TEXTE

Exploration et Exploitation de données textuelles

Effectifs au 30/06/2013 :
5 permanents (2,5 ETP)
5 doctorants (3,7 ETP)
3 autres personnels (2,5 ETP)

Nombre de thèses soutenues entre le 01/01/2008 et le 30/06/2013 : 3,5

Responsable :
Mathieu Roche
Co-responsable :
Mathieu Lafourcade

Page Internet de l'équipe :
http://www.lirmm.fr/recherche/equipes/texte

Laboratoire Informatique Robotique Microélectronique Montpellier

ANALYSE SYNTAXIQUE, SEMANTIQUE LEXICALE, FOUILLE DE TEXTES

Présentation

Les travaux de recherche de l'équipe TEXTE couvrent plusieurs spécialités du TALN (Traitement Automatique du Langage Naturel). L'équipe concentre ses problématiques de recherche, à la fois théoriques et pratiques, autour de deux axes :

1. l'analyse des données textuelles,
2. la sémantique lexicale.

Dans le premier axe, l'analyse des textes s'effectue à un niveau fin (par exemple, l'analyse syntaxique) jusqu'à une granularité plus élevée (travaux en fouille de textes). Le second axe consiste à construire ou enrichir des réseaux lexicaux. Cette acquisition peut s'effectuer par l'analyse des textes mais également de façon contributive, par exemple par l'intermédiaire des jeux sérieux. Ces derniers ont permis de construire un réseau lexical sans précédent composé de plus d'un million de relations sémantiques entre des centaines de milliers de termes. Les travaux menés autour de ces deux axes ont permis des retombées significatives en titrage, traduction ou classification de documents. Par ailleurs, l'exploration et l'exploitation des données textuelles revêt une dimension éminemment pluridisciplinaire que l'équipe TEXTE ne cesse de cultiver avec des linguistes et des géographes.

Evolution de l'équipe

Des évolutions scientifiques sont intervenues suite au départ de J. Chauché qui reste un membre associé très actif sur la thématique « syntaxe » développée par l'équipe. L'arrivée de J.P. Prost en 2010 a permis une évolution vers la logique de cette thématique importante. Par ailleurs, depuis 2011, l'équipe TAL (animée par V. Prince et et M. Roche) est devenue l'équipe TEXTE (animée par M. Roche et M. Lafourcade). Le rayonnement de l'équipe depuis sa création a permis de bâtir un nouveau projet reposant sur ses « compétences historiques » associées au premier axe tout en développant des thèmes de recherche nouveaux et difficiles liés au second axe (par exemple, les jeux sérieux pour la construction de ressources lexicales, l'analyse de sentiments, la recherche d'information biomédicale). Le dynamisme de l'équipe depuis 2008 est concrétisé par le nombre significatif de thèses et d'HDR soutenues.

Organisation et Vie de l'équipe

L'identification des deux axes de recherche de l'équipe n'empêche pas les nombreux travaux transverses qui permettent un enrichissement et une plus grande ouverture scientifique. Par exemple, l'acquisition d'information sémantique à partir de textes (axe 2) peut s'effectuer à travers des informations syntaxiques (axe 1). Ceci encourage les collaborations entre les membres de l'équipe. Les réunions d'équipe présentant les travaux en cours ou finalisés des doctorants et permanents consolident ces collaborations intra-équipes. Les séminaires bi-mensuels largement ouverts à des invités extérieurs permettent également de confronter les points de vue sur des problèmes théoriques et appliqués du TALN. En outre, dans le but d'aider à lever des verrous scientifiques, l'équipe TEXTE encourage les collaborations inter-équipes (en particulier, avec TATOO, COCONUT, SMILE, MAREL) ce qui renforce l'animation des Pôles Données et Connaissances et Intelligence Artificielle du département informatique. Ces collaborations inter-équipes se sont souvent concrétisées par des co-encadrements de thèses.

Activités scientifiques

De l'analyse des données textuelles à la sémantique lexicale : un cycle vertueux développé par l'équipe TEXTE.
Dans un premier temps, l'analyse fine des textes (analyse syntaxique par exemple) permet d'obtenir des informations précieuses dans un processus d'exploration et d'exploitation des données textuelles. De plus, une analyse par des méthodes de fouille de textes aide à la découverte de connaissances nouvelles. Dans un second temps, ces informations et connaissances permettent de construire et/ou d'enrichir des ressources lexicales (dictionnaires, réseaux sémantiques, etc). Ce type de connaissance est crucial dans un processus d'analyse des données textuelles.

