A Temporal Warehouse for Modern Luxembourgish Text Collections

Daniela Gierschek[2], Peter Gilles[2], Christoph Purschke[2],

Christoph Schommer[1] and Joshgun Sirajzade[1]

[1]MINE Lab, Computer Science, University of Luxembourg

[2]Institute for Luxembourgish Language and Literatures, University of Luxembourg

Abstract

The project *STRIPS* is a 3-years project (02/18-01/21) that concerns the implementation of a semantic search toolbox for the retrieval of similar patterns in documents written in Luxembourgish. *STRIPS* is a joint project between the Department of Computer Science and the Institute for Luxembourgish Language & Literature (both University of Luxembourg) and RTL (Radio Television Luxembourg), which acts as data provider. In STRIPS, we focus on RTL text collections of online news and their corresponding user comments (2008-2018). The aim of *STRIPS* is not only to retrieve semantically similar texts (similar news based on their topic content or comments based on their sentiment content), but also to monitor temporal patterns throughout the given time period. The term *semantic* refers here not only to the use of search terms or bag of words (e.g. names, geographical identifiers), but also to more complex structures - such as topics or sentiments of a document. Furthermore, a linguistic processing of these texts takes place by means of tokenizing, normalizing of spellings, stemming, the use of Luxembourgish dictionaries, and part-of-speech tagging (POS). The processing of Luxembourgish language is especially challenging due to the high variation in spelling. Those orthographic differences are likely to be a result of the fact that in the schools the Luxembourgish spelling rules are not sufficiently taught.

Similarity learning algorithms are used to allow fuzzy search queries. The identification of temporal cross-dependencies within the text corpus is also processed. In order to efficiently implement these applications, a *Temporal Warehouse* acts as an essential data backbone with the aim of separating data and applications. The *Temporal Warehouse* offers only two types of user access: the retrieve of data (e.g., via *XQuery*) and the load of data. In addition, each text data entry has a timestamp and sentiment information (at the moment, based on adjectives, sentences and comments). The *Temporal Warehouse* is reproducible from its original data sources through an Extract-Transform-Load (*ETL*) pipeline. At present, the texts are managed in *XML*-format, organized in *TEI*, whereas the *ETL*-part is currently implemented by scripts written in Python and Java. A workbench, which acts as a Graphical User Interface, is available for the tokenization, POS tagging, and morphological analysis. The Luxembourgish language is still a low resource language and a *Temporal Warehouse* for it is unique in its kind, as it is, to the best of our knowledge, the first one to be implemented. In future projects, it could be made available to the general public as a database for searching language specific phenomena. Another possible search query could be the change of sentiment over time. A person interested in Luxembourgish politics

could, for instance, use our data to query whether the sentiment of the population towards a specific politician has changed or not over time. As mentioned, the *Temporal Warehouse* is enriched by the estimation of the sentiment of sentences on the basis of manifold of lexical, grammatical and semantic features, e.g. dictionaries, different part of speeches and word embeddings. So-called *Marts* are intended to serve individual applications with data samples, such as for a topic modeling or a sentiment monitoring. First experiments have shown that such a *Temporal Warehouse* significantly improves the evaluation scores of a connected sentiment analysis engine. **Keywords:** Data Warehouse, XML, Luxembourgish, Sentiment Monitoring, Topic Modeling.

(1) C. Aggarwal. *Machine Learning for Text.* First edition. Springer (2018).

(2) F. Collet. *Deep Learning with Python.* First edition. Manning Publications (2017).

(3) P. Gilles. *From status to corpus: Codification and implementation of spelling norms in Luxembourgish,* In W., Davies and E., Ziegler (Eds.), Macro and micro language planning (pp. 128-149). London: Palgrave Macmillan (2015).

(3) B. Liu. *Sentiment Analysis.* First edition. Cambridge University Press (2015).

(4) R. Kimball. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.* Third Edition. Wiley (2013).

(5) D. Sarkar. *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data.* First edition. Apress (2016).