UNIVERSITÉ DE LILLE 1
Formation Informatique de Lille (FIL)



RAPPORT DE STAGE

Extraction d'informations et analyse sémantique dans le domaine du traitement automatique du langage naturel

Auteur :
Nicolas DUBRULLE

Tuteur de stage :
Julien CARTIGNY

Tuteur en entreprise :
Romain MORVAN



3 septembre 2017

Université de Lille 1
Formation Informatique de Lille (FIL)

Résumé

Extraction d'informations et analyse sémantique dans le domaine du traitement automatique du langage naturel

par Nicolas DUBRULLE

L'extraction de mots-clés étant l'une des tâches les plus importantes lorsque l'on travaille avec du texte, divers algorithmes ont été développés pour répondre à ce problème. Cependant, étant donné un texte (ou un corpus), ils nécessitent généralement des réglages fins pour être précis.

La désambiguïsation lexicale consiste quant à elle en la détermination du sens d'un mot dans une phrase lorsque que ce dernier peut avoir plusieurs sens possibles. Il s'agit d'un domaine de recherche actuellement non encore résolu.

Mon stage étant de type R&D, j'expliquerai en première partie de ce rapport ma prise en main avec la plateforme Languages.ai. Je détaillerai ensuite comment je m'y suis pris pour extraire des phrases ou mots-clés d'articles quelconques à partir de leurs URL à des fins de classifications et recommandations, pour ensuite présenter de quelle manière j'ai essayé d'obtenir les meilleurs sens et traductions de mots à partir de leur contexte dans le but de fournir des résultats plus pertinents aux utilisateurs.

Remerciements

Tout d'abord, je remercie M. Romain MORVAN et M. Alexandre CHAINER, respectivement CEO et COO de Languages.ai, pour m'avoir accueilli durant six mois dans leurs locaux.

Je remercie une fois de plus M. Romain MORVAN, qui fut également mon tuteur au sein de Languages.ai, pour m'avoir conseillé durant toute cette période.

Je remercie l'ensemble de l'équipe pédagogique de l'université de Lille 1 pour m'avoir enseigné les connaissances théoriques nécessaires à la réalisation de ce stage, et plus particulièrement M. CARTIGNY, pour m'avoir suivi durant ce stage.

Enfin je tiens à remercier M. Gilles VANWORMHOUDT, chercheur, M. Laurent BETTI, ingénieur et M. Quentin DELVALLET, stagiaire, pour l'aide et les conseils qu'ils ont tous pu me fournir durant la réalisation de ce projet.

Table des matières

Résumé **i**

Remerciements **ii**

Introduction **v**

1 Présentation de l'entreprise **1**

2 Prise en main de la plateforme **2**
2.1 Back-end . . . 2
2.2 Front-end . . . 2
2.3 Mini-projet : *Bookmarks* . . . 2

3 Extraction de mots-clés depuis un document **4**
3.1 Extraction du texte brut depuis un article web . . . 4
3.1.1 Récupération du contenu pertinent . . . 4
3.1.2 Nettoyage . . . 4
3.1.3 Persistance . . . 5
3.2 Extraction de mots-clés . . . 5
3.2.1 Terminologie . . . 5
3.2.2 Algorithmes utilisés . . . 5
3.2.3 Unification . . . 6
3.3 Recherche des paramètres optimaux . . . 8
3.4 Résultats . . . 9
3.5 Benchmark . . . 9

4 Analyse sémantique et désambiguïsation lexicale **11**
4.1 Analyse de Wikipédia . . . 11
4.2 Récupération des traductions . . . 12
4.3 Méthodes de désambiguïsation . . . 13
4.3.1 Avec Wordnet . . . 13
4.3.2 Avec word2Vec et l'API de traduction . . . 13
4.4 Résultats . . . 14

Conclusion **16**

A Schéma de l'algorithme *keywords* **17**

B Performance de l'algorithme $keywords_{W2V}$ **18**

C Capture d'écran du projet *Bookmarks* **19**

D Précision de l'algorithme $keywords_{W2V}$ **20**

Bibliographie **21**

Introduction

L'extraction de mots-clés est l'une des tâches les plus importantes lorsque l'on travaille avec du texte. Les lecteurs profitent des mots-clés parce qu'ils peuvent juger plus rapidement si le texte vaut la peine d'être lu ; les créateurs de sites web tirent avantage des mots-clés car ils peuvent regrouper les contenus similaires par sujets ; et les développeurs bénéficient des mots-clés car ces derniers réduisent la dimensionnalité du texte à ses caractéristiques les plus importantes.

