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

Master 2 internship topic

Representation Learning for Text Mining Tasks

Supervision:
Bernard Espinasse (AMU/LIS) and Rinaldo Lima (UFRPE)

Text mining increasingly uses Deep Learning techniques for Natural Language Processing (NLP) tasks such as information extraction (named entity recognition and relation extraction) or higher-level tasks such as text simplification, and automatic text summarization.

Such deep learning techniques are based on many neural network architectures, including Convolutional (CNN), Recurrent (RNN), and Long Short Term Memory Neural Networks (LSTM), and more recently Transformers with BERT (Bidirectional Encoder Representations from Transformers), that allow to reach impressive results in many NLP task.

However, as demonstrated by recent studies such performance can be improved by mainly integrating linguistic features such as syntactic dependencies (Espinasse et al., 2019). In addition, other symbolic NLP-based 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 seems to be beneficial.

This research work will address recent advances in representation learning (Škrlj et. al., 2021), a cutting-edge research area of machine learning. Representation learning refers to modern data transformation techniques that convert data of different modalities and complexity, including texts, graphs, and relations, into compact tabular representations, which effectively capture their semantic properties and relations.

More particularly, this Master's internship will focus on new hybrid software solutions combining two approaches for symbolic and embedding representation (Lavrac et al., 2021) (Škrlj et. al., 2021) *propositionalization approaches*, established in relational learning and inductive logic programming, and (ii) *embedding approaches*, which have gained popularity with recent advances in deep learning.

After having better identified the interest and limitations of these new hybrid approaches based on representation learning techniques, their implementation will be evaluated on specific tasks such as the named entity recognition, and/or relation extraction.

References

(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.

(Lavrac e al., 2021) Lavrac, N., Podpecan, V., Robnik-Sikonja, M. (2021). Representation Learning: Propositionalization and Embeddings. Springer International Publishing.

(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.

(Skrlj et al., 2021) Škrlj B., Lavrac N., Kralj J. (2019) Symbolic Graph Embedding Using Frequent Pattern Mining. In: Kralj Novak P., Šmuc T., Džeroski S. (eds) Discovery Science. DS 2019. Lecture Notes in Computer Science, vol 11828. Springer, Cham.

(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

___