Analyse des données textuelles

L'analyse syntaxique a pour but la construction de structures syntaxiques à partir des textes. L'analyse doit être robuste, c'est-à-dire qu'elle doit fournir un résultat, même partiel, dans tous les cas. L'approche définie par SYGFRAN spécifique au français repose sur une construction algorithmique d'un filtre constructiviste de structures partielles. Cette construction s'effectue par la définition de règles de réécriture sur des ensembles de structures. Le modèle théorique des systèmes de réécriture s'appuie sur les algorithmes de Markov.
La participation à différentes campagnes d'évaluation (en particulier, EASY) a montré la robustesse de l'approche SYGFRAN développée dans l'équipe. Les résultats de l'analyse syntaxique sont utilisés dans de nombreuses thèses de l'équipe (Nicolas Béchet, Alexandre Labadié, Cédric Lopez, Johan Ségura, Guillaume Tisserant) et de nombreux travaux de l'équipe décrits dans ce document.

La dimension fondamentale des travaux liés à l'analyse syntaxique repose sur deux axes de développement :

- l'un relatif à la modélisation et le développement de stratégies algorithmiques d'analyse syntaxique profonde et robuste pour du langage tout-venant, et
- l'autre relatif à l'interface syntaxe-sémantique et sa formalisation pour des cadres de représentation des connaissances linguistiques fondés sur la théorie logique des modèles.

Ces travaux font l'objet de collaborations à la fois internes à l'équipe et intra-équipes (équipe COCONUT du LIRMM et laboratoire LIFO à Orléans).

Analyse et Fouille de Textes.
Avec l'émergence des réseaux sociaux et de l'amorce possible d'une e-démocratie, l'analyse automatique des sentiments véhiculés dans les textes devient un enjeux sociétal majeur. Cette thématique est notamment abordée au sein du projet MSH-M Senterritoire (2012-2013) qui traite d'analyse de sentiments liés à l'aménagement du territoire [TKwims13]. Depuis 2008, en collaboration avec l'équipe TATOO et le laboratoire TETIS, nos travaux ont permis de proposer des méthodes originales d'acquisition automatique d'un vocabulaire d'opinion adapté à un domaine [HPdn08]. Nos approches reposent sur la combinaison de méthodes d'extraction d'Entités Nommées, de fouille de textes (identification d'associations entre mots d'opinion) et de fouille du Web pour valider le vocabulaire appris.

Le traitement des textes issus des réseaux sociaux, données volumineuses et souvent bruitées, a été mené dans le cadre d'une collaboration avec l'Université d'Ottawa (Canada) et l'équipe TATOO. Nos études, qui consistent entre autres à déterminer automatiquement des « communautés d'opinion », ont été reconnues par plusieurs publications scientifiques depuis 2010 (par exemple, [BPplead12]) et des retombées médiatiques importantes (articles dans la presse régionale et nationale). Notons que les messages (tweets, SMS, etc.) de ces réseaux de communication ont des spécificités lexicales et/ou syntaxiques qui sont également étudiées dans le cadre de projets académiques (MSH-M, DGLFL, PEPS) en collaboration avec des linguistes (Praxiling). Concrètement, cette collaboration pluridisciplinaire nous a permis de mettre en œuvre un processus complet de traitement de données SMS (collecte d'un corpus conséquent, anonymisation, transcodage, analyse). Ces approches reposent sur des méthodes semi-automatiques en plaçant l'utilisateur au cœur du processus [APli12] et par apprentissage automatique en mode supervisé [PAcicling13].

