A Search Engine for Police Press Releases to Double-Check the News

Maik Fröbe[1(✉)], Nina Schwanke[1], Matthias Hagen[1], and Martin Potthast[2]

[1] Martin-Luther-Universität Halle-Wittenberg, Halle, Germany
maik.froebe@informatik.uni-halle.de
[2] Leipzig University, Leipzig, Germany

Abstract. Many people have doubts about the factual accuracy of online news, while still trusting the press releases of police departments. To enable an easy corroboration of online news about police-related events, we build a search engine for press releases of police departments. Addressing the German "market", the search engine takes the URL of a German piece of online news as input and retrieves relevant press releases of the German police. Comparing different query-by-document strategies in a TREC-style evaluation on 105 topics, we show that our system is able to accurately identify relevant press releases if there are any.

1 Introduction

We introduce Police PR Search[1], a search system that allows for easily double-checking a piece of online news about police-related events (e.g., serious crimes but also accidents or demonstrations) against the content of relevant police press releases. Our current prototype indexes and retrieves press releases of virtually the entire German police force when queried with the URL of some (German) online news article. Readers of online news about police-related events can use the system to retrieve official background information.

An illustrating example is given in Fig. 1. On June 23, 2016, a hostage situation occurred in a cinema in Viernheim, a small town south of Frankfurt and Darmstadt, Germany. The yellow press German newspaper BILD quickly picked up the incident (cf. Fig. 1 (left)), reporting about a rampage and the involvement of explosives. Soon after, the BBC tweeted about 20 casualties in the cinema (cf. Fig. 1 (right)). Indeed, the police did shoot and kill the hostage-taker, but luckily no one else was injured, and no explosives were involved, as explained in the official press release of the police, which was published later after the incident. The sensationally exaggerated and erroneous facts were then removed from the online articles.

Not every piece of (online) news is wrong or sensationally exaggerated. Still, wrong news articles are published frequently [8], and sometimes even intentionally [7]. Readers might thus want to double-check news articles against official

[1] Demo: https://demo.webis.de/police-pr.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 454–458, 2020.
https://doi.org/10.1007/978-3-030-45442-5_57

Fig. 1. Excerpts from the coverage of a hostage taking in a cinema in Viernheim on June 23, 2016, in the German newspaper BILD (left) and a tweet from the BBC (right).

statements, which can be a rather time-consuming process of searching for other trustworthy sources. Manually selecting text fragments as queries, for instance, may still yield the same wrong information in the search results.

To offer some (semi-) automatic support in such situations, we have developed a search engine that can be queried directly with the URL of an online news article. As results to be retrieved, we index official press releases from police and fire departments. They offer information on a lot of local events—topics that many readers are interested in anyway [6]—and the police is a trusted source of information for many [3]. Currently addressing the German market with the prototype, we have crawled and indexed press releases from the German press portal Blaulicht. In our evaluation, we compare different strategies of formulating search queries given a URL. In a TREC-style setup [5] on 105 topics covering 7 classes of police-related events, we show that even the most simple querying strategy (searching the title of the news article in the titles and bodies of the press releases) substantially outperforms the search facility offered by the press portal itself. It turns out that the best (and more involved) automatic querying strategy implemented in our system achieves precision@1 and nDCG@5 scores of about 0.9—a performance clearly indicating practical applicability.

2 Search System and Query-by-Document Strategies

We extract the title, body, date, and police department location from all press releases as fields for retrieval with Elasticsearch's BM25F implementation. As querying strategies against this index, we basically follow a query-by-document approach (the news article as the "query"). Somewhat following previous works that try to identify the most important keyphrases from an input query document to find similar content [2,9], we compare three query formulation strategies.

Our three "query-by-document" strategies extract information from an input news article and combine them as follows: (1) Only the title of the article, (2) title and main content of the article, and, (3) title, main content, and publication date and locations mentioned in the article (if any). Since publication dates and locations can not be extracted accurately for all potential input articles (third strategy), we resort to only title and body information in such cases.

The queries against the Elasticsearch index are formulated as follows: The title of a given news article is queried against the title and body of the police press releases. When a news article has a body, it is queried against that of the indexed press releases. When a publication date for the news article can be extracted, it is queried against the body of the press releases and used as a filter to remove press releases that were published more than two weeks before, and more than eight weeks after that date. The potential locations extracted from a news article are used as queries against the police department name field as well as against the title and body of the police press releases.

3 Evaluation

To test our system, we follow the TREC evaluation paradigm [5]: 1,172,703 press releases form the document collection[2], covering virtually all German police departments. Topics are formed by news articles about police-related incidents, and relevance judgments are obtained in a depth-5 pooling of different search system rankings.

To create topics representative of the "importance" of police-related incidents, we use the German crime statistics of 2018 [1]. The seven categories we select to cover are murder, theft, migration, related to sports events, thunderstorms, traffic accidents, and general capital offenses. As per the results of a G*Power t-test [4], based on a small pilot experiment, we create 15 topics (105 in total).

The individual topics in the form of news articles were compiled as follows: Given a random police press release from one of the aforementioned categories, an expert tried to identify a related news article using various online news search engines. The expert also was instructed to rate a topic's difficulty during the creation. A difficulty of "Level 1" indicates that the news article and the press release use very similar vocabulary, "Level 2" that the titles greatly differ but there are similarities in the bodies, and "Level 3" for larger differences in the titles and the bodies. If no news article was found for some press release after 5 min, the expert continued with another random press release.

