A Framework for Historical Russian Flu Epidemic Exploration from German Newspapers

Tran Van Canh
ctran@l3s.de
L3S Research Center, Hannover, Germany

Katja Markert
markert@cl.uni-heidelberg.de
Heidelberg University, Germany

Wolfgang Nejdl
nejdl@l3s.de
L3S Research Center, Hannover, Germany

Introduction

The Russian flu 1889-1893 epidemic reached Europe from the East in November and December of 1889 and spread over the whole globe in the space of a few months. It was one of the first epidemics of influenza that occurred during the period of the rapid development of bacteriology. In addition, it was the first ever epidemic that was publicly and intensively narrated in the developing daily press, especially those published in German located in Germany and Austria (Mirosławska et al., 2013). However, as stated in (Valtat et al., 2011), very limited information about the epidemiology of this influenza has been found, which was based on materials published in English. While a large amount of news about the flu was published in German, it is hard to find a study on the epidemic based on German documents. These motivate our goal in this work, which is to build a framework from German materials to support research community getting more insights into the disease. Our framework consists of different components including data collection and cleaning, corpus creation, and associated tools for analysis. The framework is pictorially shown in Figure 1.

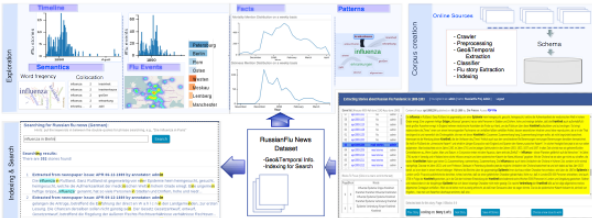

Figure 1. Russian flu exploration framework

Related work

There is limited information about the epidemiology of the Russian flu epidemic 1889-1893. In (Mirosławska et al., 2013), the authors conducted an analysis to examine the impact of the epidemic in 14 cities in Europe. Their results showed that the epidemic spread quickly from Saint Petersburg, Russia to other parts of Europe with a speed of around 400 km/week and reached the American continent only 70 days after the original peak in Saint Petersburg. In addition, some detailed information about case fatality ratio and the median basic reproduction was given also. However, their work was based on reports of only two local daily newspapers in Poznań, which implies some uncertainty due to the lack of data coverage. Valleron et al., 2010 presented a case study on the transmissibility and geographic spread of the Russian flu. A similar approach was followed by Valtat et al., 2011 to examine the age distribution of the affected people and the mortality rate of this flu event. In a recent study, Ewing et al., 2016 collected contemporary reports and explored a digital humanities approach to interpret information dissemination regarding this particular epidemic. The limitations common to all of these studies are the heterogeneity and lack of coverage of data used.

Data preparation

ID	Keyword	Variation
1	Influenza	Jnfluenza,Jnsolvenza
2	Epidemie	
3	Influenza-Epidemie	Influenzaepidemie
4	Grippe	
5	erkrankt	ertrankt
6	Pathologie	

Table 1: Keywords used to collect high recall collection of newspaper issues containing stories about influenza epidemic

Data collection

Data used in this work was collected from the Austrian Newspapers Online (ANNO) repository. ANNO contains almost all issues from many newspapers in Austria and Germany during the time the Russian flu epidemic took place. The data are accessible in both scanned PDF and OCR formats. These are appropriate for our goal in terms of extracting Russian flu related stories from noisy OCR text and checking against the scanned PDF content for validity. To establish the data collection, the keywords listed in Table 1 (along with some misspelt variations of keywords, which were included due to OCR misrecognition) were used to search the ANNO repository The search query was constrained by the time interval from 1889 to 1893. After preprocessing the search results we obtained 4,806 issues, which become the candidates to extract stories about the Russian flu.

Noise reduction

Due to the low quality of the scanned images of newspaper issues, a lot of noise is present in the corresponding OCR texts. The word error rate (WER) computed on sample texts is around 18.9%. Our goals here were to remove noise and correct misrecognized words as much as possible but at the same time manage keep the language as it was so that the derived corpus pertains its historical perspective. It is noted that modern German is rather different in writing and usage of many words due to the language's evolution. To cope with these issues, we adopted a snapshot of the Google-2-gram dataset for German from 1885 to 1895. The dataset was used to train our bigram-based model for word segmentation and spell checking. After running the model, the word error rate was reduced to 5.5%.

