Notes de lecture

Rubrique préparée par Denis Maurel

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

Atefeh FARZINDAR, Diana INKPEN. Natural Language Processing for Social Media. *Morgan & Claypool publishers*. 2015. 143 pages. ISBN 978-1-62705-388-4.

Lu par **Christian BOITET**

GETALP – Université Joseph Fourier

Natural Language Processing for SOCIAL MEDIA *est vraiment un ouvrage qui tient ses promesses, et qu'on ne peut que conseiller aux étudiants et chercheurs en TAL, particulièrement à ceux qui veulent travailler sur le contenu textuel au sens large*[1] *des médias*[2] *sociaux : tweets, microblogs, blogs, forums de discussion*[3]*, sites de partage de contenus, tchats, etc. (vu la rapidité de l'évolution de ce domaine, il y en a sans doute de nouveaux dont le livre ne parle pas). De plus, il est présenté de façon parfaite, avec une mise en page et une structure particulièrement claires, ce qui le rend très agréable à lire.*

L'ouvrage s'ouvre par un bref rappel sur l'explosion récente des réseaux sociaux et plus généralement des médias sociaux, qui mène à la nécessité de construire des applications capables de traiter les énormes masses de données générées[4]. Les auteurs insistent sur l'idée qu'il ne suffit plus de faire de l'analyse externe des flux de messages, en étudiant le ou les graphes formés par les acteurs (les nœuds) et les messages (les arcs), mais qu'il faut aussi faire une analyse du contenu de ces messages. Le domaine de recherche qui s'est ouvert récemment est alors celui de l'analyse sémantique de ces messages et groupes de messages. Il s'agit bien de traitements linguistiques, mais ils doivent être non classiques (les messages sont très souvent bruités, disfluents, agrammaticaux, etc.) : il doit s'agir non seulement d'extraction de contenu, nécessairement adossée à une représentation plus ou moins

1 *i.e.*, éventuellement augmenté d'émoticônes.

2 D'après les dictionnaires modernes assez complets, *une média* est un terme médical, et *un média* est un moyen de diffusion d'informations et plus généralement de communication. On ne dit pas (ou plus) « un médium, des média ».

3 Cela comprend sans doute aussi les échanges par courriels, liés à un projet précis.

4 Par exemple, plus de 500 M de tweets (ou retweets) sont envoyés chaque jour dans le monde, dont 500 k en Inde, dont 5 % environ sont multilingues (deux ou trois langues mélangées dans un même tweet).

explicite de la sémantique du domaine (le contexte de ces échanges), mais aussi de synergie avec la représentation au niveau du graphe des acteurs et des messages (simples ou formant des structures plus complexes comme des conversations). Il s'agit bien d'un domaine de recherche nouveau, qui a déjà suscité l'organisation de plusieurs ateliers en marge de conférences de l'ACL (SASM-2012, LASM-2013, LASM-2014), et de nombreuses publications.

Le chapitre 1 fait un excellent tour d'horizon des plates-formes de médias sociaux et de leurs caractéristiques (réseaux sociaux, blogs et commentaires de blogs, microblogs, forums, signets sociaux[5],wikis, journaux sociaux[6], échange de contenu[7]). Les traitements à effectuer ne sont pas seulement de l'analyse de contenu, il faut d'abord « suivre » ou « surveiller »[8], puis « capturer » ou « filtrer » ces énormes quantités de données. Le but est souvent (en e-commerce par exemple) de prédire le comportement des intervenants, et il peut aussi être d'extraire bien d'autres types d'information (analyses d'opinions, de sentiments, etc.). La recherche sur les nouvelles méthodes à utiliser foisonne (en témoignent les ateliers déjà cités, ainsi que la série d'ateliers *Making Sense of Microposts* associés aux conférences W3C depuis 2012). Le chapitre se poursuit par une présentation assez exhaustive des applications des médias sociaux, non seulement potentielles, mais aussi réellement utilisées, par des gouvernements, dans l'industrie, dans les médias et le journalisme, pour la santé, en politique, et pour la défense et la sécurité intérieure. Enfin, on trouve une présentation synthétique des défis scientifiques et techniques posés par le traitement des données des médias sociaux, et en particulier par les traitements innovants : analyse ou fouille d'opinions, extraction d'informations, résumé, traduction.

