# Note de lecture

Rubrique préparée par Denis Maurel

*Université François Rabelais Tours, LI (Laboratoire d'informatique)*

**Violaine PRINCE, Mathieu ROCHE, Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration, *Medical Information Science Reference*, 2009, 460 pages, ISBN 978-1-60566-274-9.**

Lu par **Lina F. SOUALMIA**

*Université Paris XIII, laboratoire LIM & Bio*

*L'ouvrage est un panorama des travaux récents en recherche d'information (RI) dans les domaines de la biologie et de la médecine. Ici, la RI est décrite en tant que thématique générale englobant toutes les problématiques de l'indexation et de l'extraction de connaissances (EC) à partir de textes : leur structuration, la représentation des connaissances – extraites sous forme d'ontologie – et enfin l'exploitation des ressources termino-ontologiques dans ces processus. La méthode utilisée est le traitement automatique du langage qu'il soit général ou de spécialité. Le principal message du livre est que le TAL de spécialité n'est pas une sous-partie du TAL, mais un domaine de recherche pluridisciplinaire.*

Il est avéré qu'Internet est un réservoir de connaissances, majoritairement disponibles sous forme textuelle, ce qui confère au TAL toute sa place en tant que principale méthode de RI et d'EC, par opposition aux techniques purement statistiques. Les contributions rassemblées dans le présent ouvrage recouvrent un ensemble de recherches allant de l'étude des mots en biologie et médecine, à celle de la pragmatique du discours des textes, jusqu'à l'implémentation de systèmes de RI exploitant des structures de connaissances préexistantes ou extraites. La littérature dans le domaine biomédical est très abondante. Les problématiques de RI, d'indexation de documents, de codage de dossiers patients et de comptes rendus d'hospitalisation à l'aide d'un vocabulaire contrôlé ou d'une ontologie demeurent non résolues. Ici le TAL traite les textes comme entrée du processus de RI et d'EC avec des degrés de granularité différents : le mot, le terme, la phrase et le texte, voire le corpus de textes, vecteur de discours.

Le livre se décline en six sections de vingt chapitres : une section introductive, quatre principales sections, dont trois sont dépendantes du niveau de granularité du TAL, et enfin une section conclusion. La préface de l'ouvrage, rédigée par les deux éditeurs, présente l'ensemble des chapitres de manière cohérente sur une dizaine de pages. Les chapitres ne sont pas résumés mais les choix de classification sont systématiquement justifiés.

Le chapitre I est une introduction générale de l'ouvrage sur les besoins, la nature des résultats et les réalisations de l'état de l'art. La fouille de textes y est définie comme composée de trois traitements : RI, EC et fouille de données. Le lecteur reste néanmoins sur sa faim et doit se reporter aux différentes références : pour chacun des traitements, des outils sont cités sans plus de détails sur les principes théoriques sur lesquels ils reposent. Ce chapitre liste un ensemble d'applications logicielles qu'offre le *UK National Center for Text Mining*.

Cette introduction aurait pu être écrite par d'autres chercheurs également connus en TAL dédié au biomédical : Aronson (outils lexicaux de l'UMLS), Bodenreider (ontologies et EC), Cimiano (ontologies), Friedman (codage de textes), Hersh (RI), Hearst (extraction de relations sémantiques) ou encore Zweigenbaum (pour le français). Néanmoins leurs travaux sont largement cités dans l'ouvrage.

La section I, la plus dense de l'ouvrage, est dédiée au *niveau lexical* du TAL et de la gestion des connaissances ontologiques. L'ambiguïté se situe au niveau des mots. Les neuf chapitres qui la composent peuvent se lire de manière totalement indépendante. Ils sont ordonnés en fonction de trois axes : l'exploitation de ressources préexistantes, le traitement du multilinguisme et l'enrichissement de terminologies/ontologies.

Le chapitre II étudie l'ambiguïté des variations des mots et leur projection, dans le but d'*indexation*, vers le thésaurus MeSH (suédois). Les chapitres III et IV traitent respectivement de la *catégorisation* de documents et de la *RI* en utilisant le métathésaurus UMLS (qui englobe une centaine de terminologies médicales mises en correspondance). On y explique pourquoi un moteur de recherche généraliste ne peut pas être efficace pour la littérature biomédicale. Ici l'expansion de requêtes est résolue en exploitant les différents libellés existant dans plusieurs classifications médicales. Les chapitres V et VI abordent les besoins de *multilinguisme*. Dans le chapitre V les auteurs décrivent un système d'alignement automatique pour le français entre une terminologie spécialisée/générale en fonction des similarités (différences de vocabulaire entre néophyte et spécialiste). Le chapitre VI est relatif au besoin de traduction de mots inconnus d'un langage à un autre (de l'anglais vers sept autres langues) par l'apprentissage de règles de réécriture. Les chapitres VII et VIII sont plus orientés TAL. Ils se complètent et démontrent respectivement l'ambiguïté dans les textes et dans les ontologies. Cette ambiguïté est réduite par l'enrichissement sémantique des terminologies et des ontologies effectué également par apprentissage. Remarquons, néanmoins, que l'ambiguïté de l'ontologie pourrait être résolue si elle était représentée avec un langage formel de type logique de description qui a comme avantage principal une sémantique dénotationnelle associée, écartant tous les problèmes de nature syntaxique.

