Machine Exploration of Secondary Literature with Literary Exploration Machine

Maciej Maryl
Institute of Literary Research,
Polish Academy of Sciences
maciej.maryl@ibl.waw.pl

Maciej Piasecki
Wrocław University
of Science and Technology
maciej.piasecki@pwr.edu.pl

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

Abstract

This paper presents a design of a web-based application for textual scholars. The goal of this project is to create a complex and stable research environment allowing scholars to upload the texts they are analysing and either explore with a suite of dedicated tools or transform them into another format (text, table, list). This latter functionality is especially important for research into Polish texts, because it allows for further processing with the tools built for English. This project utilises the already existing CLARIN-PL applications and supplements them with new functionalities.

1. Challenge

Digital literary studies seem to be one of the most vividly developing strand of Digital Humanities. Yet, the main obstacle for the development of the field lies in the users' lack of programming skills and insufficient knowledge of how to use digital methods and operate the existing tools. All of the authors of this paper were involved in various educational activities as organisers and instructors of workshops aimed at acknowledging Polish scholars with digital methods of textual scholarship. The main lesson learned from those endeavours is that although there is a genuine interest in computational literary criticism, the learning curve is steep and scholars do not incorporate the tools they were exposed to into their research workflows because they find them too complicated. The results of the 2015 survey into digital methods and practices, conducted by DARIAH DiMPO Working Group, seems to corroborate those claims: scholars articulate a need for technological support and guidance concerning the existing tools (Dallas et al., 2017: 7).

In order to address those challenges we are developing a web-based system, called *Literary Exploration Machine* (LEM)[1], which does not require installation and programming skills. LEM has a component-based architecture, remains open for further expansion with new components, implements natural language processing on different levels and is planned to support several different paradigms of the text analysis.

LEM system is similar to the DARIAH-DE/Topics[2] package in terms of providing an interface and tailored workflows for various analytical tools. However LEM is meant to be more heterogeneous with respect to the tools and techniques used, as well as it is built iteratively through research use cases. Popular *Voyant*[3] offers mainly tracing word forms and their relative frequencies across texts, supplemented by a range of NLP tools added on the a basis of the Stanford CoreNLP, e.g. PN recognition, but limited to English. Moreover, it provides solely simple statistical measures. However, *Voyant* visualisation methods and a friendly GUI serve as an example of good practice for LEM.

2. Design of the system

The project is based on a close cooperation between IT professionals, linguists and literary scholars, which ensures that the tools will suit actual researchers' needs. The interdisciplinary team is working on selected use-cases which will be implemented into the system, listed below.

---

[1] http://ws.clarin-pl.eu/lem.shtml
[2] https://github.com/DARIAH-DE/Topics
[3] Voyant: http://docs.voyant-tools.org, CoreNLP: https://nlp.stanford.edu/software/

1. Research on secondary literature, i.e. scientific journal articles (a comparative study of the transformation in Polish literary scholarship 1989-2014 with a matching study on historical research in that period).
2. Evolution of a collective genre on a web portal (analysis of web content and it evolution overtime on the example of selected portals and weblogs).
3. Working with primary sources (on the example of Radio Free Europe materials).

Each of those case studies is aimed at producing the following content:

1. research questions which could be answered through computational analysis;
2. relevant textual source material;
3. operationalisation of research problems into computational tools;
4. actual tools;
5. rich description of the tool with a hands-on guide for other scholars.

LEM will expand *WebSty*[4] - an open stylometric system, with tools for extracting statistical description of text corpora, as well comparison of texts and subcorpora from the perspective of statistically significant similarities and differences in linguistic structures. In addition to WebSty initial LEM prototype offers functions for extracting features describing texts on the level of: segments (lengths of documents, paragraphs and sentences), morphology (word forms, punctuations, pseudo-suffixes and lemmas), grammatical classes and categories (e.g. from the Polish National Corpus tagset (Przepiórkowski et al., 2012)) and their n-grams.

The **processing paradigms** share the following workflow:

1. *Uploading* a corpus of documents together with metadata in CMDI format.
2. *Text extraction* and cleaning. OCR-ed documents usually contain many language errors that should be corrected at this stage through a set of predefined (and customisable) procedures.
3. *Selecting features* for the description of documents - done manually by users. This step is based on hands-on guide. Users are not expected to have advanced knowledge of Natural Language Engineering or Data Mining.
4. *Setting up* the parameters for processing - manually, but default settings of parameters will be provided. More advanced users will be able to tune the tool to their needs.
5. *Pre-processing texts* with language tools provided by CLARIN-PL. Each text is analysed by a part-of-speech tagger (e.g. *WCRFT2*) and next piped to a Name Entity Recognizer (e.g. *Liner2*), temporal expression recognition, word sense recognition (*WoSeDon*), etc.
6. *Calculating feature values* for the pre-processed texts. Extraction of features encompasses counting frequencies, but also annotations matching patterns for every position in a document.
7. *Filtering and/or transforming* the original feature values. Most filtering and transformation functions is provided by WebSty and its components, further data analysis packages, and transformations specific for comparison of corpora will be added.
8. *Data mining* - several processing paradigms are employed, namely: *topic modelling* - (e.g. *Mallet*) representing document in terms of stochastic processes generating word occurrences from topic-related subsets in text, *unsupervised clustering* - revealing document groups by analysing similarity of document feature vectors (WebSty expanded with semantic features) and supervised classification (a prototype) - application of Machine Learning (e.g. *Weka*[5], *scikit-learn*, and *SciPy* packages) to documents annotated by user categories.
9. *Presenting the results*: visualisation or export of data - in the latter case rankings of features can be interactively browsing or downloaded in formats allowing for further exploration in available tools (e.g. spreadsheets or *Gephi*).

