

# Rapport de missions 2014

Natalia Grabar

En 2014, le laboratoire STL a financé une partie de trois missions effectuées (TALN 2014 à Marseille, ICHI 2014 à Verona, Italy et CompuTerm 2014 à Dublin, Ireland). Ce financement concernait les frais de transport entre Paris et le lieu des missions. La conférence MIE 2014 figure dans ce rapport parce que Mai Thi Tran a bénéficié de l'aide du laboratoire pour effectuer sa mission à l'occasion de cette conférence.

## 1 TALN 2014

La conférence Traitement Automatique des Langues Naturelles (TALN) a eu lieu à Marseille, en juin 2014. La conférence est consacrée aux travaux sur les différentes thématiques de TAL (morphologie, lexique, syntaxe, dictionnaires, traduction automatique, émotions, etc.). J'ai participé à cette conférence pour assurer la présentation d'un travail dédié à la détection automatique des paraphrases dans les corpus oraux :

– Iris Eshkol-Taravella, Natalia Grabar

*Repérage et analyse de la reformulation paraphrastique dans les corpus oraux*

TALN 2014. Marseille, France.

L'approche proposée est une approche syntagmatique qui tient compte des marqueurs de reformulation paraphrastique et des spécificités de l'oral. L'annotation manuelle effectuée par deux annotateurs permet d'obtenir une description fine et multidimensionnelle des données de référence. Une méthode automatique est proposée afin de décider si les tours de parole comportent ou ne comportent pas des reformulations paraphrastiques. Les résultats obtenus montrent jusqu'à 66,4 % de précision. L'analyse de l'annotation manuelle indique qu'il existe peu de segments paraphrastiques avec des modifications morphologiques (flexion, dérivation ou composition) ou de segments qui montrent l'équivalence syntaxique.

À l'occasion de cette conférence, j'ai rencontré et discuté avec plusieurs collègues, y compris les collègues industriels. Un des bénéfices de ces rencontres est la réactivation des relations scientifiques et la motivation de préparer et soumettre des projets de recherche.

Les actes de la conférence sont accessibles en ligne :

<http://www.aclweb.org/anthology/F/F14/>

## 2 ICHI 2014

La conférence International Conference on Healthcare Informatics (ICHI) a eu lieu à Verona, Italy en septembre 2014. Cette conférence a lieu sous l'égide de IEEE. La conférence est dédiée aux applications proposées par plusieurs domaines de recherche (informatique, sciences de l'information et de communication) afin de satisfaire des problèmes et des besoins de la santé et des soins, de la santé publique, et du bien-être des patients. Lors de cette conférence, j'ai présenté un travail lié

à la lisibilité des documents médicaux par les patients et, plus particulièrement, à l'acquisition de paraphrases compréhensibles par les patients pour des termes médicaux :

- Natalia Grabar, Thierry Hamon  
*Automatic extraction of layman names for technical medical terms.*  
ICHI 2014. Pavia, Italy

La motivation principale est que les informations médicales et de santé sont répandues dans la société moderne (problèmes de santé individuels, média, romans...). Cependant, le domaine médicale véhicule de nombreuses notions spécifiques et souvent opaques, comme *myocardial infarction*, *cholecystectomy*, *abdominal strangulated hernia*, *galactose urine*. De telles notions restent difficiles à comprendre par les non spécialistes. Nous proposons une méthode automatique pour l'acquisition de paraphrases pour les termes médicaux techniques. Nous voulons ainsi que de tels paraphrases soient plus faciles à comprendre que les termes d'origine. La méthode est basée sur l'analyse morphologique des termes et sur la fouille de texte provenant des médias sociaux (forums, wikipedia) en français. L'analyse des résultats et leur évaluation indiquent que les paraphrases pour des termes médicaux peuvent effectivement être trouvées dans les textes non spécialisés et que le niveau de lisibilité, ou de compréhension, de ces termes est plus abordable :

