LIS UMR CNRS 7020, Aix-Marseille Université, Marseille, France.

Proposition de stage de Master financé
Funded Master's internship proposal

Fouille de textes par Machines Relationnelles Profondes
Text Mining with Deep Relational Machines

Direction/*Supervision* : Bernard Espinasse, Sébastien Fournier, Adrian Chifu
5 mois/*months* – 550 euros par/by mois/*month*
Contact : bernard.espinasse@lis-lab.fr

La fouille de textes (Text-Mining) utilise de plus en plus de techniques issues de l'apprentissage profond pour des tâches de traitement automatique des langues (TAL) de très bas niveau comme l'extraction d'information (entités nommées ou relations) ou des tâches de plus haut niveau comme la simplification de textes, le résumé automatique.

Ces techniques d'apprentissage profond utilisant diverses architectures de réseaux de neurones (CNN, RCC, LSTM, …) permettent d'atteindre des performances intéressantes. Ces performances peuvent être améliorées par l'intégration de caractéristiques linguistiques comme les dépendances syntaxiques (Espinasse et al., 2019). Cependant les performances de ces techniques relevant de l'apprentissage profond semblent plafonner. D'autres techniques de TAL, symboliques tirent mieux partie de la linguistique, de ressources sémantiques externes (ontologies), avec notamment l'usage d'un apprentissage relationnel comme dans (Lima et al., 2019) (Verbeke et al., 2014). Pour outrepasser les limites des techniques par apprentissage profond, leur combinaison avec ces techniques symboliques s'avère judicieuse.

Ce stage de Master de 5 mois s'intéressera plus particulièrement à de nouvelles solutions logicielles hybrides notamment les Machines Relationnelles Profondes (*Deep Relational Machines* – DRM – Dash et al, 2018) combinant ces deux types de techniques. Les DRM sont des réseaux neuronaux profonds avec des fonctions booléennes de premier ordre à la couche d'entrée. Ils offrent un moyen d'intégrer des connaissances complexes dans des réseaux profonds, au travers de caractéristiques relationnelles. Dans la formulation originale de Huma Lodhi (Lodhi, 2013), ces caractéristiques relationnelles sont sélectionnées par un moteur de Programmation Logique Inductive (ILP) utilisant la connaissance de domaine encodée sous forme de programmes logiques. Les DRM nous apparaissent constituer un axe de recherche pertinent novateur, motivé par le fait que des travaux très récents sur les DRM ont donné des résultats très encourageant. Ainsi les études de (Dash et al., 2018) et (Vig et al., 2018) évaluées sur plusieurs ensembles de données de référence ont permis d'améliorer considérablement la performance prédictive grâce à l'intégration des connaissances du domaine.

Après avoir mieux cerné l'intérêt et les limites de ces nouvelles approches hybrides à base de Machines Relationnelles Profondes pour la fouille de texte, leur mise en oeuvre sera faite sur une tâche spécifique comme l'extraction et la classification de relations.

Text mining increasingly uses deep learning techniques for very low-level automatic language processing (NLP) tasks such as information extraction (named entities or relationships) or higher-level tasks such as text simplification, automatic summarization.

These deep learning techniques using various neural network architectures (CNN, RCC, LSTM, ...) allow to reach interesting performances. These performances can be improved by integrating linguistic features such as syntactic dependencies (Espinasse et al., 2019). However, the performance of these deep learning techniques seems to be levelling off. Other symbolic NLT techniques make better use of linguistics and external semantic resources (ontologies), including the use of relational learning as in (Lima et al., 2019) (Verbeke et al., 2014). In order to go beyond the limits of deep learning techniques, their combination with these symbolic techniques proves to be judicious.

This Master's internship of 5 month will focus on new hybrid software solutions such as Deep Relational Machines (DRM - Dash et al, 2018) combining these two types of techniques. DRMs are deep neural networks with first-order Boolean functions at the input layer. They offer a way to integrate complex knowledge into deep networks through relational features. In the original formulation of Huma Lodhi (Lodhi, 2013), these relational characteristics are selected by an Inductive Logic Programming (ILP) engine using domain knowledge encoded in the form of logic programs. DRMs appear to us to be an innovative and relevant line of research, motivated by the fact that very recent work on DRMs has yielded very encouraging results. For example, studies by (Dash et al., 2018) and (Vig et al., 2018) evaluated on several reference data sets have led to a significant improvement in predictive performance through the integration of domain knowledge.

After having better identified the interest and limitations of these new hybrid approaches based on Deep Relational Machines for text mining, their implementation will be done on a specific task such as the extraction and classification of relations.

Références

(Dash et al., 2018) Tirtharaj Dash, Ashwin Srinivasan, Lovekesh Vig, Oghenejokpeme I. Orhobor, Ross D. King: Large-Scale Assessment of Deep Relational Machines. ILP 2018: 22-37

(Espinasse et al., 2019), B. Espinasse, S. Fournier, A. Chifu, G. Guibon, R. Azcurra, V. Mace, « On the Use of Dependencies in Relation Classification of Text with Deep Learning », 20 International Conference on Computational Linguistics and Intelligent Text Processing, long paper, CICLing 2019, La Rochelle, France, April 7 to 13, 2019.

(Lima et al., 2019) R. Lima, S. B. Espinasse, F. Freitas, « The Impact of Semantic Linguistic Features in Relation Extraction: A Logical Relational Learning Approach », Recent

Advances in Natural Language Processing, long paper, RANLP 2019, Varna, Bulgaria, September 2-4, 2019.

(Lodhi, 2013) Lodhi, H.: Deep Relational Machines. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) ICONIP (2013). LNCS, vol. 8227, pp. 212–219. Springer, Heidelberg (2013). https://doi.org/10.1007/978- 3-642-42042-9 27

(Verbeke et al., 2014) Mathias Verbeke, Paolo Frasconi, Kurt De Grave, Fabrizio Costa, Luc De Raedt (2014) kLogNLP: Graph Kernel–based Relational Learning of Natural Language, Conference: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, January 2014 - DOI: 10.3115/v1/P14-5015

(Vig et al., 2018) Vig L., Srinivasan A., Bain M., Verma A. (2018) An Investigation into the Role of Domain-Knowledge on the Use of Embeddings. In: Lachiche N., Vrain C. (eds) Inductive Logic Programming. ILP 2017. Lecture Notes in Computer Science, vol 10759. Springer.

—