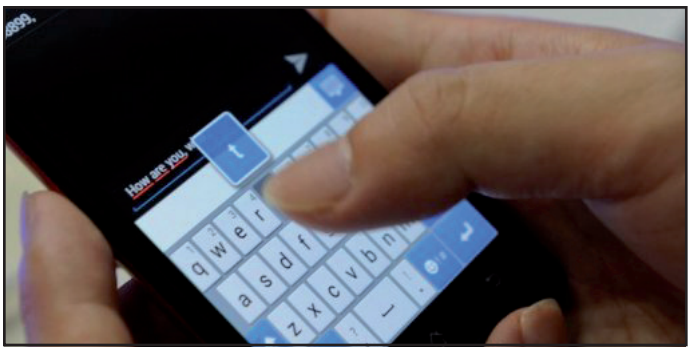

Figure 1 : TAL et données SMS

Sémantique Lexicale

La possibilité effective d'une acquisition d'informations lexicales via une activité ludique a été démontrée à travers le projet JeuxDeMots [LJtal09]. Cette acquisition prend la forme de la construction incrémentale d'un réseau lexical où les relations sont orientées, typées et pondérées. L'activité ludique est ici la motivation qui pousse les utilisateurs à aboutir à une construction par consensus populaire, sans négociation. Les participants n'ont pas besoin d'avoir conscience qu'ils contribuent à la construction d'une ressource lexicale. En effet, lors d'une partie, les joueurs ne sont pas en contact et ne peuvent donc pas négocier leurs réponses. PtiCLic (travaux avec le LIDILEM) est une variante de JeuxDeMots mettant l'accent sur la consolidation du réseau via une activité de réattribution de relations pour des couples de termes [ZLsticef11]. Contrairement à une approche à partir de définitions, le réseau lexical de JeuxDeMots ne fournit pas directement de sens pour les termes. Toutefois, il est possible de déduire au moins partiellement des usages pour chaque terme.

En 5 ans, plus de 2.200.000 relations entre 260.000 termes ont été collectés et sont disponibles librement pour la communauté scientifique. L'équipe texte utilise cette ressource dans de nombreux projets (aussi bien internes qu'en coopération académique/industrielle nationale et internationale) et notamment ceux ayant trait à la sémantique lexicale et analyse sémantique de textes. La ressource lexicale créée (RL-JDM) contient plus de 50 types de relations différentes aussi bien lexicales (synonymes, termes dérivés, etc.), qu'ontologiques (hyperonymes, partie/tout/...), de rôle sémantiques (agent, patient, instrument, etc.), de valeurs temporelle, causale, de polarité et sentiments. Depuis 2008, plus de 20 articles en revues et conférences ont été publiés relativement a ce projet. Une douzaine de versions de JeuxDeMots ont été construites pour d'autres langues (aussi bien par le LIRMM que d'autres partenaire : LIG, Thaïlande, etc.). Enfin, dans le but d'évaluer à grande échelle les données RL JDM, un contre-jeu appelé AKI (cf. Figure 2) a été proposé.

Figure 2 : AKI

Outre l'acquisition lexicale de manière contributive différents travaux développés dans l'équipe TEXTE consistent à enrichir des ontologies à partir de termes extraits dans les textes. Cet enrichissement s'appuie sur des méthodes combinant l'exploitation d'informations morpho-syntaxiques (par exemple, utilisation du verbe comme élément pivot) et d'informations statistiques. Ces travaux sont menés dans le cadre d'une thèse en co-direction avec l'IRIT (thèse soutenue en 2010) et d'une thèse en collaboration avec l'INRA et AgroParisTech (en cours).

Enfin, des travaux fondamentaux de l'équipe TEXTE liés à la sémantique lexicale vise sur le long terme à unir deux types de modèles sémantiques du langage naturel, les modèles sémantiques vectoriels fondés sur les probabilités et les modèles fonctionnels qui s'appuient sur la logique. Les espaces vectoriels sont munis d'une variante de la logique quantique adaptée aux vecteurs linguistiques et transforme tout modèle logique en un espace vectoriel suivant la probabilité choisie. Ceci permet de démontrer sous quelles conditions la logique quantique coïncide avec la logique des prédicats, réunissant ainsi le traitement efficace de grandes quantités de données à la déduction automatisée de la logique mathématique.

Applications

Les applications décrites dans cette section reposent sur des travaux transverses aux deux axes de recherche de l'équipe.

Titrage automatique. Alors que de nombreux travaux ont été publiés concernant le résumé automatique, le titrage automatique demeurait jusqu'alors discret et connaissait quelques difficultés quant à son positionnement dans le domaine du TALN. La thèse de Cédric Lopez (2009-2012) a permis de mettre en exergue les spécificités de cette tâche de titrage. Concrètement, une série de méthodes originales reposant sur des approches mixtes (statistiques et linguistiques) ont permis de produire des ensembles de titres à la fois informatifs (dans 81% des cas) et accrocheurs (dans 78% des cas). L'évaluation est réalisée par externalisation ouverte (crowdsourcing). Les méthodes mises en place s'appuient sur des connaissances syntaxiques (SYGFRAN) et sémantiques (JeuxDeMots) développées dans l'équipe. Ces travaux ont donné lieu à des publications dans les meilleures conférences du domaine (NAACL, EACL, RANLP).

Traduction automatique. Les travaux liés à la traduction automatique issus de la thèse de Johan Ségura (2009-2012) traitent plus précisément de l'alignement sous-phrastique bilingue. Les travaux reposent sur l'analyseur syntaxique SYGFRAN et un processus évolutif à base d'exemples. Une partie importante du travail mené a consisté à définir un cadre formel sous-tendant une architecture originale à base d'exemples alignés. De nouvelles méthodes d'alignement ont été proposées et comparées à des références reconnues.