- *dorsalgie* : douleur dans le dos
- *myélocyte* : cellules dans la moelle osseuse
- *glycohémoglobine* : quantité de sucre par litre de sang
- *gastralgie* : douleurs à l'estomac
- *desmorrhexie* : rupture des ligaments
- *cardiospasme* : maladie qui spasme les coronaires et n'alimente plus le coeur
- *hépatite* : inflammation du foie
- *immunosupprimé* : diminue le système immunitaire pour éviter le rejet

En fonction de la sémantique des termes (maladies, procédures, traitements), la précision varie entre 6 et 100 %. Ce type de ressources est utile pour plusieurs tâches du TAL, comme la recherche et l'extraction d'information, la simplification de textes, la lisibilité des textes de santé, systèmes de questions/réponses.

### 3 CompuTerm 2014

CompuTerm est un workshop accepté par la conférence internationale COLING (Computational Linguistics) et hébergé par cette conférence. Avec d'autres collègues, nous avons organisé ce workshop :

- Patrick Drouin, Natalia Grabar, Thierry Hamon, Kyo Kageura, editors.  
*4th International Workshop on Computational Terminology (Computerm)*  
<http://perso.limsi.fr/hamon/Computerm2014/>

Le workshop est dédié à la terminologie computationnelle (Computational terminology). Il a eu lieu à Dublin, Irlande en septembre 2014. Le workshop s'est déroulé avec succès, avec plus de 30 personnes qui sont venues y participer. Les actes de ce workshop sont accessibles librement en ligne :

<http://www.aclweb.org/anthology/W/W14/#4800>

En plus de l'animation de ce workshop, nous avons aussi présenté deux travaux :

1. Natalia Grabar, Thierry Hamon.  
*Unsupervised method for the acquisition of general language paraphrases for medical compounds.*

4th International Workshop on Computational Terminology (Computerm)  
Dublin, Ireland. 2014. Poster.

2. Ornella Wandji Tchami, Natalia Grabar.

*Towards Automatic Distinction between Specialized and Non-Specialized Occurrences of Verbs in Medical Corpora*

4th International Workshop on Computational Terminology (Computerm)  
Dublin, Ireland. 2014. Poster.

Le premier travail est lié à celui présenté pendant la conférence ICHI 2014. Quant au deuxième travail, il est aussi lié à la lisibilité des textes médicaux. Ce travail est orienté sur la syntaxe des textes, et plus particulièrement sur l'emploi des verbes. Dans ce travail, nous décrivons une méthode pour l'analyse automatique et contrastive des verbes dans les corpus médicaux. L'objectif principal est de proposer les critères pour la distinction de la spécialisation des sens de verbes et leur implémentation. La méthode proposée est basée sur l'annotation sémantique des co-occurents nominaux des verbes. Les corpus médicaux traités sont différenciés par leur niveau de spécialisation : niveau élevé (documents écrits par les chercheurs) et niveau faible (discussions des forums de discussion écrits par les patients). L'annotation sémantique de ces corpus est effectuée avec une terminologie médicale. Les résultats indiquent que les mêmes verbes employés dans ces deux corpus montrent des niveaux de spécialisation différents : ceci transparait avec les co-occurants nominaux.

Ce workshop a donné lieu au lancement d'un numéro spécial de la revue Terminology :

<http://natalia.grabar.perso.sfr.fr/TERMINO2015/>

## 4 MIE 2014

La conférence MIE (Medical Informatics in Europe) a eu lieu à Istanbul, Turquie en août 2014. Cette conférence d'informatique médicale est dédiée aux questions diverses liées aux besoins des praticiens et des patients. Entre autres, une place importante est dédiée aux aspects liés à la modélisation du domaine et aux terminologies et ontologies médicales. Lors de cette conférence, nous avons présenté plusieurs travaux :

1. Ornella Wandji Tchami, Marie-Claude L'Homme, Natalia Grabar.

*Frame Semantics-based study of verbs across medical genres.*

MIE 2014. Istanbul, Turkey.

