Observations in Students' Theses – A Critical Analysis of Use-Cases, Models and Problems in Natural Language Processing

Natural Language Processing (NLP) has seen increased interest, not only due to the rise of Large Language Models (LLMs) such as ChatGPT. Scholarly work as well as advances in the private sector are majorly based on models trained on huge datasets and requiring immense processing power, especially in recent years (Sharir et al. 2020). Research groups and students, researching and implementing systems in smaller projects or for their final theses, were benefitted as well as restricted due these developments in NLP. This paper will present observations gathered across several bachelor's and master's theses from 2020 until 2023.

The theses were supervised by the Chair of Media Informatics at the University of Bamberg. Since no dedicated chair of NLP was present in Bamberg at the time, a wide-ranging set of topics including text classification/summarization/simplification, information extraction/retrieval, research data management, generation of misinformation and more were submitted. The authors of those theses also represent a diverse group of students, majoring in study subjects such as computer science, information systems as well as computing in the humanities.

ChatGPT and LLMs have changed the NLP-landscape since the announcement of ChatGPT in 2022. The capabilities of LLMs are in many cases and applications demonstrably impressive. Problems such as hallucinations in the generated texts or reproducibility persist and even with advancements by using larger models, or the development of competing products such as Bing Chat or Bard, could not be solved adequately (Augenstein et al 2023, Sallam 2023). Additionally, the execution of state-of-the-art models is often only available via subscriptions and therefore prohibitive for students or projects with limited budgets.

Thus, a look at more traditional and sometimes simpler models was not only imaginable but sometimes necessary to comply with resource restrictions. We use the terms traditional or simpler models as, for example, support vector machines (SVMs) when compared to large neural networks, or extractive summarization when compared to abstractive/generative summarization, or even rule-based approaches. Due to the long history of such well-established techniques, documentation as well as instructions for setup, customization for the respective use-case and evaluation are widely available and often of high quality. Our students, who were usually not experts in NLP-approaches, needed clear directions as well as understandable models. Traditional models requiring only few dependencies and no large development environments were well suited for beginners.

A wide-ranging collection of NLP-applications were developed in the supervised theses. Hallucinations were apparent in students' work across text simplification and text generation in the application scenario of misinformation (Fruth 2022, Matschat 2022). Both theses used generative models based on the transformer architecture (Vaswani et al. 2017). Grammatical errors were also present in the texts generated in the same theses, due to the generative nature of the models. An even bigger problem was present because of the prevalence of the English language in NLP-research (Khurana et al. 2023). There are multi-lingual as well as language-specific models targeting other languages, but the performance when compared to their English counterparts is often inferior (Aumiller et al. 2023). This was noticed in nearly all theses that were working with German texts (Fruth 2022, Matschat 2022, Pulver 2023, Raab 2023). The size of training datasets plays a role here, as well as adaptations that are necessary for languages other than English, when customizing an approach and evaluation metrics originally intended for English to the specifics of other languages.

Two theses will be presented briefly to showcase specific observations, from which we have drawn our conclusions. The first dealt with information extraction from a large database with an industry partner, incorporating information on different companies (Pulver 2023). The goal was to identify the sector of each company and its legal form, e.g., "GmbH" – German for limited liability company, from the official name of the companies. After first experimenting with neural nets to extract both meta-information, a rule-based approach was used to extract the legal form. The variations for declaring the legal form are finite and can thus be implemented using rules with higher accuracy than approaches using neural nets, showing the advantages of traditional models in this case.

The second thesis focused on text classification on domain-specific languages and on domains with little training data, in this case legal and insurance data, again with an industry partner (Raab 2023). Models based on the transformer architecture were used here once more, however SVMs were evaluated as a point of comparison and matched the performance of modern models, thus showing the relevant position of traditional models even in this use-case of text classification.

Bibliography

Augenstein, Isabelle, et al. "Factuality challenges in the era of large language models." arXiv preprint arXiv:2310.05189 (2023).

Aumiller, Dennis, Jing Fan, and Michael Gertz. "On the State of German (Abstractive) Text Summarization." arXiv preprint arXiv:2301.07095 (2023).

Fruth, Leon. „An Approach Towards Unsupervised Text Simplification on Paragraph-Level for German Texts“. Master thesis, Media Informatics, University of Bamberg (2022).

Matschat, David. „Automatic Generation of Misinformation - Natural Language Generators in the Application Scenario of Wikipedia“. Master thesis, Media Informatics, University of Bamberg (2022).

Pulver, Niklas. „Information Extraction from Company Names Using Current Text Mining Techniques“. Master thesis, Media Informatics, University of Bamberg (2023).

Raab, Stefan Wolfgang. „Few Shot Text Classification for Domain Specific Languages“. Master thesis, Media Informatics, University of Bamberg (2023).

Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." Multimedia tools and applications 82.3 (2023): 3713-3744.

Sallam, Malik. "The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations." medRxiv (2023): 2023-02.

Sharir, Or, Barak Peleg, and Yoav Shoham. "The cost of training nlp models: A concise overview." arXiv preprint arXiv:2004.08900 (2020).

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).