Le chapitre 2 est consacré aux outils de prétraitement des données, et à leur adaptation aux données des médias sociaux (segmenteurs, baliseurs morphosyntaxiques et lemmatiseurs, normaliseurs, reconnaisseurs d'entités nommées, analyseurs). Les auteurs attirent l'attention sur le manque de bonnes mesures pour l'évaluation des performances de ces outils, et même des plus simples *a priori*, comme les segmenteurs en mots (typographiques). Suivent deux sections très intéressantes et informatives sur les boîtes à outils[9] existant pour l'anglais, et sur l'ajout (nécessaire) de l'aspect multilingue, avec des informations très détaillées sur ce qui existe pour l'identification de langues et encore plus pour l'identification de dialectes.

Le chapitre 3 est, comme le disent les auteurs, le cœur de l'ouvrage. Il présente les méthodes utilisées dans diverses applications pour effectuer l'analyse sémantique

5 *Social bookmarks.*

6 *Social news.*

7 *Media sharing.*

8 *To monitor.*

9 *Toolkits.*

des textes des réseaux sociaux, pour analyser leurs flux[10], pour extraire de l'information, et pour classer les textes. Toute la première section est consacrée à la géolocalisation. Il est intéressant d'apprendre que les méthodes utilisées sur nos PC, portables et tablettes ne marchent pas parfaitement, et qu'elles doivent, si possible, être complétées par des méthodes fondées sur l'analyse du contenu textuel. La section suivante présente les techniques liées aux entités nommées : repérage, lien avec une base de connaissances, et désambiguïsation. Viennent ensuite trois sections beaucoup plus longues et détaillées. La quatrième section est dédiée aux méthodes de fouille d'opinions et d'analyse de sentiments, comprenant l'analyse de l'humeur[11] et des émotions. La section suivante, consacrée à la détection d'événements et de thèmes d'intérêt, est une des plus longues. Il faut en effet distinguer les événements spécifiés, connus à l'avance, et les événements non spécifiés. S'y ajoute la reconnaissance de la survenue de situations d'urgence[12]. La section d'après, sur le résumé, est très intéressante, car elle montre que les médias sociaux suscitent différents types de résumés : le résumé de mise à jour, qui doit produire uniquement les informations nouvelles concernant un thème donné, le résumé sur l'activité d'un réseau social, le résumé d'un événement et le résumé sur l'opinion d'un ensemble de personnes. On parle enfin des mesures utilisées pour évaluer les résumés, avec quelques exemples tirés des mesures utilisées en TA… mais qui sont encore moins pertinentes, à notre avis, pour le résumé que pour la TA. *A contrario*, on voit bien la nécessité d'en trouver de meilleures. Ensuite, la septième section présente les possibilités d'utilisation de la TA pour les médias sociaux. En bref, ça peut très bien fonctionner si les tweets sont propres et homogènes, comme dans le cas où ils sont émis par des agences gouvernementales, au Canada par exemple, et pas du tout s'il s'agit de tweets très bruités. On termine par une sous-section dédiée à la TA dans le cas de l'arabe, où il s'agit en fait de traduire entre une douzaine de dialectes, et par une autre sous-section, très courte, sur les mesures d'évaluation de résultats de TA.

Le chapitre 4 montre des applications de haut niveau des méthodes présentées au chapitre précédent. Il s'agit d'applications dans les domaines de la santé, des finances, de la politique (prédiction de votes), du suivi de réseaux sociaux, de la sécurité et défense, en profilage d'utilisateurs fondé sur le contenu des messages, en visualisation d'informations tirées des médias sociaux dans des situations de crise, en réponse à des catastrophes (tremblements de terre, typhons…), et pour le divertissement.

Le chapitre 5 introduit différents aspects complémentaires, souvent considérés comme secondaires, mais qui peuvent s'avérer cruciaux, comme la collecte et l'annotation de ressources (en particulier pour les méthodes empiriques, fondées sur l'apprentissage automatique), le respect du caractère privé des données (d'où la nécessité d'anonymiser les données), les techniques d'élimination des spams, ainsi que les méthodes et les bancs d'essai pour l'évaluation.

---

10 *Social media analytics*

11 *Mood.*

12 *Emergency situation awareness.*

Enfin, le chapitre 6 résume l'ouvrage, et se termine en détaillant le très grand « potentiel de recherche » en TAL, présenté en neuf groupes de questions, soit vingt-quatre au total… et il y en a sûrement d'autres. Enfin, une annexe présente un système réalisé par la firme d'A. Fazindar (qui enseigne aussi à l'université de Montréal), un glossaire très bien venu, et une très riche bibliographie (trente-trois pages).

On ne peut que conseiller aux « Talistes » de lire ce livre !

Iryna GUREVYCH, Judith ECKLE-KOHLER, Michael MATUSCHEK. Linked Lexical Knowledge Bases – Foundations and Applications. *Morgan & Claypool publishers*. 2016. 121 pages. ISBN 978-1-62705-974-9.

