Providing Digital Answers to Disciplinary Questions with Graph Literary Exploration Machine

Maryl, Maciej

maciej.maryl@ibl.waw.pl
Institute of Literary Research of the Polish Academy of Sciences

Karlińska, Agnieszka

agnieszka.karlinska@ibl.waw.pl
NASK National Research Institute

Walentynowicz, Wiktor

wiktor.walentynowicz@pwr.edu.pl
Wrocław University of Science and Technology

Walkowiak, Tomasz

tomasz.walkowiak@pwr.edu.pl
Wrocław University of Science and Technology

This paper introduces Graph Literary Exploration Machine (GoLEM), a new web-based application for literary scholars, which will become fully operational by the end of 2023 (https://ws.clarin-pl.eu/GoLEM). GoLEM has been developed through close cooperation between literary researchers and computational linguists to tailor the tool to the specificity of literary studies at large and to the current debates and trends in humanities research. We will demonstrate a beta version of the application, discussing its theoretical and methodological considerations, as well as its architecture and applications.

GoLEM addresses the criticism towards Digital Humanities (DH), in particular computational literary studies, revolving around the gap between the development of methods and tools and the use of their potential to formulate new research questions or discover new phenomena, reductionism, ahistoricism and disregard for the socio-cultural circumstances (Bode, 2017). Responding to the call arising from the above criticism, to embed computational analyses within broader disciplinary contexts and knowledge (Underwood, 2019), the cooperation between the researchers and IT professionals was driven by challenges and hypotheses created in a particular disciplinary context, what differs it from its more universal predecessor LEM (Maryl et al, 2018).

The application brings together tools offered by CLARIN-PL and the resources and tools developed in the ongoing DARIAH-Lab project. They are combined in a comprehensive workflow attuned to the needs of literary scholars deriving from specific research questions and the underlying theoretical frameworks, in particular the study of "travelling" concepts (Bal, 2002), cultural analysis, historical semantics, philosophy of science or sociology of knowledge.

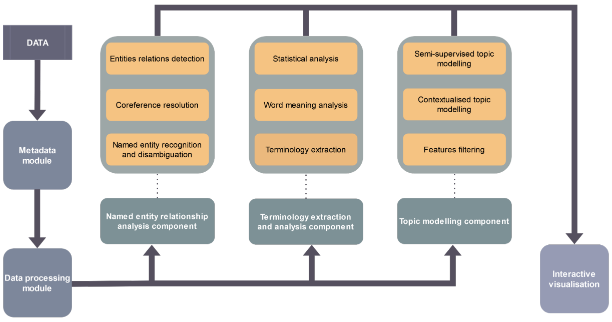
DATA
Metadata module
Data processing module
Entities relations detection
Coreference resolution
Named entity recognition and disambiguation
Named entity relationship analysis component
Statistical analysis
Word meaning analysis
Terminology extraction
Terminology extraction and analysis component
Semi-supervised topic modelling
Contextualised topic modelling
Features filtering
Topic modelling component
Interactive visualisation

Fig.1 GoLEM workflow

The workflow consists of three components allowing for the identification of constellations of terms and concepts and the comparison of their flows over time and space:

1. **Named entity relationship analysis component** performs the tasks of named entity extraction based on a system using pre-trained language models in Transformer technology (Devlin et al., 2019; Conneau et al., 2020), named entity disambiguation with available knowledge bases and linking entities with relations based on a neural network-based classification system. GoLEM extracts named entities and determines the occurrence of relationships of a given type based on contexts in which they are located.
2. **Terminology mining component** uses pattern search methods to extract terminologies described in knowledge bases. It also performs statistical analysis of changes in the meaning of terms between subcorpora.
3. **Topic modelling component** allows topic identification, keyphrase extraction and stylometric comparison. In addition to the LDA technique, semi-supervised topic modelling and contextual topic modelling that uses pre-trained representations of language to support topic modelling (Bianchi et al., 2021a; Bianchi et al., 2021b) are implemented.

GoLEM allows for research designs aimed at showing how the reception of a given author has evolved, reconstructing the terminology networks of selected scholars, grasping differences in their understanding of key concepts, analysing the semantic field of a selected term and exploring the links between authors based on the topics discussed in their papers. Figure 2 shows the results of topic modelling of 1000 texts from 24 Polish anthologies with a literary studies profile, which was performed on the basis of a controlled list of literary terms (due to the space limit and exemplary nature of those figures, we are not elaborating here on sources and interpretation of the results interpretation). Figure 3 provides a chronological visualisation of the results. An interactive visualisation of those results could be found at https://ws.clarin-pl.eu/GoLEM/topic/.

