le cnam

Sujet de stage inter-équipes CEDRIC

Extraction d'entités nommées visant à l'amélioration d'un système de recommandation

Cédric du Mouza, Quentin Grossetti (ISID) &
Nicolas Travers [1] (Vertigo)

Mots-clés : Entités nommées, NER, Systèmes de recommandation, Big Data, Scalability, Microblog

29 janvier 2018

1. mailto:nicolas.travers@cnam.fr

1 Contexte

L'extraction d'entités nommées à partir d'un document est une tâche ardue clairement identifiée depuis plusieurs années par la communauté de systèmes d'information. L'objectif consiste à extraire des objets textuels facilement catégorisables (Personne, Endroit, Raison Sociale, etc..) à partir d'un texte. De nombreuses évolutions technologiques ont permis d'améliorer drastiquement les performances de ces systèmes. On a ainsi pu voir ces systèmes qui reposaient sur des bases de connaissances expertes [1] devenir progressivement moins supervisés [2]. Les meilleures performances sont aujourd'hui obtenues par des systèmes non-supervisés utilisant des réseaux de neurones (deep-learning) [3]. Si ces systèmes sont efficaces, ils montrent leurs limites lorsque le contenu à analyser est multilingue, sans domaine précis et avec un contexte très léger tel que les plateformes de micro-blogging [4], [5]. Utiliser le réseau social des individus, ainsi que le timing des publications semble être un axe porteur pour réussir à identifier les utilisateurs parlant d'un même accident d'avion mais séparés par des dizaines de milliers de kilomètres.

SimGraph [6, 7, 8] est un système de recommandation reposant sur un graphe d'utilisateurs inter-connectés, ainsi que l'historique des publications propagé sur ce graphe. Notre système propose donc un méta-graphe capable d'améliorer la qualité des recommandations tout en restant efficace grâce à une réduction drastique et pertinente de l'espace de recherche. Toutefois, ce méta-graphe repose en partie sur les historiques des publications, qui de fait s'intéresse au chemin d'un tweet sur le graphe. Intégrer une notion d'extraction d'entité entre les publications permettrait de rapprocher des tweets sémantiquement et améliorer la pertinence des recommandations.

L'enjeu du stage est donc double, dans un premier temps tenter de raffiner un modèle existant d'extraction d'entités nommés afin de l'adapter à la structure d'une plateforme de microblogging (Twitter). Dans un second temps, une fois ces entités efficacement extraites, il s'agira d'évaluer la pertinence de leur utilisation dans l'amélioration de notre système de recommandation. Cette approche est originale dans le fait de produite une similarité sémantique pour un méta-graphe avec un modèle de propagation pour du micro-blogging.

2 Sujet de stage

Le stage s'appuie sur un large jeu de données déjà collecté issu de Twitter [9] comportant plus de 3 milliards de messages, 2 182 867 utilisateurs et 325 451 980 arcs. À partir de ce jeu, il s'agira de choisir une méthode qui permette d'extraire efficacement les entités nommées telle que la méthode de Stanford par exemple [10] et de l'adapter au contexte d'une plateforme de micro-blogging. Il s'agira ensuite d'évaluer la pertinence de l'utilisation de celles-ci afin d'améliorer le système de recommandation de notre équipe qui a déjà donné des résultats satisfaisants en dépassant les performances du système de recommandation actuellement utilisé par Twitter [11]. Ce stage pourra déboucher sur une offre de thèse.

Références

[1] Jun'ichi Kazama and Kentaro Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707, 2007.

[2] David Nadeau. *Semi-supervised named entity recognition : learning to recognize 100 entity types with little supervision.* PhD thesis, University of Ottawa, 2007.

[3] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv :1603.01360*, 2016.

[4] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2) :32–49, 2015.

[5] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets : an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[6] Q. Grossetti, C. Constantin, C. du Mouza, and N. Travers. An Homophily-based Approach for Fast Post Recommendation in Microblogging Systems. In *EDBT*, pages 1–12, Vienna, Austria, March 2018.

[7] Q. Grossetti, C. du Mouza, and N. Travers. Tweet, Retweet et Follower : que recommander et à qui ? In *Atelier interdisciplinaire sur les systèmes de recommandation*, pages 1–6, Paris, France, May 2017.

[8] Q. Grossetti, C. du Mouza, and N. Travers. Enhance micro-blogging recommendations of posts with an homophily-based graph. In *BDA'17*, pages 1–10, Nancy, France, November 2017.

[9] Camelia Constantin, Ryadh Dahimene, Quentin Grossetti, and Cédric Du Mouza. Finding Users of Interest in Micro-blogging Systems. In *International Conference on Extending Database Technology, EDBT*, Bordeaux, France, March 2016.

[10] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[11] Aneesh Sharma, Jerry Jiang, Praveen Bommannavar, Brian Larson, and Jimmy Lin. GraphJet : Real-time Content Recommendations at Twitter. *Proc. VLDB Endow.*, 9(13) :1281–1292, 2016.