Workshop on Text Mining and Generation (TMG)

Mirko Lenz[1,*,†], Lorik Dumani[1,†], Alexander Bondarenko[2,†] and Shahbaz Syed[3]

[1] *Trier University, Universitätsring 15, 54296 Trier, Germany*

[2] *Friedrich-Schiller-Universität Jena, Fürstengraben 1, 07743 Jena, Germany*

[3] *Leipzig University, Augustusplatz 10, 04109 Leipzig, Germany*

Abstract

This paper is a report on the first Text Mining and Generation Workshop (TMG), which was a one-day virtual event hosted at the German Conference on Artificial Intelligence (KI 2022) in Trier, Germany. In addition to four accepted original papers, there were three invited talks by speakers who presented their works already published at high-ranked conferences as well as one keynote by a pioneer in the two research fields relevant to the workshop.

Keywords

Text mining, Text generation, Natural language processing, Workshop

1. Introduction and Motivation

Digital text data is available in large amounts and different granularities. Typical sources of this data are social media posts, books, news articles, web pages, company reports, etc. A major challenge this text data imposes is that it is unstructured and has to first be processed to make further analysis possible. At the same time, there are also many situations in which only structured data is available that is to be verbally explained, for instance, by Explainable AI. These contrasting scenarios lead to two complementary application areas: *text mining* and *text generation*. The aim of text mining is to analyze the content of unstructured text and extract (useful) structured information. In contrast, text generation attempts to (automatically) create text from structured information or knowledge that is for example stored in large language models. The goal of the TMG workshop is to bring these two perspectives together by eliciting research paper submissions that aim for bridging the gap between knowledge extraction and text generation. Since recent approaches to text mining and text generation are predominantly based on artificial intelligence (AI) methodologies, KI 2022 has been a great venue to bring together AI researchers working on these two tasks.

KI 2022, September 19–23, 2022, Trier, Germany

[*] Corresponding author.

[†] These authors contributed equally.

info@mirko-lenz.de (M. Lenz); dumani@uni-trier.de (L. Dumani); alexander.bondarenko@uni-jena.de (A. Bondarenko); shahbaz.syed@uni-leipzig.de (S. Syed)

https://www.mirko-lenz.de (M. Lenz)

0000-0002-7720-0436 (M. Lenz); 0000-0001-9567-1699 (L. Dumani); 0000-0002-1678-0094 (A. Bondarenko); 0000-0002-4821-1507 (S. Syed)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Accepted Papers

In total, we accepted four papers for publication. In the following, we provide a brief overview of their main contributions. The paper "German to English: Fake News Detection with Machine Translation" applies translation (essentially a *text generation* task) to mitigate the fact that languages other than English mostly have worse ML models available. Their evaluation shows that translating such texts beforehand and then using the better English models is a valuable processing step. "IRT2: Inductive Linking and Ranking in Knowledge Graphs of Varying Scale" is concerned with knowledge graph completion based on natural language text and thus covers the *text mining* task. The authors propose two models for predicting links in knowledge graphs and provide initial results based on an experimental evaluation. The third paper "Explaining Hate Speech Classification with Model-Agnostic Methods" proposes a pipeline to create interpretable explanations for black-box models like BERT classifiers. By leveraging *text generation*, they aim to assist users in detecting hate speech in a model-agnostic way. Finally, "Comparing Unsupervised Algorithms to Construct Argument Graphs" is concerned with generating relations between argumentative statements that are extracted from plain texts.

3. Keynote Talk

A keynote talk titled "*Detect—Verify—Communicate: Combating Misinformation with More Realistic NLP*" was given by Prof. Dr. Iryna Gurevych (Technical University (TU) of Darmstadt, Germany), who addressed the omnipresent problem of misinformation in our society, and elaborated on how to identify and debunk misinformation. In particular, she addressed the fact that current NLP systems cannot be used for fact-checking in real-life scenarios and presented her work where they are exploring solutions for this. While the main problems reside with resources, their research is dedicated to the most harmful, novel examples. Besides two corpora constructed for this purpose, they compare the capabilities of automated NLP-based approaches to the methods people use for fact-checking. In order to include different perspectives, they collaborate with cognitive scientists and psychologists.

4. Invited Talks

At the workshop, we also had three invited talks to spark a discussion and exchange of knowledge and ideas. We selected three previously published papers at top-tier conferences (EACL, SIGIR, and WWW) that cover the topics of text mining and generation [1, 2, 3]. In their talk, Adrian Ulges presented a multi-task approach for entity-level relation extraction from texts that combines entity mention localization with coreference resolution [3]. Further, Milad Alshomary presented a query-independent graph-based extractive summarization approach for argumentative web documents [1] that utilizes the PageRank algorithm [4] to estimate the importance of each sentence in arguments retrieved from a given web document. Finally, Wei-Fan Chen talked about a query-biased abstractive summarization approach for snippet generation [2] and showed that their bidirectional model based on pointer-generator networks could generate fluent snippets with low text reuse from the source document while preserving query terms.

5. Conclusion and Discussion

In the first Text Mining and Generation Workshop, a total of four papers were accepted for publication that addressed various topics such as fake news, knowledge graphs, explanations, and relation generation in the context of text mining and generation. In addition, our keynote and invited talks allowed students and young scientists to connect with experienced researchers, ask questions, and participate in discussions.

We expect the topics of this workshop will be becoming increasingly important in the future, particularly with the emergence of new large language models (LLMs). The significant advancements in (retrieval-augmented) text generation introduced by LLMs like GPT-3, ChatGPT, and BLOOM make it crucial that we extensively study their impact on downstream tasks and benchmark key findings. On the one hand, for example, GPT-3 has been found to outperform all supervised models on the news summarization task [5]. On the other hand, LLMs trained on the data that contains misinformation may tend to repeat it [6]. This highlights the need for efficient methods that leverage text mining to extract information from structured (credible) documents, "guide" generation, and provide the sources of evidence. Moreover, new evaluation methodologies are needed that consider the *purpose* and *suitability* of generated texts, not just their similarity to the ground truth. In future iterations of our workshop, we will explore this along with text mining and generation techniques to gain a comprehensive understanding of LLMs' potential applications and limitations.

Acknowledgments

This work has been funded by the DFG within the projects "ReCAP-II" (No. 375342983) and "ACQuA 2.0: Answering Comparative Questions with Arguments" (No. 376430233) as part of the priority program RATIO (SPP-1999) as well as the "Studienstiftung".

References

[1] M. Alshomary, N. Düsterhus, H. Wachsmuth, Extractive snippet generation for arguments, in: Proceedings of SIGIR 2020, ACM, 2020, pp. 1969–1972.

[2] W. Chen, S. Syed, B. Stein, M. Hagen, M. Potthast, Abstractive snippet generation, in: Proceedings of WWW 2020, ACM / IW3C2, 2020, pp. 1309–1319.

[3] M. Eberts, A. Ulges, An end-to-end model for entity-level relation extraction using multi-instance learning, in: Proceedings of EACL 2021, ACL, 2021, pp. 3650–3660.

[4] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web, Technical Report, Stanford InfoLab, 1999.

[5] T. Goyal, J. J. Li, G. Durrett, News summarization and evaluation in the era of GPT-3, CoRR abs/2209.12356 (2022).

[6] S. Lin, J. Hilton, O. Evans, TruthfulQA: Measuring how models mimic human falsehoods, in: Proceedings of ACL 2022, ACL, 2022, pp. 3214–3252.