Bien que particulièrement complexe, l'extraction de phrases ou mots-clés est primordiale lorsque l'on souhaite classifier un texte, effectuer des recherches sur un ensemble de textes (i.e. un corpus), ou, dans le cas d'une approche basée sur le contenu, de pouvoir proposer des recommandations de textes à partir d'un profil d'utilisateur. Aujourd'hui, de nombreux algorithmes existent et permettent d'extraire des mots-clés à partir d'un texte donné. Cependant, en les observants un à un, on peut rapidement constater qu'ils fournissent soit **(a)** des résultats raisonnablement précis, mais trop peu nombreux pour permettre des approches de type *recommandations* ; ou **(b)** trop de résultats, résultant en un nuage de termes peu précis dont on ne sait finalement pas quoi faire.

La solution que j'ai apporté afin de palier à ces problèmes part du principe que différents algorithmes peuvent se *soutenir* entre-eux. On peut y voir une sorte d'*ensemble learning*, mais sans la notion de *machine learning* (il n'y a pas d'apprentissage). Ainsi, si plusieurs algorithmes décident qu'un terme est plausible (il a été extrait) et pertinent (son score n'est pas mauvais), il a de fortes chances d'être valide.

La désambiguïsation lexicale consiste en la détermination du sens d'un mot dans une phrase alors que ce dernier peut avoir plusieurs sens possibles. Cette tâche semble aisée pour l'Homme, mais est d'une grande complexité pour la machine. Aujourd'hui, aucune méthode ne permet de connaître de façon certaine le sens d'un mot dans une phrase. Les recherches sont actives dans ce domaine, et les systèmes les plus précis affichent un score d'environ 82-83 % (NAVIGLI, 2012).

De manière à approcher au mieux le sens des mots ambigus d'une phrase, je propose ici une méthode basée d'une part sur le modèle Word2Vec (MIKOLOV et al., 2013) afin d'effectuer des tests de similarité entre les définitions possibles des mots ambigus et leurs voisins dans leur phrase d'origine ; et d'autre part, sur Wikipédia, afin de trouver les natures des mots les plus probables dans les cas où il y-a également ambiguïté sur la nature d'un mot.

Bibliographie

GONZALEZ, Aitor, German RIGAU et Mauro CASTILLO (2012). « A Graph-Based Method to Improve WordNet Domains ». In : *Computational Linguistics and Intelligent Text Processing : 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I*. Sous la dir. d'Alexander GELBUKH. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 17–28. ISBN : 978-3-642-28604-9. DOI : 10.1007/978-3-642-28604-9_2. URL : https://doi.org/10.1007/978-3-642-28604-9_2.

GONZALEZ-AGIRRE, Aitor, Mauro CASTILLO et German RIGAU (2012). « A proposal for improving WordNet Domains ». Anglais. In : *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Sous la dir. de Nicoletta Calzolari (Conference CHAIR) et al. Istanbul, Turkey : European Language Resources Association (ELRA). ISBN : 978-2-9517408-7-7.

HARRIS, Zellig S. (1954). « Distributional Structure ». In : WORD 10.2-3, p. 146–162. DOI : 10.1080/00437956.1954.11659520. eprint : http://dx.doi.org/10.1080/00437956.1954.11659520. URL : http://dx.doi.org/10.1080/00437956.1954.11659520.

JONES, KAREN SPARCK (1972). « A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL ». In : *Journal of Documentation* 28.1, p. 11–21. DOI : 10.1108/eb026526. eprint : https://doi.org/10.1108/eb026526. URL : https://doi.org/10.1108/eb026526.

KUSNER, M. J. et al. (2015). « From Word Embeddings To Document Distances ». In : *ICML*.