Lu par **Franck SAJOUS**

CLLE-ERSS, CNRS, Toulouse

« La connaissance sémantique lexicale est cruciale pour la plupart des tâches de TAL ». *Un corollaire de cette première phrase de la préface est qu'à traitement constant, de meilleures bases de connaissances permettront d'atteindre de meilleures performances. Motivé par la création de meilleures ressources, l'ouvrage traite des méthodes d'alignement qui permettent de « lier » plusieurs ressources lexicales, complémentaires en termes de couverture ou par la nature différente des données décrites. Outre les techniques d'alignement, le livre, conçu pour les étudiants et les chercheurs en TAL, entend présenter plus largement les fondements des bases de connaissances lexicales liées, leurs applications (la désambiguïsation d'unités textuelles), les aspects multilingues, les outils et les interfaces d'exploration de ces bases.*

L'ouvrage est organisé en huit chapitres que l'on peut regrouper comme suit : les deux premiers introduisent les bases de connaissances lexicales *(Lexical Knowledge Bases*, ou LKB) et les bases de connaissances lexicales liées (LLKB, pour *Linked LKB*). Les chapitres centraux présentent les techniques d'alignement et d'utilisation de ces ressources. Les deux derniers chapitres, plus périphériques, traitent du multilinguisme, des outils et des interfaces.

Le premier chapitre introduit les LKB et illustre la définition qui en est donnée par un panorama de bases existantes, classées selon leur mode de construction (*« par des experts* vs *collaborativement »* – comprendre : par des experts *vs* crowdsourcées), puis selon le type d'information décrite. Les auteurs choisissent de centrer leur étude sur les bases fournissant des informations ancrées au niveau des sens, laissant de côté les aspects morphologiques. Sont également écartés les dictionnaires électroniques, conçus pour une utilisation humaine, dont la structure est jugée trop lâche pour une utilisation en TAL. Poursuivant la discussion sur le rôle des formats et des standards, le chapitre se conclut par une présentation de la norme dictionnairique Lexical Markup Framework LMF et quelques lignes sur le Web sémantique, dont il ne sera plus question par la suite.

Le deuxième chapitre présente la notion de LKB liées, définissant tout d'abord ce qu'est un alignement de sens, puis l'alignement de ressources lexicales.

L'alignement de LKB est motivé par la production de ressources plus riches (en combinant des informations hétérogènes comme celles de WordNet et FrameNet), multilingues (par exemple en alignant plusieurs WordNets, technique également utilisée pour la construction d'un WordNet donné en profitant de la structure des autres réseaux), ou dotées d'une meilleure couverture. Dans cette dernière catégorie, deux LLKB sont présentées : UBY et BabelNet. La première, développée à l'université de Darmstadt, combine onze ressources de nature différente. BabelNet, composé également de l'alignement d'un grand nombre de ressources (initialement WordNet et Wikipédia), se concentre sur les définitions et les liens de traduction.

Le chapitre 3 présente différentes familles d'algorithmes permettant d'apparier les sens équivalents de différentes ressources, ainsi que les métriques traditionnelles d'évaluation des alignements produits. Ces méthodes d'alignement sont réparties en trois catégories. Les premières sont fondées sur le calcul de similarité entre gloses, ainsi qu'une version modifiée du PageRank Personnalisé (PPR). La deuxième famille de méthodes exploite la similarité structurelle des ressources, *i.e.* leur structure de graphe sous-jacente. Parmi ces méthodes, on trouve celles qui alignent des ressources à la hiérarchie des catégories de Wikipédia et d'autres, fondées sur les algorithmes de recherche des plus courts chemins. Une section est enfin dédiée aux méthodes hybrides combinant les deux familles d'alignements précédentes.

Le chapitre 4 décrit la tâche de *désambiguïsation* d'unités textuelles relativement à une ressource de référence. Cette tâche prise au sens large comprend la désambiguïsation de sens en contexte, le repérage d'entités nommées ou de structures prédicatives, l'attribution de rôles sémantiques. Certaines de ces méthodes d'appariement entre unités textuelles et ressources sont les mêmes que celles présentées au chapitre 3 (utilisées pour apparier les entrées de deux ressources), D'autres sont fondées sur des algorithmes d'apprentissage : supervisés ou semi-supervisés à partir de corpus annotés, ou non supervisés, par induction de sens.

