Making sense of non-sense. Tracing topics in a historical corpus on psychiatry facing low OCR quality

Maria Biryukov – Historical Consulting Luxembourg; Lars Wieneke, Eva Andersen – C^{2}DH, University of Luxembourg

In this paper, we discuss our experiences in evaluating the transnational dissemination of ideas within the psychiatric community as expressed in publications of psychiatric associations (mid 19th to early 20th century). While the underlying research questions are part of an ongoing PhD project that covers periodicals from psychiatric associations in multiple countries using different languages (French, Dutch, English and German) our experimental probe was intentionally limited to a collection of journals published by the "Royal College of Psychiatrists"[1] between 1853 and 1925. This limitation became necessary for different reasons: first due to time constraints. The probing exercise is part of an experiment at our institution that teams up historians with computer scientists for up to six weeks in order to boost their digital research activities. Thereby the full corpus would have easily exceeded the available time without yielding relevant results for the involved parties. Second, an English language corpus offered easy access to already existing tools and toolchains while English was also the transfer language for all partners working on the project.

As opposed to the structure of modern scientific journals, our probe contains a wide variety of entries, among which "original articles" formed only a small part. In order to identify scientific ideas and follow their evolution, one would have to separate articles from asylum reports, board member elections, obituaries, etc. However, the irregular structure of the scanned volumes (e.g. presence or absence of table of contents) and the limited quality of OCR, did not leave stable ground to rely on and to exclude irrelevant material in a simple automatic way. Therefore, we decided to apply a technique that would help us to split the entire content into semantically coherent portions which we would use to re-shape the corpus for further analysis using topic modelling. Before applying the topic modelling, we pre-processed the corpus to maximize content-bearing pages. The main goal at this stage was to get rid of long membership lists, indexes and badly OCR-ed pages. To this end, we split the entire collection into pages and evaluated their relevance using stop-word/content word statistics while dropping pages that didn't match our relevance criteria.

Using this approach, we arrived at 47.085 pages[2] that we analyzed through Non-negative matrix factorization (NMF – Wang et al. 2012) following the approach introduced by Greene and Cross (2017) which relies on NMF for the topic modelling and offers an on-the-fly evaluation of the partitioning into topics. Besides the originally used word2vec technique (Mikolov et al. 2013) we experimented also with its modification, Paragraph2Vec (Le and Mikolov 2014) due to its claimed strength in accounting for context and better capturing the meaning of words. Both models were trained on the entire corpus and we

[1] see https://www.rcpsych.ac.uk/about-us/library-and-archives/our-history for details. The association was created as the "Association of Medical Officers of Asylums and Hospitals for the Insane" in 1841, changed its name to "Royal Medico Psychological Association" in 1926 and took its current name in 1971.

[2] starting from an original volume of 51.479 pages

generated topic sets of 4 to 10 topics for each year allowing us to create "window topics". The algorithm also performed an estimation of the most coherent sets of topics.

Based on the window topics, we generated various sets of dynamic topics, considering time blocks of 5 years until the entire time span. We also experimented with tracking specific topics only, thus addressing the problem of mixed content of the original corpus.

Evaluation

A subset of window topics (one year for each decade, 1853-1923) as well as dynamic topics have been reviewed by a domain expert. The goal here was to evaluate the extent to which generated topics were representative for the field in the given span of time.

While the topics have been considered meaningful, the best number of topics estimated by the system did not always correspond to the expert's view. In case of disagreement, the overall tendency of the expert was to choose a higher number of topics which allowed for a more fine-grained explanation of the subject, highlighting aspects that otherwise would have been overlooked. It leads us to an important conclusion, that even though the automatically generated results make sense and provide insight into a large collection of data, an expert analysis of the topic scope has to be made every time in order to reach sound interpretation of the material.

Conclusion & Future work

The probing exercise allowed us to identify different topics and their evolution over time using an imperfect corpus. However, as it is the nature of explorative research, it led to more questions than answers. While we need to better understand how OCR quality influenced topic modelling in our corpus it also raised issues on the necessary granularity of topics to understand when topics become meaningful without creating an abbreviated perspective on the content. We therefore plan to follow up on this initial testing by combining the initial topic modelling with a visual interface that facilitates the evaluation of the created topics and fosters the exploration of topic evolution in time. We hope to showcase some of the outcomes during the conference. To go beyond topic modelling and evolution tracking, we currently implement a two-stage approach that first retrieves documents (pages), relevant to a selected topic, and then performs a multi-document (multi-page) summarization with respect to the topic (see Biryukov, Ageluta, Moens 2005; Radev, Jing and Budzikowska 2000; Rossiello, Basile and Semeraro 2017).

In the next step, we also want to tackle our main challenge of tracing the transnational dissemination of ideas by identifying persons and their relationships as documented in the corpus building on our previous work in entity extraction (see Novak et al. 2014) as well as work by Green, Feinerer and Burman (2015).

Bibliography

Maria Biryukov, R. Angeluta and M.-F. Moens (2005). Multidocument question answering text summarization using topic signatures. In Journal of Digital Information Management, Vol. 3, Number 1

Derek Greene and James P. Cross (2017), "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modelling Approach". In: Political Analysis, Volume 25, Issue 1, pp. 77-94.

Dragomir R. Radev, H. Jing and M. Budzikowska (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. NAACL-ANLP workshop, vol2.

Gaetano Rossiello, P. Basile and G. Semeraro (2017). Centroid-based text summarization through compositionality of word embeddings. In Proceedings of the MultiLing Workshop on Summarization and Summary Evaluation across source types and genres, pp. 12-21.

Christopher D. Green, Ingo Feinerer, Jeremy T. Burman, Searching for the structure of early American psychology: Networking Psychological Review, 1894–1908. History of Psychology, Vol 18(1), Feb 2015, 15-31

T. Mikolov, K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. CoRR abs/1301.3781.

Jasminko Novak, Lars Wieneke, Marten Düring, Isabel Micheel, Mark Melenhorst, Javier Garcia Morón, Chiara Pasini, Marco Tagliasacchi, and Piero Fraternali. 2014. "HistoGraph: A Visualization Tool for Collaborative Analysis of Historical Social Networks from Multimedia Collections." IV2014 - DHKV: Cultural Heritage Knowledge Visualisation, 2014.

Q. Le, T. Mikolov (2014). Distributed Representations of Sentences and Documents. CoRR abs/1405.4053.v2

Q. Wang, Z. Cao, J. Xu, H. Li(2012): Group matrix factorization for scalable topic modelling. In: Proc. 35th SIGIR conference on Research and development in Information Retrieval. pp. 375–384. ACM