LOPER, Edward et Steven BIRD (2002). « NLTK : The Natural Language Toolkit ». In : *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. ETMTNLP '02. Philadelphia, Pennsylvania : Association for Computational Linguistics, p. 63–70. DOI : 10.3115/1118108.1118117. URL : http://dx.doi.org/10.3115/1118108.1118117.

LUHN, H. P. (1957). « A Statistical Approach to Mechanized Encoding and Searching of Literary Information ». In : *IBM Journal of Research and Development* 1.4, p. 309–317. ISSN : 0018-8646. DOI : 10.1147/rd.14.0309.

MARUJO, Luis et al. (2012). « Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization ». In : *Proceedings of LREC 2012*. Istanbul, Turkey : ELRA.

MEDELYAN, Olena, Eibe FRANK et Ian H. WITTEN (2009). « Human-competitive Tagging Using Automatic Keyphrase Extraction ». In : *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3 - Volume 3*. EMNLP '09. Singapore : Association for Computational Linguistics, p. 1318–1327. ISBN : 978-1-932432-63-3. URL : http://dl.acm.org/citation.cfm?id=1699648.1699678.

MIHALCEA, Rada et Paul TARAU (2004). « TextRank : Bringing Order into Texts ». In : *Proceedings of EMNLP 2004*. Sous la dir. de Dekang LIN et Dekai WU. Barcelona, Spain : Association for Computational Linguistics, p. 404–411. URL : http://www.aclweb.org/anthology/W04-3252.

MIKOLOV, Tomas et al. (2013). « Efficient Estimation of Word Representations in Vector Space ». In : *CoRR* abs/1301.3781. URL : http://arxiv.org/abs/1301.3781.

MILLER, George A. (1995). « WordNet : A Lexical Database for English ». In : *Commun. ACM* 38.11, p. 39–41. ISSN : 0001-0782. DOI : 10.1145/219717.219748. URL : http://doi.acm.org/10.1145/219717.219748.

NAVIGLI, Roberto (2012). « A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches ». In : *SOFSEM 2012 : Theory and Practice of Computer Science : 38th Conference on Current Trends in Theory and Practice of Computer Science, Špindlerův Mlýn, Czech Republic, January 21-27, 2012. Proceedings*. Sous la dir. de Mária BIELIKOVÁ et al. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 115–129. ISBN : 978-3-642-27660-6. DOI : 10.1007/978-3-642-27660-6_10. URL : https://doi.org/10.1007/978-3-642-27660-6_10.

NGUYEN, Thuy Dung et Min-Yen KAN (2007). « Key phrase Extraction in Scientific Publications ». In : *Proceeding of International Conference on Asian Digital Libraries*, p. 317–326.

PAGE, Lawrence et al. (1999). *The PageRank Citation Ranking : Bringing Order to the Web.* Technical Report 1999-66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab. URL : http://ilpubs.stanford.edu:8090/422/.

PEDREGOSA, F. et al. (2011). « Scikit-learn : Machine Learning in Python ». In : *Journal of Machine Learning Research* 12, p. 2825–2830.

ŘEHŮŘEK, Radim et Petr SOJKA (2010). « Software Framework for Topic Modelling with Large Corpora ». English. In : *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta : ELRA, p. 45–50.

ROSE, Stuart et al. (2010). « Automatic Keyword Extraction from Individual Documents ». In : *Text Mining. Applications and Theory*. Sous la dir. de Michael W. BERRY et Jacob KOGAN. John Wiley et Sons, Ltd, p. 1–20. ISBN : 9780470689646. DOI : 10.1002/9780470689646.ch1. URL : http://dx.doi.org/10.1002/9780470689646.ch1.

Liste des acronymes

AJAX	asynchronous Javascript and XML
AMT	Amazon's Mechanical Turk
API	Application Programming Interface
BoW	Bag of Words
CSS	Cascading Style Sheets
DOM	Document Object Model
FAO	Food and Agriculture Organization of the United Nations
HTML	HyperText Markup Language
JSON	JavaScript Object Notation
NE	Named Entity
NLP	Natural Language Processing
PoS	Part of Speech
RAKE	Rapid Automatic Keyword Extraction
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System
SQL	Structured Query Language
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator
W2V	Word2Vec
WMD	Word Mover's Distance
XML	Extensible Markup Language
XWND	Extended Wordnet Domains