[4] WebSty: http://websty.clarin-pl.eu/, Mallet: http://mallet.cs.umass.edu/

[5] Weka: http://www.cs.waikato.ac.nz/ml/weka/, scikit-learn: http://scikit-learn.org/stable/, SciPy: https://www.scipy.org, Gephi: https://gephi.org

3. Use Case

LEM prototype was developed by the team working with a particular textual corpus of 2,553 Polish texts, published in *Teksty Drugie*[6], an academic journal dedicated to literary studies. The corpus is consisted two parts: OCRd scans (1990-1998) and digital files (1999-2014). Given the aim of this paper (software presentation) and the shortage of space, we will treat the results only as examples of the method, without getting into too much detail.

The work on the prototype was divided into stages, conceived as a feedback loop for the developing team: on every stage a new service was added to the application and the test run was performed. After the analysis of the result, the step was repeated or the team moved to the next phase.

Phase 1. The OCR-ed corpus has been cleaned (e.g. word breaks and headers were removed).

Phase 2. The corpus was lemmatized and parts of speech were tagged. Frequency lists were created that enabled the search for patterns in the textual output. This allowed, for example, tracing patterns of interest in particular Polish poets throughout 25 years, based on lemmatized mentions.

Phase 3. The analysis of the word frequencies, especially experiments ML-based training of classifiers revealed problems with the word list, e.g. occurrences of numbers, years and city names, which were preserved in bibliographic references. A functionality of adopting a custom stopword list was added. The exclusion of corpus-specific problematic words and general meaningless words (e.g. *a*, *this*, *that*, *if*) allowed for visualisation of the most frequent words in *Teksty Drugie* (Fig. 1)

Figure 1. 300 most frequent words from *Teksty Drugie* (1990-2014) (meaningless words excluded) visualised with *Wordle*.

Phase 4. The texts were then grouped into clusters of 20, 50 and 100 in a series of experiments with WebSty. Each grouping revealed a bit different level of generalization about the texts.

By choosing the level of granularity (20, 50 or 100 clusters) we may analyse diverse patterns of discursive similarities between texts. Table 1 shows the differences in clustering of the same sample. The first option (20) shows the similarity between texts on a rather general level, that could be described as stylistic or genre similarity (e.g. formal vocabulary). Other options allow for more detailed exploration of general research approach (50) or particular topics analyzed in articles (100). Semantics of clusters is described by the identified characteristic features.

[6] http://tekstydrugie.pl/en/

Number of clusters	100	50	20
Cluster size (mean)	25.33	50.66	56.65
Cluster size (median)	24	47	51,5
Smallest cluster size	13	25	2
Largest cluster size	51	91	96

Table 1. Differences between the clustering options (numbers reflect the quantity of texts assigned to particular cluster)

Researchers may explore all options and analyse the vocabulary responsible for classifying particular texts into a certain group by a virtue of being over- or under-represented in comparison to the entire sample. Full interpretation of the results of *Teksty Drugie* analysis with LEM were published in (Maryl 2016).

4. Further Development

Currently LEM's GUI is being developed in cooperation with potential users, literary scholars working on various types of texts (fiction, journal articles, blog posts). That is also why we call this software "literary", because further development will address the issues pertinent for literary theory, exceeding a purely linguistic perspective. Some literary-specific issues and functions will be expanded on the later stage of development, e.g. with adding language tools for Word Sense Disambiguation and partial analysis of the text structure, like anaphor resolution and discourse structure recognition. LEM's architecture is open for such extensions. Yet, in this paper we have focused on the current stage of development.

LEM will be fully implemented and made available as a web application to the scholarly audience working on Polish. Next, it will be extended with with tools for other languages (e.g. English and German). As LEM has a modular architecture, it would require mostly linking new processing Web Services and adding converters. LEM is available on open licences and we will be happy to share our tools, code and *know-how*. Export options to other formats will be added, so researchers can easily create the output in a particular format (list, text, table) and upload it to other applications (e.g. Mallet) for further processing.

References (selected)

Dallas, Costis, Chatzidiakou, Nephelie, Benardou, Agiatis, Bender, Michael, Berra, Aurélien, Clivaz, Claire, … Zebec, Tvrtko. (2017). European survey on scholarly practices and digital needs in the arts and humanities - Highlights Report. Zenodo.

Maryl, M. (2016) „Tekstów świat. Przyczynek do makroanalitycznej monografii czasopisma literaturoznawczego" [World of Texts. Take on a Macroanalitical Monograph of a Scholarly Journal] In Nasiłowska, A. & Łapiński, Z. (Eds.), *Projekt na daleką metę. Prace ofiarowane Ryszardowi Nyczowi*, Warszawa: Wyd. IBL, pp. 443-462.

Piasecki, M.; Walkowiak, T. & Eder, M. (2016) WebSty –- an Open Web-based System for Exploring Stylometric Structures in Document Collections. In Eder, M. & Rybicki, J. (Eds.) Digital Humanities 2016 Conference Abstracts, Jagiellonian University and Pedagogical University, 2016, 859-861.

Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.