université d'Angers, examinateur),.

Résumé *: La simulation du processus de compréhension de la langue naturelle (LN) est l'une des ambitions qui ont motivé l'Intelligence Artificielle (IA). Les premières investigations dans ce domaine ont mis en évidence les sérieuses difficultés du problème, notamment pour les aspects liés à la sémantique et la pragmatique. Aujourd'hui, les recherches dans ce domaine sont partagées entre un courant formel qui cherche à fonder entièrement la sémantique de la LN sur des bases mathématiques mais dont l'apport à la compréhension de textes réels reste plutôt limité, et un autre courant qui s'oriente plus vers une ingénierie de la LN et essaie de développer des heuristiques efficaces pour des problèmes particuliers mais pour de gros corpus. On s'est éloigné ainsi des objectifs initiaux de l'IA qui, en parallèle, a réalisé des progrès importants dans le développement d'outils pour la formalisation du raisonnement de bon sens.*

Dans cette thèse, nous nous proposons de profiter des avancées de l'IA en terme de formalisation du raisonnement pour reconsidérer le problème de la compréhension automatique de la LN. Les bases de notre approche sont, cependant, très différentes de celles utilisées dans le cadre des méthodes formelles. En considérant que le but principal d'une compréhension automatique est de parvenir, à partir d'un texte, aux mêmes conclusions qu'un lecteur humain ordinaire, nous défendons l'idée d'une sémantique inférentielle dans laquelle les mots déclenchent des inférences visant à satisfaire simultanément différentes contraintes provenant de multiples niveaux : lexical, syntaxique, sémantique et pragmatique. Nous montrons que ces inférences sont essentiellement bâties sur nos connaissances communes des normes du domaine (le mot « norme » réfère ici au déroulement normal des événements). Le sens d'un énoncé déclaratif n'est plus assimilé à ses conditions de vérité sur un

univers de discours ; il résulte d'un équilibre dans le processus de satisfaction des contraintes linguistiques et extralinguistiques. Puisque les conclusions accessibles grâce aux normes peuvent être remises en cause, le raisonnement que nous utilisons est de nature non monotone.
*Nous travaillons sur un corpus de textes décrivant des accidents de la route et nous cherchons à répondre automatiquement à la question de savoir la cause de ces accidents. Bien que le système que nous ayons développé garde une certaine séquentialité dans ses traitements et ne corresponde donc pas entièrement à l'idée d'un équilibre satisfaisant simultanément des contraintes de natures très différentes, il est en grande partie basé sur un processus inférentiel et ses différentes étapes se servent de connaissances d'ordre sémantique. La question à laquelle nous cherchons à répondre nous conduit à être confrontés à la notion de causalité qui est très intuitive mais dont la nature intrinsèque est sujette à controverses. Loin de se lancer dans des débats sur cette nature, nous la considérons d'un point de vue qui, d'une part, se rapproche de l'aspect action de la cause, aspect privilégié par l'IA, et d'autre part et surtout, établit une liaison entre les notions de cause et de norme. Nous postulons ainsi, que parmi les différentes causes possibles d'un événement anormal, celle qui nous paraît la plus plausible est celle qui correspond à la violation de la norme la plus spécifique évoquée dans le texte. Pour la représentation des connaissances et le raisonnement, nous avons défini un langage de premier ordre réifié prenant en compte les modalités utiles, l'aspect temporel et les inférences non monotones exprimées à l'aide du formalisme des défauts de Reiter. Pour la mise en œuvre, nous avons utilisé le paradigme de programmation par ensemble de réponses (*Answer Set programming*) : nos règles d'inférence sont traduites en programmes logiques étendus exprimés dans le langage Smodels.*

URL où la thèse pourra être téléchargée :

http : //www-lipn.univ-paris13.fr/~nouioua/These_Nouioua_F.pdf

---

Thibaud ROY (roy.thibault@wanadoo.fr)

Titre : Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires

Mots-clés : traitement automatique des langues naturelles, sémantique, interfaces utilisateurs, cartographie, informatique, gestion électronique de documents, représentations de connaissances, logiciels interactifs et individu-centrés, accès au contenu d'ensembles de textes

Title : *Interactive visualizations for personal help in interpretation of sets of documents*

Keywords : *natural langage processing, semantics, user interface, cartography – computer science, electronic management of documents, knowledge representation, interactive and user-centered softwares, accessing the content of sets of texts.*

Thèse de doctorat en Informatique, Université de Caen – Basse-Normandie, GREYC (Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen), UMR 6072, Département d'informatique/UFR de Sciences, sous la direction de M. Jacques Vergne, Pr et M. Pierre Beust, MC. Soutenue le 17/10/2007.

Jury : Mme Adeline Nazarenko (Pr, Université Paris XIII, présidente), M. Jacques Vergne (Pr, Université de Caen, codirecteur), M. Pierre Beust (MC , Université de Caen, codirecteur), M. Benoît Habert (Pr, ENS LSH Lyon, rapporteur), M. Pierre Zweigenbaum, (DR, LIMSI, rapporteur).