Dans cette première section, les techniques d'indexation et de RI décrites sont très classiques (expansion de requêtes à l'aide de synonymes, dictionnaires, techniques d'apprentissage). De fait, elles ne sont pas spécifiques au biomédical même si de bons résultats sont obtenus. Ces méthodes peuvent être appliquées à

d'autres domaines dès lors que l'on possède des ressources termino-ontologiques de type UMLS ou bibliographiques de type PubMed. On peut regretter notamment les descriptions redondantes des ressources termino-ontologiques existantes (MeSH, UMLS, Specialist Lexicon, etc.).

Dans la section II, composée de trois chapitres, on passe à un niveau de granularité linguistique supérieur : la *phrase* comme unité élémentaire. Dans le chapitre IX les auteurs partent de résumés PubMed contenant des phrases qui expriment des informations sur la « phosphorylation des protéines ». Le système, à base de règles, considère la phrase comme décrivant un type sémantique (« procédure » ou « événement ») dans le but de découvrir des concepts. Le chapitre X est consacré à la présentation de l'outil CorTag utilisé sur le même thème de phosphorylation, et les traitements du chapitre IX sont complétés par la mise en relations des concepts découverts : les phrases et les mots permettent l'extraction de concepts et la structure lexicale d'une phrase les met en relation sémantique. À un niveau intermédiaire entre la phrase et tout le discours, un système de question-réponse est décrit dans le chapitre XI. Ici la phrase est une question (ou une réponse) qu'un médecin peut se poser en situation clinique.

La section III regroupe trois chapitres axés sur la pragmatique du discours. Le chapitre XII décrit le discours, la nature et les effets des structures des textes. Il n'est pas vraiment en relation avec le domaine biomédical mais constitue une critique des méthodes qui ne se fondent que sur des techniques statistiques. D'après l'auteure, la complexité du langage ne peut pas être cernée par des statistiques et la fouille de textes doit être enrichie avec la linguistique (le texte n'étant pas un ensemble de phrases). Les contributions de cette section sont, à mon sens, plus difficiles à appréhender selon que l'on est familier ou pas avec la pragmatique du discours. Les applications sont la RI et l'analyse des dossiers médicaux de patients (diagnostics et symptômes).

Dans la section IV, les cinq chapitres sont dédiés à la description d'applications logicielles pour la RI. Quelques-uns des outils décrits sont accessibles librement sur le Web. Ils reposent soit sur des bases de connaissances de type termino-ontologie soit sur des traitements lexicaux. Cette section est un état de l'art des logiciels ainsi que des techniques existantes que l'on peut tester et comparer entre elles.

La dernière section est un chapitre conclusion, assez intéressant, traitant de l'analyse de notes prises par des cliniciens en unités de soins intensifs. Un panorama de toutes les techniques TAL de ces trente dernières années y est dressé avec en trame de fond cette application particulière dans la pratique quotidienne des médecins ayant des contraintes de temps de traitement. Enfin, une compilation des références bibliographiques (trente-six pages) ainsi qu'un index (huit pages) closent l'ouvrage.

Les phénomènes linguistiques liés au domaine biomédical sont assez identiques à ceux du domaine général (polysémie, synonymie) avec une difficulté supplémentaire qui est l'utilisation systématique d'acronymes (textes scientifiques,

dossiers cliniques des patients). L'intervention d'un expert du domaine identifiant les besoins, mais surtout évaluant les résultats obtenus en terme d'EC, est indispensable si l'on souhaite obtenir des applications de qualité. Dans de nombreux chapitres de cet ouvrage, on a le sentiment que ce sont des techniques de TAL qui sont appliquées au biomédical et non pas des techniques pluridisciplinaires dédiées. En effet, en analysant les biographies des auteurs (incomplètes) ou encore les différentes méthodes d'évaluation, seuls les chapitres III, X, XIV et XVIII font intervenir des acteurs du domaine. L'évaluation est automatique (précision/rappel) dans les chapitres II, XI, XII et XVII. Dans le chapitre V cela peut se justifier par le fait que l'outil est destiné à un utilisateur lambda pouvant être amené à rechercher des informations sur la santé avec un axe tout particulier concernant les définitions de termes médicaux. Les chapitres VII et VIII manquent d'une partie évaluation/validation conséquente. Dans le chapitre VII les auteurs détaillent une méthode connue (enrichissement de ressources médicales avec des ressources de la langue commune) sans un réel apport original au domaine.

Du point de vue forme, on peut regretter le volume encyclopédique de l'ouvrage. Il est difficile à lire : articles sur deux colonnes, figures en noir et blanc et copies d'écrans très peu lisibles. En revanche, cette organisation des textes peut être utile si l'on veut le parcourir de manière diagonale en s'arrêtant sur un paragraphe ou un chapitre particulier.