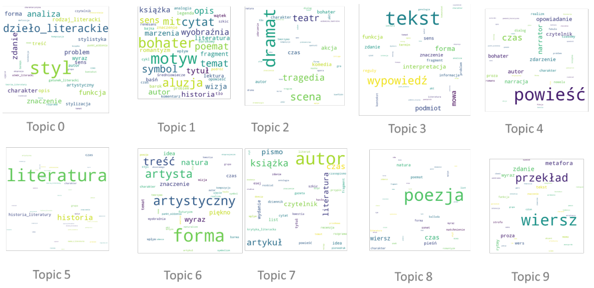

Fig.2 Example of topics generated by GoLEM

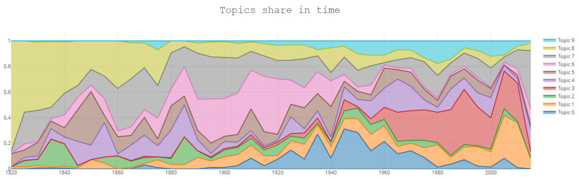

Fig.3 The same group of topics represented chronologically

A strong emphasis is put on the visualisation of results in GoLEM, e.g. as graphs, time series, maps or scatter plots, which allow to take into account local circumstances and to trace the impact of historical and spatial factors on the dynamics of literary processes. GoLEM also allows for determining and visualising relationships between named entities in the corpus. Fig.4 shows the Gephi-made visualisation of co-appearance relationships of persons in the corpus.

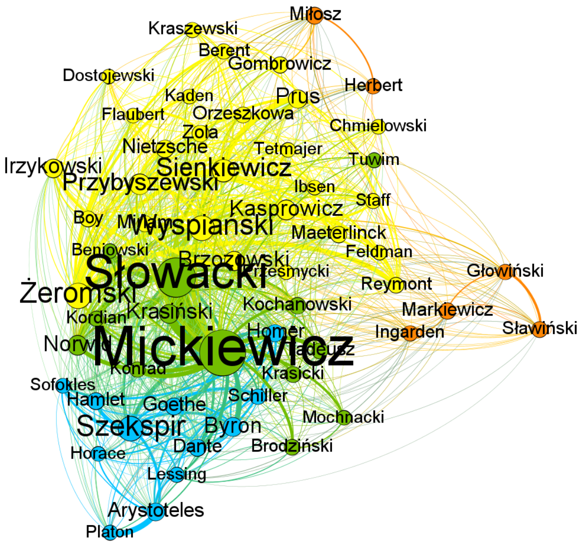

Fig.4 Clustered visualisation of persons from the literary-studies anthologies

GoLEM stands out from other web-based text analysis tools thanks to its close relation between uploaded texts and their metadata: users can add metadata to documents in bulk, filter texts by metadata and group texts based on metadata or keyword searches to perform comparative analyses of subcorpora. Moreover, GoLEM takes into account different levels of user experience, from basic to expert. The system also allows user intervention and iterative adaptation of individual sub-processes of the processing pipeline.

Acknowledgements

This work was conducted in the project *DARIAH-PL: digital research infrastructure for arts & humanities*, Polish Smart Growth Operational Programme 2014-2020 #POIR.04.02.00-00-D006/20.

Bibliography

Bal, M. (2002) *Travelling Concepts in the Humanities. A Rough Guide*. Toronto.

Bianchi, F. / Terragni, S. / and Hovy, D. (2021a). “Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence”, in *Proceedings of the 59th Annual Meeting of the ACL*, Volume 2: 759–766.

Bianchi, F. / Terragni, S. / Hovy, D. / Nozza, D. / Fersini, E. (2021b) “Cross-lingual Contextualized Topic Models with Zero-shot Learning”. In *Proceedings of the 16th Conference of the European Chapter of the ACL*,1676–1683.

Bode, K (2017) “The equivalence of “close” and “distant” reading; or, toward a new object for data rich literary history” *MLQ*, 78: 77–106.

Conneau, A. / Khandelwal, K. / et al (2020). “Unsupervised Cross-lingual Representation Learning at Scale”, in *Proceedings of the 58th Annual Meeting of the ACL*: 8440–8451.

Devlin, J. / Chang, M. / Lee K. / Toutanova, K. (2019) “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in *Proceedings of the 2019 Conference of the North American Chapter of ACL*, Minneapolis, Minnesota: 4171–4186 .

Maryl, M. / Piasecki, M. / Walkowiak, T. (2018) “ Literary Exploration Machine A Web-Based Application for Textual Scholars”, in *Selected Papers from the CLARIN Annual Conference 2017*. Linköping, 128–44.

Underwood, T. (2019) *Distant Horizons: Digital Evidence and Literary Change*., Chicago, IL.