Résumé *Cette thèse prend place en informatique et plus particulièrement en Traitement Automatique des Langues. À l'heure actuelle, avec la multiplication des documents électroniques sur différents réseaux informatiques, et plus particulièrement sur Internet, les utilisateurs y cherchant une information se retrouvent très souvent face à une véritable montagne de documents difficile à gravir. Ce travail a l'objectif d'aider les utilisateurs dans de telles situations en leur proposant des aides logicielles pour appréhender le contenu d'ensembles documentaires. Afin de fournir aux utilisateurs de nouvelles aides logicielles pour un tel accès, nous partons d'un double constat : les systèmes traditionnellement proposés (tels par exemple les moteurs de recherche sur Internet ou des outils de recherche dans les ouvrages des bibliothèques) ne prennent que très peu en considération le point de vue de l'utilisateur et ne lui permettent pas de visualiser globalement et interactivement le résultat de sa recherche.*

En suivant ce constat, nous faisons l'hypothèse que de nouvelles aides logicielles dans des tâches d'accès au contenu d'ensemble de documents doivent tout d'abord prendre en considération le point de vue de leur utilisateur, comme ses domaines d'intérêt sur la tâche qu'il vise. Ensuite, pour que ce dernier puisse réellement accéder au contenu d'un ensemble documentaire, nous faisons comme seconde hypothèse qu'il est nécessaire de le laisser visualiser son ensemble documentaire et interagir avec cet ensemble, en passant, par exemple, d'une vue globale de l'ensemble documentaire, à une vue plus locale, comme un sous-groupe de documents aux contenus jugés similaires. Ainsi, nous définissons dans cette thèse le modèle AidED (Analyse Interprétative d'Ensembles Documentaires) proposant tout d'abord de représenter le point de vue de chaque utilisateur sur les domaines de son intérêt par des ensembles de termes décrits et organisés selon un modèle centré-utilisateur de sémantique lexicale différentielle (le modèle LUCIA développé depuis plusieurs années au sein du laboratoire d'informatique de l'Université de Caen, avec notamment les travaux de Pierre Beust et Vincent Perlerin). AIdED propose ensuite d'utiliser les représentations des domaines d'intérêt de l'utilisateur afin de

produire des cartes personnalisées interactives de l'ensemble documentaire mettant en évidence des regroupements, des liens et des différences entre documents de l'ensemble.

Dans le but d'opérationnaliser de telles propositions, nous avons mis au point une instrumentation logicielle du modèle AIdED, afin de permettre à l'utilisateur de construire ses domaines d'intérêt sur sa tâche d'accès au contenu et de les projeter en ensembles documentaires pour obtenir des cartes personnalisées et interactives d'ensembles documentaires. La construction de telles cartes se fait via *la plate-forme ProxiDocs. À partir d'un ensemble documentaire et d'ensembles de termes décrivant les domaines d'intérêt de l'utilisateur, la plate-forme réalise un certain nombre de traitements statistiques et linguistiques afin de produire différentes vues cartographiques de l'ensemble. Ces cartes proposent à la fois différents niveaux de visualisation : ensemble documentaire dans sa globalité, sous-ensemble de documents constitué automatiquement par la plate-forme selon les domaines de l'utilisateur, document original où les termes des ressources de l'utilisateur sont mis en évidence, ainsi que des vues temporelles mettant en évidence l'évolution des domaines dans l'ensemble documentaire au fil du temps. L'utilisateur est alors laissé libre de passer interactivement d'un niveau de visualisation à un autre et d'une période à une autre, ceci afin de lui permettre de réaliser sa propre appropriation de l'ensemble documentaire et ainsi réaliser ses propres parcours lui permettant de se faire son interprétation de l'ensemble de documents.*

Différentes utilisations du modèle AidED et de son instrumentation ont été réalisées afin de mettre à l'épreuve nos hypothèses. Ces utilisations ont pris place dans des contextes expérimentaux très variés, allant notamment de la recherche d'information sur Internet à l'étude linguistique d'expressions métaphoriques. Une de ces utilisations a consisté à produire des cartes d'une partie des pages retournées par un moteur de recherche sur Internet à partir de domaines d'intérêt de l'utilisateur. Le parcours visuel et interactif des cartes a permis à l'utilisateur d'appréhender globalement le contenu de l'ensemble des documents retournés, de visualiser différentes catégories thématiques de pages, et ainsi d'orienter efficacement la recherche d'information. Une deuxième expérimentation, prenant place dans un contexte très différent, a cherché à illustrer la pertinence de nos hypothèses pour l'analyse linguistique d'expressions métaphoriques. En représentant finement les domaines d'étude de ces expressions, nous avons pu à la fois mettre en évidence des sous-ensembles de documents partageant de mêmes métaphores, ainsi que des liens étroits entre la présence de certaines métaphores et l'actualité. D'autres exploitations du modèle ont également été réalisées, notamment dans le cadre de collaborations, et dans des domaines d'applications variés tels l'analyse de forums de discussions ou encore l'indexation de documents médicaux.

Cette thèse met ainsi en évidence le grand besoin de personnalisation et d'interaction dans des tâches très variées d'accès au contenu d'ensemble de documents. Les principales perspectives offertes par ce travail de recherche consistent à aller encore plus loin dans la prise en considération de l'utilisateur et

de ses interactions pour l'interprétation d'ensemble de documents en lui proposant, par exemple, de nouvelles visualisations et interactions. Une véritable évaluation du couple utilisateur/outil-logiciel devra être également réalisée. Cette évaluation n'est pas sans poser problème, car il n'est pas question ici d'évaluer comment fonctionnent nos logiciels, en terme de complexité algorithmique, par exemple, mais plutôt de s'intéresser à la valeur ajoutée de notre système pour un utilisateur dans des tâches d'accès aux contenus d'ensembles documentaires.

URL où la thèse pourra être téléchargée :

http : //users.info.unicaen.fr/~troy/these/manuscrit/These_Thibault_Roy.pdf