2. Natalia Grabar, Ornella Wandji Tchami, Laura Maxim.

*Machine learning-based detection of chemical risk.*

MIE 2014. Istanbul, Turkey.

3. Maïté Boyé, Natalia Grabar, Mai Thi Tran

*Contrastive conversational analysis of language production by Alzheimer and control people.*

MIE 2014. Istanbul, Turkey.

4. Amine Abdaoui, Jérôme Azé, Sandra Bringay, Natalia Grabar, Pascal Poncelet.

*Analysis of forum messages written by health professionals and patients.*

MIE 2014. Istanbul, Turkey. Poster.

Le premier travail est effectué avec Ornella Wandji Tchami et Marie-Claude L'Homme. Il s'agit de la première étude effectuée autour de l'étude des verbes dans les textes médicaux. Ce travail a été préliminaire à celui présenté lors du workshop Computerm et a permis de produire des fondements pour de nouveaux travaux.

Le deuxième travail concerne la détection du risque chimique. Le risque chimique couvre les situations où les produits chimiques sont ou peuvent être dangereux pour la santé humaine, animale et pour l'environnement. Comme souvent il n'existe pas de preuves irréfutables sur la nocivité des substances, cela crée des situations de controverse. Les informations sur l'état des connaissances, et sur le risque chimique potentiel, se trouvent typiquement dans la littérature scientifique. Nous proposons d'exploiter la littérature scientifique afin de détecter les phrases liées au risque chimique. La tâche est considérée comme un problème de catégorisation supervisée. Selon la granularité et les types de risque chimique, nous obtenons entre 0,64 et 0,95 de F-mesure.

Le troisième travail est issu de la collaboration avec Mai Thi Tran et Maïté Boyé de l'Institut d'Orthophonie. Ce travail est dédié à une étude contrastive du langage des patients atteints de la maladie d'Alzheimer grâce à une comparaison avec les patients témoins (non atteints de la maladie d'Alzheimer). Il s'agit d'une maladie du vieillissement qui constitue la première cause de démence neurodégénérative. L'installation progressive de troubles cognitifs, émotionnels et comportementaux conduit à une perte de l'autonomie et un état de dépendance correspondant à la phase démentielle. Les troubles du langage font partie des premiers signes cliniques cognitifs de la maladie. Notre objectif est d'étudier la communication verbale de sujets atteints de la maladie d'Alzheimer à un stade débutant à modérément sévère. Une des particularités de notre démarche consiste à placer les sujets dans une situation écologique de conversation face à une personne connue. La conversation étant un phénomène complexe à analyser, nous proposons une étude contrastive des productions de cinq sujets Alzheimer et de cinq sujets témoins âgés de plus de 80 ans. Les corpus oraux des dix sujets sont transcrits et traités avec des méthodes de TAL. Plus de trente critères repartis en quatre groupes sont étudiés. Les résultats confirment que les sujets Alzheimer présentent le déficit lexico-sémantique à la phase initiale de la maladie et que leur langage conversationnel diffère, quantitativement et qualitativement, de celui des sujets témoins.

Finalement, le quatrième travail a été réalisé avec les collègues du LIRMM, Université Montpellier 2 dans le cadre du projet inter-MSH Patients' mind. L'objectif de ce travail consiste à faire une distinction automatique des documents en fonction de leur niveau de spécialisation. Il s'agit typiquement des professionnels de santé (médecins, chercheurs) et des patients. Un autre intérêt de ce travail consiste à étudier les discussions des forums de santé. L'approche proposée est basée sur la classification supervisée et exploite la modalité des écrits. Les résultats obtenus montrent que cette tâche peut être bien traitée par des méthodes automatiques.

Les publications de MIE 2014 sont indexées dans le portail PubMed.