Biomédecine. Dans le but d'enrichir des ressources de biomédecine, l'équipe TEXTE a développé des méthodes de fouille de textes spécifiques aux données biomédicales. Ces approches ont, par exemple, permis d'extraire et désambiguïser les sigles très présents dans ce domaine. Par ailleurs, un travail dédié au traitement des textes biomédicaux en français afin d'en extraire une terminologie a été initié depuis 2012 dans le cadre :

- d'une thèse co-encadrée avec l'équipe SMILE et le laboratoire TETIS,
- de notre participation à une ANR Jeunes Chercheurs.

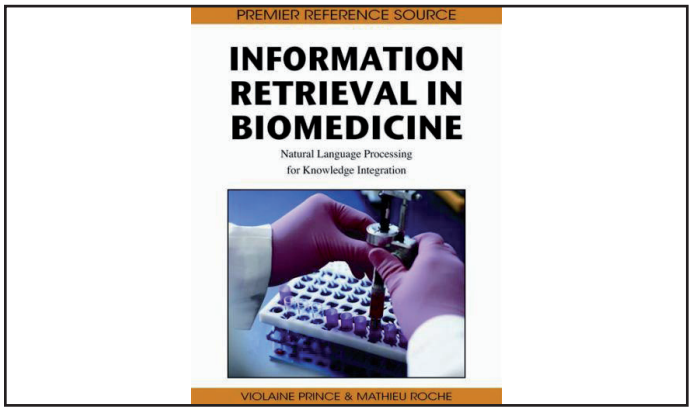

Figure 3 : V. Prince, M. Roche (Eds) (2009) Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration. Medical Information Science Reference, IGI Gobal, 460 pages

Applications industrielles. L'équipe TEXTE a établi de nombreuses collaborations avec des entreprises locales et nationales : Itesoft, Lingua et Machina, NatutalPad, Success Together, Namae concept, Expernova, Satin Technologies, Web Report, Twiam, Orange Lab, Darryl, Nout, Viseo, etc. Ces collaborations se sont concrétisées par plusieurs publications (démonstrations logiciels notamment) et des co-encadrements de stages Recherche de M2 et de 2 thèses en collaboration avec l'équipe TATOO (CIFRE et thèse financée sur contrat).

Faits marquants

- Organisation de 2 conférences (plus de 320 participants) en 2011 : TALN (Traitement Automatique des Langues Naturelles), L ACL (Logical Aspects of Computational Linguistics)
- JeuxDeMots : Production de ressources sémantiques massives et libres : plus de 2.200.000 relations entre 260.000 termes
- Coordination de projets pluridisciplinaires : 1 projet MASTODONS CNRS en 2013, 1 projet MSH-M en 2012
- Développement d'un groupe de travail de dimension internationale « Senterritoire » en collaboration avec l'UMR TETIS : 1 projet MSH-M, 1 Post-Doc Labex Numev, invitation de chercheurs invités internationaux

Collaborations externes

Collaborations internationales :

- Université d'Ottawa - Canada (visite/invitation) – 3 publications communes
- Université d'Oxford - Grande Bretagne (visite/invitation) - 4 publications communes
- Faculty of Business Informatics, Moscow, Russia (invitation UM2 d'un professeur)

Collaborations nationales :

- LIDILEM (Grenoble), 6 publications communes
- INRA - AgroParisTech (Montpellier - Paris) : 1 thèse en commun depuis 2012 (Labex Numev/INRA)

Publications majeures

- M. Lafourcade, A. Joubert (2009). Similitude entre les sens d'usage d'un terme dans un réseau lexical. Dans Traitement Automatique des Langues (TAL), Vol. 50, Numéro 1. Varia, pp. 179-200.
- R. Kessler, N. Béchet, M. Roche, J.M. Torres-Moreno, M. El-Bèze (2012). A hybrid approach to managing job offers and candidates. Information Processing & Management (IPM), Elsevier, Vol. 48, Issue 6, pp. 1124-1135.
- C. Lopez, V. Prince, M. Roche (2012). NOMIT: Automatic Titling by Nominalizing. In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Full paper, pp. 274-283
- S. Pogodalla, J.-P. Prost (Eds) (2011). Proceedings of Logical Aspects of Computational Linguistics. Lecture Notes in Computer Science (LNCS), Springer Verlag
- A. Preller, M. Sadrzadeh (2011). Semantic Vector Models and Functional Models for Pregroup Grammars. Journal of Logic, Language and Information, Vol. 20, Issue 4, pp. 419-443