The qrels for the 105 topics (i.e., news articles) have been created as follows: The initial press release used to create the topic is judged as highly relevant (score of 2). Then a depth-5 pool of the rankings returned by different querying strategies and retrieval systems is completely judged: the Blaulicht portal's original search facility using the title or the title and the body as query, and our

[2] They are publicly available under www.presseportal.de/blaulicht/.

Table 1. Evaluation of the query-strategies: title (T), title and body (TB), and a combination of title, body, place, and date (TBPD) for various difficulty levels. We compare our search-engine (PPR) with the search engine of the German police press portal (ORI), reporting nDCG@5 and precision@1 (P@1).

Method	Time	All levels		Level 1		Level 2		Level 3	
		nDCG	P@1	nDCG	P@1	nDCG	P@1	nDCG	P@1
T@ORI	0.7 s	0.13	0.14	0.28	0.31	0.01	0.01	0.00	0.00
TB@ORI	0.9 s	0.04	0.04	0.10	0.10	0.02	0.02	0.00	0.00
T@PPR	0.6 s	0.21	0.21	0.44	0.48	0.14	0.11	0.04	0.07
TB@PPR	9.2 s	0.75	0.76	0.81	0.86	0.78	0.79	0.47	0.47
TBPD@PPR	9.1 s	0.92	0.88	0.93	0.90	0.94	0.98	0.59	0.60

three strategies detailed above. The top-5 results of each ranking are judged on a graded scale from 0 (irrelevant) to 2 (highly relevant). A press release is judged as "highly relevant" (score 2) if it directly deals with the event described in the news article, and as "relevant" (score 1) if the news article's event refers to the police press release. Most topics only have one highly relevant press release.

Table 1 shows the aggregated effectiveness of the different systems/strategies in terms of precision@1 and nDCG@5. The performance of the Blaulicht portal's search facility is rather low: even for easy topics (Level 1), hardly any relevant documents are found using a news article's title as the query. A reason might be that some exact match retrieval is used since for many topics no result is returned. This trend gets even worse if additional information in the form of the bodies of the news articles is incorporated into the query.

Our Elasticsearch-based system outperforms the portal's search facility by far on every category (differences significant) even when only the title is used as the query. Adding more information to the query (body, location, date) results in slower search as the news articles' body texts produce very long queries. Still, the effectiveness greatly improves by adding body as well as location and date information to the query with precision@1 and nDCG@5 reaching 0.9 on average. However, this gain in effectiveness comes at the cost of an increased average response time of more than 9 s. Testing and implementing strategies selecting the most informative keywords and phrases from the body to reduce query length thus form an interesting direction for future efficiency improvements.

4 Conclusion and Future Work

Our prototype shows that using a news article as the query against an index of police press releases can often very accurately deliver background information about police-related incidents. Facts can directly be double-checked against official statements, a source of trust for many. For future work, we envision improvements in a number of directions. The efficiency for long queries involving a news article's body text can possibly be improved by keyphrase extraction

methods, reducing the queries to maybe several tens of words only. It would also be interesting to more closely analyze cases where no press release can be identified; in our user study, this often was the case when the vocabulary greatly differs and broadening this analysis might help to avoid showing only irrelevant or no results at all.

Acknowledgements. We thank Ahmad Dawar Hakimi, Anita Susheva, Brennan Nicholson, Christopher Pfeiffer, Christoph Traser, and David Sturm for their work on the first prototype and for crawling the press releases. Our special thanks go to the Blaulicht press portal for supplying us with their collection of police press releases for research.

References

1. Polizeiliche Kriminalstatistik, Bundeskriminalamt (2018). https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/PolizeilicheKriminalstatistik/2018/pks2018Jahrbuch1Faelle.pdf?__blob=publicationFile&v=6. Accessed 23 Sept 2019
2. Dasdan, A., D'Alberto, P., Kolay, S., Drome, C.: Automatic retrieval of similar content using search engine query interface. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, 2–6 November 2009, pp. 701–710 (2009)
3. European Commission: Wie sehr vertrauen Sie der Polizei? (2019). https://de.statista.com/statistik/daten/studie/377233/umfrage/umfrage-in-deutschland-zum-vertrauen-in-die-polizei/. Accessed 19 Jan 2020
4. Faul, F., Erdfelder, E., Buchner, A., Lang, A.G.: Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav. Res. Methods **41**, 1149–1160 (2009)
5. Harman, D.: TREC-style evaluations. In: Agosti, M., Ferro, N., Forner, P., Müller, H., Santucci, G. (eds.) PROMISE 2012. LNCS, vol. 7757, pp. 97–115. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36415-0_7
6. Schrøder, K.: What do news readers really want to read about? How relevance works for news audiences. Digital News Publications (2019). http://www.digitalnewsreport.org/publications/2019/news-readers-really-want-read-relevance-works-news-audiences/. Accessed 19 Jan 2020
7. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
8. Tandoc Jr., E., Lim, Z., Ling, R.: Defining "fake news". Digit. Journal. **6**(2), 137–153 (2018)
9. Yang, Y., Bansal, N., Dakka, W., Ipeirotis, P.G., Koudas, N., Papadias, D.: Query by document. In: Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, 9–11 February 2009, pp. 34–43 (2009)