News Representation with Multi-Word Features

Mihail Minev[1] and Christoph Schommer[2]

[1] University of Luxembourg mihail.minev@uni.lu
[2] University of Luxembourg christoph.schommer@uni.lu

Abstract. Information is commonly reflected in news articles. However, texts are unstructured and thus demanding to analyze automatically. To identify and capture the facts in a news story we propose a novel approach, which utilizes natural language engineering. A combination of selected linguistic and statistical criteria enables the identification of grammatical units such as noun, verb, adjective, and adverb phrases. In literature, these entities are presumed to carry the meaning and the information expressed in English texts. In our study, we focus on determining multi-word features in articles related to the monetary policy conducted by the central bank in the USA, FED. The features are composed as attribute-value pairs, where the attributes represent grammatical units, which quantify the major event characteristics. The corresponding values are conditional expressions, which vary over time as facts evolve. The final set is aggregated over the corpus by the application of heuristic and syntax-based rules. Financial experts contributed to the project by providing expertise for the document interpretation.

References

LI, X. (2010): Understanding the semantic structure of noun phrase queries. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1337–1345, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

SHAFIEI, M. and WANG, S. and ZHANG, R. and MILIOS, E. and TANG, B. and TOUGAS, J. and SPITERI, R. (2007): Document representation and dimension reduction for text clustering. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pages 770 –779.

Keywords

INFORMATION EXTRACTION, FEATURE REPRESENTATION