Le cinquième chapitre intitulé « Méthodes de désambiguïsation avancées » recense les travaux les plus récents sur les tâches de désambiguïsation d'unités textuelles et la complétion automatique de bases de connaissances. Bien que la première partie s'intitule « Construction automatique de bases de connaissances », ce sont en réalité les principes d'extension de bases existantes qui y sont énoncés. Une méthode d'apprentissage semi-supervisé est présentée : la « supervision distante ». Elle consiste à désambiguïser les unités textuelles connues de la base (*i.e.* aligner automatiquement ces unités avec les entrées de la ressource), puis à transférer les traits de ces entrées de la ressource vers le texte. Le corpus enrichi de ces nouvelles annotations servira de données d'entraînement à de nouveaux algorithmes d'apprentissage, qui pourront à leur tour enrichir la base. Des modèles vectoriels continus connus comme les *word embeddings* sont appliqués aux LKB. Appris à partir de bases de connaissances et/ou de corpus, ces modèles donnent des représentations (denses et de faibles dimensions) d'entités et de relations (concepts et relations ontologiques dans le cas d'ontologies, sens et relations lexico-sémantiques pour les bases lexicales), et sont utilisés pour inférer de nouvelles connaissances.

Le chapitre 6 se compose de deux volets. Le premier, très court, signale la possibilité qu'offrent les LLKB d'étendre la notion de calcul de similarité sémantique au niveau multilingue. Le second traite de traduction assistée par ordinateur (par opposition à traduction automatique), réalisée par des traducteurs humains tirant bénéfice de l'utilisation de LLKB multilingues.

Le dernier chapitre (avant conclusion) traite des interfaces d'exploration des ressources (dont l'intérêt apparaît après lecture de la section sur la traduction manuelle), de celles permettant la maintenance des bases, et des API d'interrogation.

Commentaire

L'objectif du livre est ambitieux et l'on peut d'emblée questionner l'adéquation entre un ouvrage sous-titré « Fondements et applications » et la collection, conçue pour accueillir des monographies synthétiques offrant un survol des travaux récents dans une thématique donnée. De fait, l'ouvrage hésite entre manuel et état de l'art qui balaie large, de manière superficielle : le résultat s'apparente souvent à une énumération de travaux (31 pages de bibliographie pour 89 pages de texte) placés sur le même plan, insuffisamment exemplifiés et commentés (lacune étonnante en regard de la visée didactique annoncée par les auteurs), dont les pertinences relatives n'apparaissent pas toujours nettement. Or on voit mal comment un même lecteur, dont « *seulement des connaissances limitées en TAL sont attendues* », pourrait tout à la fois découvrir avec profit des notions de base, parfois non spécifiques au sujet traité, comme les mesures d'association et de similarité (Dice, Jaccard, tf-idf, cosinus) ou les métriques d'évaluation (précision, rappel, F-mesure), et la section sur le PPR, décrite en moins d'une page, dont la lecture impose une familiarité préalable avec les marches aléatoires. De la même manière, le lecteur dépourvu d'une solide connaissance en UML et ignorant la notion, non introduite, de métamodèle, pourra rester perplexe devant le long développement sur LMF, au premier chapitre.

On pourra en revanche apprécier l'organisation claire de l'ouvrage, au moins pour la première partie. Chaque chapitre débute par une présentation concise de sa problématique, se conclut par une synthèse, et l'articulation des chapitres entre eux est explicite. Le cinquième chapitre, moins bien introduit, prend tout son intérêt une fois que l'on a compris que le processus de désambiguïsation textuelle et celui de complétion des LKB sont intiment liés, dans un processus cyclique.

On regrettera finalement l'absence de discussion sur des questions que l'on pressent problématiques. Que penser de l'affirmation initiale sur la nécessité pour les applications de TAL de disposer de connaissances lexicales et du constat (p. 65-66), suspendu, que le bénéfice de l'utilisation des LKB dans les systèmes de TAL n'est souvent pas clairement perceptible ? Concernant les formats standard, on peut se demander en quoi la notion de métamodèles comme LMF, dont chacun peut donner une instanciation différente va dans le sens de l'interopérabilité, et en quoi UBY-LMF, une implémentation particulière, présente un avantage sur d'autres. En quoi (et comment) UBY parvient à aligner onze ressources, alors que les auteurs pointent la difficulté d'effectuer un alignement composé de plus de deux ressources ? Quelles sont les conditions de réussite de méthodes telles que la supervision distante (chapitre 5), qui apprend automatiquement sur un corpus annoté

automatiquement ? Le lecteur aurait pu s'attendre à bénéficier de l'éclairage que le recul des auteurs, sur de tels sujets, pourrait lui apporter.