SHAX: A Semantic Historical Archive eXplorer

Michael Feldman[1], Shen Gao[1], Marc Novel[2], Katerina Papaioannou[1], and Abraham Bernstein[1]

[1] Department of Informatics, University of Zurich, Zurich, Switzerland
[2] Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

Abstract. Newspaper archives are some of the richest historical document collections. Their study is, however, very tedious: one needs to physically visit the archives, search through reams of old, very fragile paper, and manually assemble cross-references. We present SHAX, a visual newspaper-archive exploration tool that takes large, historical archives as an input and allows interested parties to browse the information included in a chronological or geographic manner so as to re-discover history.
We used SHAX on a selection of the Neue Zürcher Zeitung (NZZ)—the longest continuously published German newspaper in Switzerland with archives going back to 1780. Specifically, we took the highly noisy OCRed text segments, extracted pertinent entities, geolocation, as well as temporal information, linked them with the Linked Open Data cloud, and built a browser-based exploration platform.
This platform enables users to interactively browse the 111906 newspaper pages published from 1910 to 1920 and containing historic events such as World War I (WWI) and the Russian Revolution. Note that SHAX is neither limited to this newspaper nor to this time-period or language but exemplifies the power in combining semantic technologies with an exceptional dataset.

1 Introduction

During the past decade, many newspapers (most notably the New York Times [3] but see also [1] for an overview) have digitalized their archive in order to make it searchable and publicly available. Usually, the scanned newspapers are converted into text via the use of Optical Character Recognition (OCR). The received output contains a great degree of noise and makes knowledge discovery from historical newspaper archives a challenging task. Approaches like data cleaning with specialized Information Retrieval (IR) tools are commonly used for this task but require substantial human involvement and domain-specific knowledge [4].

Alternatively, we develop a Semantic-Web based, data-driven approach, which effectively retrieves information from a large volume of newspaper issues. Our methodology was applied to a part of the digitalized archive of the Neue Zürcher

[3] http://open.blogs.nytimes.com/2013/07/11/introducing-the-new-timesmachine/

Zeitung (NZZ) for the years ranging from 1910 to 1920 and is applicable to different news corpora in various languages. The interactive visualization of our results enables the user to browse and discover historical events with the related geographic information.

2 Dataset

The NZZ kindly provided us with a part of the archive covering the issues published from 1910 to 1920. This period covers historic events such as WWI and the Russian Revolution. The scanning, digitizing, and OCRing of the NZZ archive was conducted by the Fraunhofer Institute[4]. The dataset we use consists of 354 GB scanned PDFs and 111906 OCRed pages in XML format (one XML file per newspaper page).

One of the biggest problems when processing the data is noise. The OCR struggled with the Gothic font, which is used during the longest period in the archive, including the one under discussion. Additionally, during wartime, when printing resources were scarce, ink and paper quality decreased: some pages are simply not readable and others were printed on thin paper, causing the text of the backside to shimmer through the front side in the scans. The recognized text also contains unavoidable errors, such as different word-spelling due to language change. However, the names of high-frequency entities remain the same during this time period. These errors cannot affect our results considerably.

As a result only a part of the text was correctly recognized. Using a spell-checker, we found that only 64% of the words were correctly recognized. As we assume a random distribution of the errors, our results contain insignificant biases.

3 The Application

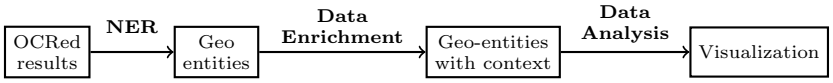


Fig. 1: The Semantic-Web based approach for analysing newspaper corpora

3.1 The Semantic Web Approach

Existing ways of dealing with historical corpora rely on Information Retrieval methods requiring a substantial amount of human effort. Based on the assumption that the locations of important historical events are explicitly mentioned in the newspaper, we develop a purely data-driven approach that leverages Semantic Web technologies to analyze the noisy dataset (Fig. 1).

Specifically, we first perform the Named Entity Recognition (NER) on the dataset with DBpedia Spotlight [3], trying out different confidence values to

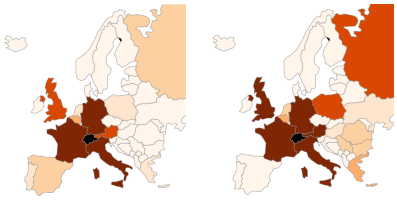
[4] https://www.iais.fraunhofer.de/nzz.html

improve the accuracy. We focus on the correlation between temporal and geographical information, hence, we only extract the geographic entities (e.g., the city of Sarajevo). Each entity is linked with its corresponding meta-data as well as information retrieved from DBpedia and GeoNames. For example, we query the longitude and latitude form DBpedia and use GeoNames to find its county code by reverse geo-indexing. The result of this process includes tuples in the following format: (entity name, longitude, latitude, country code, DBpedia link, date of mention, issue ID). Finally, we perform data analysis on the results on a monthly basis by aggregating on the country code or the entity name. In both cases, we compute the sum of counts in every group.