Text block classification

A difficult challenge for the task of extracting complete stories is that recognized OCR text blocks are very often not aligned in the same order as they were in the original image of an issue. Our approach was to automatically pre-classify OCR text blocks to identify the ones that are more likely part of some flu-related stories. Then we developed a tool to effectively help annotators extract complete Russian flu stories. For this, we adopted the KL-divergence based technique developed in (Schneider, 2004) to build a classifier. The model was trained with 245 OCR text paragraphs and obtained recall of 81.5% and precision of 68.6%. Basically, the output of the classifier can be used to help annotators start working on an issue by looking at suggested text blocks first, from which they can then select paragraphs that are part of the same story.

Extraction tool

After completing the high-recall automatic pre-extraction, we implemented a Web-based tool for annotators to help build our corpus collaboratively. The main GUI of our tool is shown in Figure 2. After having annotators work through the whole collection, we obtained a corpus of 639 news articles about Russian flu from 42 newspapers, identified with 85.7% agreement between annotators.

Figure 2. Main GUI of our tool for Russian flu story extraction

Geo and temporal information extraction

Given that location and time are helpful features for exploring the development of the epidemic, we extracted and normalized geographic names and temporal expressions occurring in the corpus. For geographic names, the Geodict tool created by Pete Warden (2011) was adapted to work with country and city names in German. HeidelTime (Strötgen and Gertz, 2013) was used for temporal information extraction and normalization.

Indexing and search engine

We created a search engine on the corpus to support research community in searching for information. The searching GUI is shown in Figure 3.

Figure 3. Russian flu story searching module

Exploration tools and sample results

We provided associated tools along with the corpus. The corpus timeline provides statistics on the number of stories in the corpus across time and news outlet. In addition, it provides interactive visualization. As an example shown in Figure 4, during the peak time in late December 1889 and January 1890, extensive news about the influenza was published.

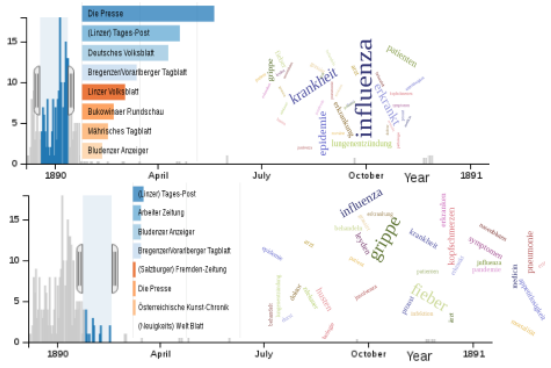

Figure 4. Press attention on the flu and topic changes over time

Newspapers were trying to narrate the outbreak as fast as possible. Words that appear significantly in the stories include *influenza*, *epidemic*, *krankheit (disease)* and *erkrankt (sick)*. A short time after this peak period, fewer reports were published about the outbreak of the flu and communities started discussing the treatment more. Names of doctors appearing in the news (e.g., Leyden, Proust) together with words describing symptoms such as *fieber (fever)*, *kopfschmerzen (headache)*, *appetitlosigkeit (anorexia)*. Furthermore, by exploring word collocations one can find even more interesting information. Figure 5 shows a frequent pattern of word collocation describing the influenza. The words *heute (today)* and *gestern (yesterday)* indicate that news about the flu is updated every day*; and* the word *jänner (January)* implies that the flu outbreaks happened during winter.

Figure 5: A frequent pattern of word collocation extracted from our corpus

Figure 6 presents the co-occurrences of three words *infuenza, erkrankt, and london* over time. It shows a similar trending pattern of the words *infuenza* and *erkrankt* being used to note about the flu. In addition, one can observe that the peak time of the flu in London was from late December 1889 to early January 1890 as indicated in (Honigsbaum, 2010; Goff, 2011). This suggests that the temporal distribution of terms can give us more insights into the evolution of the epidemic.