3.2 The Interactive Visualization

In this section, we briefly introduce the functions of our exploration platform which is available at https://files.ifi.uzh.ch/ddis/nzz-demo/WebContent/.

Function 1: Country Mentions over time A choropleth-map of Europe was generated for each year based on the country counts. As shown in Fig. 2(a) and 2(b), the color intensity of a country is in proportion to its counts (i.e. the darker the color, the more the counts). By navigating through the years, the way the colors change provides an overview of the popularity of each country. For example, the Balkan countries are mentioned more often at the beginning of WWI. In order to avoid biases, such as countries being mentioned extensively due to higher geographical proximity to Switzerland or due to larger population, the annual counts of each country were also normalized by relative distance ([5]) to Zürich and population estimated in 1910.



(a) 1910 (b) 1915

Fig. 2: Color Changing of EU in 1910 and 1915

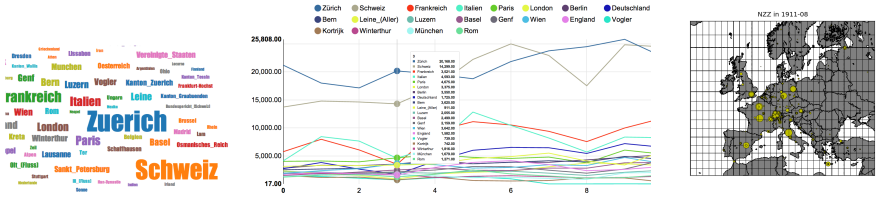
Function 2: Linking countries, issues and historical events By constructing an inverse index that links the countries to the issues where they are mentioned, users can further explore the reasons behind the change in the colors. By clicking on a country, they can see the historical events it was involved in Fig. 3 as well as the relevant newspaper's PDFs. The historical events presented are systematically extracted from DBpedia [2] by querying the category "Event" with the corresponding country. A newspaper's PDF is considered relevant if the country or a place within its borders was mentioned.

Function 3: Entity Mentions over time A more detailed analysis of entity mentions is visualized using the word cloud (Fig. 4(a)) and trend line (Fig.

Date	Event	Wikipedia Link
13-3-1915	Battle of Neuve Chapelle	http://en.wikipedia.org/wiki/Battle_of_Neuve_Chapelle
9-5-1915	Battle of Aubers Ridge	http://en.wikipedia.org/wiki/Battle_of_Aubers_Ridge
27-5-1915	Battle of Festubert	http://en.wikipedia.org/wiki/Battle_of_Festubert
14-10-1915	Battle of Loos	http://en.wikipedia.org/wiki/Battle_of_Loos
4-11-1915	Third Battle of Artois	http://en.wikipedia.org/wiki/Third_Battle_of_Artois
6-11-1915	Second Battle of Champagne	http://en.wikipedia.org/wiki/Second_Battle_of_Champagne

Fig. 3: Linking countries, issues and historical events

4(b)) of all geographic entities per year. It is possible to directly observe the changes in the popularity of each entity over time, as well as correlations among them. Additionally, we plot each entity as a bubble on the map based on its coordinates and number of mentions (Fig. 4(c)). Thus, users could discover the actual location of a historical event.



(a) Word Cloud (b) Trend Line (c) Bubble Map

Fig. 4: Entity Mentions over time

4 Conclusion

SHAX is a browser-based exploration platform, which highlights the major role of Semantic Web tools in extracting entities mentioned in highly noisy newspaper archives. Moreover, it shows that interactive visualization is necessary not only in presenting the information within the newspaper in a user-friendly way, but also in discovering implicit knowledge from the corpus. In the future, we plan to apply our method on the whole 250 years of NZZ archives and try to extend it to explore other kinds of entities such as notable people.

Acknowledgements

We would like to thank the Neue Zürcher Zeitung, Thilo Haas, Thomas Scharrenbach, and Daniel Spicar.

References

[1] Doerr, M., Markakis, G., Theodoridou, M., Tsikritzis, M.: Diathesis: Ocr based semantic annotation of newspapers (2007)
[2] Hienert, D., Luciano, F.: Extraction of historical events from wikipedia. In: CoRR (2012)
[3] Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proc. of the 7th i-Semantics (2011)
[4] Places, Á., Fariña, A., Luaces, M., Pedreira, Ó., Seco, D.: A workflow management system to feed digital libraries: proposal and case study. Multimedia Tools and Applications (2014)
[5] Worboys, M.F.: Metrics and topologies for geographic space. In: Proc. 7th Intl. Symp. Spatial Data Handling (1996)