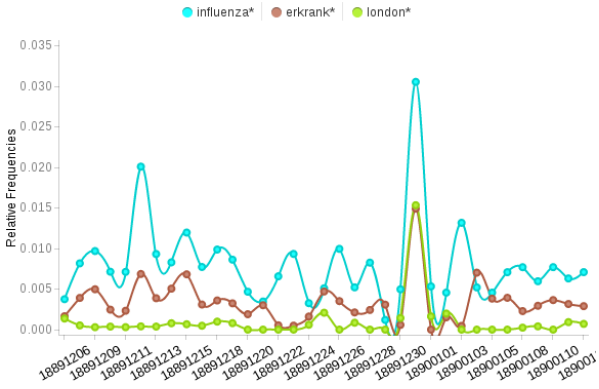

Figure 6: Similar trending pattern of the two words influenza and erkrankt, and the peak time of the flu in London

The framework also supports studying the evolution of the flu over time and geographic regions. We employed the method introduced in (Abdelhaq et al., 2013) for localized event detection. Figure 7 shows three snapshots describing the development of the epidemic over cities in Europe during the peak time from late November 1889 to January 1890.

Summary

We have introduced a framework for research communities to explore the historical Russian flu 1889-

1893 from German newspapers. We developed a tool for collaborative annotators to help build our corpus. We further presented some interesting insights that we achieved from analyzing articles in the corpus. By making the corpus and associated tools available, we provide useful contributions to the community in support of conducting studies on influenza epidemics and evaluating temporal IR models.

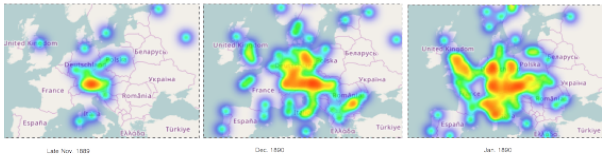

Figure 7. Evolution of the Russian flu over geographic regions during its peak time

Acknowledgements

This research is supported by the German Research Foundation (DFG) for the project "Tracking the Russian Flu in U.S. and German Medical and Popular Reports, 1889-1893" on Grant No. NE 638/13-1. We also thank you the Austrian National Library for supporting us in collecting data.

Bibliography

Abdelhaq, H., Sengstock, C., Gertz, M..(2013). "EvenTweet: Online Localized Event Detection from Twitter." Proc. VLDB Endowment Journal, 6(12):1326-4.

Aramaki, F., Maskawa, S., Morita, M..(2011). "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter." In proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1568-9.

Austrian National Library. (2011) Austrian Newspapers Online. Repository online at http://anno.onb.ac.at

Ewing, E.T., Kimmerly, V. and Ewing-Nelson, S. (2016). "Look Out for 'La Grippe': Using Digital Humanities Tools to Interpret Information Dissemination during the Russian Flu, 1889—90." Medical history,60(1):129-3.

Honigsbaum, M.(2010). "The Great Dread: Cultural and Psychological Impacts and Responses to the Russian Influenza in the United Kingdom 1889–1893".Social history of medicine,23: 299-21.

Kempińska-Mirosławska, B., and WoŸniak-Kosek, A.(2013). "The influenza epidemic of 1889–90 in selected European cities – a picture based on the reports of two Poznań daily newspapers from the second half of the nineteenth century." Med Science Monitor,19:1131-11.

Le Goff, J.M.(2011). "Diffusion of influenza during the winter of 1889-1890 in Switzerland." Jenus, 67(2): 77-23.

Paul, M.J. and Dredze, M.. (2011). "You Are What You Tweet: Analyzing Twitter for Public Health." In proceedings of the Fifth International Conference on Weblogs and Social Media. pp. 265-8.

Schneider, K.-M.(2004). "A New Feature Selection Score for Multinomial Naive Bayes Text Classification Based on KL-divergence." Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. pp. 186-4.

Strötgen, J. and Gertz, M..(2013). "Multilingual and cross-domain temporal tagging." Language Resources and Evaluation, 47(2): 269-30.

Valleron, A.J., Cori, A., Valtat, S., Meurisse, S., Fabrice Carrat, F., and Boëlle, P.Y.. (2010). "Transmissibility and geographic spread of the 1889 influenza pandemic." In proceedings of the National Academy of Sciences of the United States of America (PNAS). pp. 8778-4.

Valtat, S., Cori, A., Carrat, F., and Valleron, A.-J. (2011). "Age distribution of cases and deaths during the 1889 influenza pandemic."Vaccine, 29(2): B6-B10.

Warden, P. (2010) GeoDict. Accessible via Github at https://github.com/petewarden/geodict