Tatiana Anikina

German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus, Germany
tatiana.anikina@dfki.de
www.anikina.de

1 Research Interests

As a computational linguist I have been working on a variety of natural language processing (NLP) topics including coreference resolution, dialogue processing, explainability, and synthetic data generation. My earlier research includes e.g. the topic of **coreference resolution** in Abstract Meaning Representation which was based on my Master's thesis (Anikina et al., 2020). I have also participated in multiple shared tasks on coreference resolution in dialogue (Anikina et al., 2021, 2022; Skachkova et al., 2023), dialogical argument mining (Binder et al., 2024), and multilingual claim normalization (Anikina et al., 2025b).

My current research focuses mostly on efficient data augmentation strategies and low-resource scenarios. For instance, in my submission to the Student Research Workshop at ACL (Anikina, 2023) I focused on the task of adapting NLP models to **dialogue processing in the emergency response domain**. Having a system for dialogue intent classification and slot tagging that is robust, flexible, and data efficient is a challenging task, especially given that the emergency response domain must cover a wide variety of scenarios (e.g. fires, floods, building collapse etc.) and the available annotations are typically very scarce. In this work I evaluated different approaches towards using adapters with Transformer models and applied different **data augmentation** techniques in a **low-resource setting**, e.g. backtranslation and token replacements as well as contextual augmentation with speaker information, previous turns and dialogue summary.

Another recent work that has been recently accepted at EMNLP (Anikina et al., 2025a) compares various generation strategies for producing synthetic data in low-resource languages. Here we performed a systematic evaluation of demonstrations in different languages, label-based summaries, and self-revision that can be used for **synthetic data generation**, and evaluated the performance of different prompting strategies across 11 typologically diverse languages, including several extremely low-resource ones such as Welsh and Azerbaijani. This work shows that it is possible to achieve very good performance when fine-tuning models on the synthetic data but such data should be carefully selected and ideally revised by language models.

Currently, I am doing a research internship in the JetBrains AI Writing Assistance team, preparing a dataset with synthetic data for training a commit completion model. The goal of this project is to find an optimal way of **generating new good quality data for commit messages** that describe the code changes. Synthetic data are needed here since the existing messages collected from the sources like *GitHub* are often noisy, ungrammatical or they may not fully reflect all the relevant changes (e.g. "fix the bug" is not a good commit message compared to "fix the logic of error handling in ProjectClass"). Ideally, good messages would be fluent, coherent with the commit history and correct with respect to the code diff. Together with the team I have been experimenting with various models and approaches, including code diff summarization, chain-of-thought prompting (Fu et al., 2022), and different ways of selecting the demonstrations.

2 Data Generation and Low-resource Scenarios

Modern Large language models (LLMs) such as *ChatGPT*, *Llama*, *Qwen*, *Gemini*, and many others demonstrate impressive capabilities of producing natural language text that is fluent and typically follows the user instruction encoded in a prompt. Such LLMs are often used to generate new data for training smaller downstream models in an efficient way, they can be used for both data augmentation and generation (Dai et al., 2023; Piedboeuf and Langlais, 2023; Cegin et al., 2025). LLM-generated texts can be useful for a variety of tasks from intent recognition (Cegin et al., 2024) and news classification (Piedboeuf and Langlais, 2023) to hate speech detection (Sen et al., 2023) and sentiment analysis (Onan, 2023). Increasingly more complex tasks can be tackled by LLMs, Pranida et al. (2025) compared the data that were generated by humans, with machine translation, and LLM-generated texts for culturally nuanced commonsense reasoning in low-resource languages and found that LLM-generated data were more beneficial for downstream classification than machine translation.

However, LLMs are mostly trained on the English data and even though some multilingual versions exist, e.g. *Aya* or *Gemma 3*, they do not cover many languages, es-

Proceedings of the 1st Workshop for Young Researchers in Natural Language Generation, pages 7–9

pecially the low-resourced ones. Although some works explore generation also for low-resource languages (Zotova et al., 2021; Chung et al., 2025) the general trend is to expect lower quality of the synthetic data for those languages that were unseen (or barely seen) during training and more research needs to be done on finding the optimal generation strategies for such languages, potentially combining cross-lingual transfer, demonstration-based prompting, and LLM revision, e.g. expanding on the optimal strategies identified in (Anikina et al., 2025a).

Another interesting and underexplored venue of research is the low-resource setting with respect to the task, not necessarily the language. For instance, Ben Abacha et al. (2023) work towards summarization of doctor-patient conversations in a clinical setting. They found that the generated summaries show high fluency score, but still struggle with hallucinations and miss 33% important medical facts. Although using synthetic data like automatically generated summaries of doctor-patient conversations is very attractive as it can reduce the documentation burden for medical professionals, there is still a large room for improvement. Another low-resource domain where synthetic data could be especially valuable is emergency response. Privacy constraints and the wide variety of possible scenarios make real data scarce, yet dialogue between first responders still needs to be summarized and classified. To enable this, LLMs must first be trained on synthetic datasets that capture diverse emergency situations and incorporate the relevant terminology.

3 Questions and Discussion Topics

I would like to suggest the following discussion topics and questions to senior researchers:

- **Evaluation:** What are the best was of evaluating the quality of generated data? Often we do not have the gold standard data that the generated data can be compared to and have to resort to either reference-free metrics or rely solely on the performance of a downstream model that was trained on the generated data. Nowadays researchers also use various LLMs to score the responses, applying techniques like e.g. LLM-as-a-Judge (Wang et al., 2024) but how reliable are these judgments? How can we estimate and avoid potential bias?

- **Controlled Generation:** What are the best practices in generating outputs conditioned on specific knowledge or inputs? In many real life situations there are some initial requirements and constraints on the data that need to be respected during the generation. How can we integrate such constraints and check that they are followed in an efficient way?

- **Synthetic Data Recycling:** An increasing amount of web content is now generated by language models. As new training datasets are collected for the next generation of models, it is inevitable that some of this synthetic content will be included. How problematic is this phenomenon, and what long-term consequences might it have, e.g., could it reduce the diversity and quality of model outputs? To what extent is recycling synthetic data a viable strategy, and when should it be avoided?

Biographical Sketch

I hold an M.Sc. in Language Science and Technology from Saarland University. After graduation, I joined the German Research Center for Artificial Intelligence (DFKI) as a PhD student and researcher. Currently, I am on leave from my PhD program to pursue an internship at JetBrains, where I am part of the AI Writing Assistance team. My research interests include efficient data augmentation, dialogue processing, multilingual and low-resource scenarios.

References

Tatiana Anikina. 2023. Towards efficient dialogue processing in the emergency response domain. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 212–225, Toronto, Canada. Association for Computational Linguistics.

Tatiana Anikina, Ján Cegin, Jakub Simko, and Simon Ostermann. 2025a. A rigorous evaluation of llm data generation strategies for low-resource languages. *ArXiv*, abs/2506.12158.

Tatiana Anikina, Alexander Koller, and Michael Roth. 2020. Predicting coreference in Abstract Meaning Representations. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–38, Barcelona, Spain (online). Association for Computational Linguistics.

Tatiana Anikina, Cennet Oguz, Natalia Skachkova, Siyu Tao, Sharmila Upadhyaya, and Ivana Kruijff-Korbayova. 2021. Anaphora resolution in dialogue: Description of the DFKI-TalkingRobots system for the CODI-CRAC 2021 shared-task. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 32–42, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tatiana Anikina, Natalia Skachkova, Joseph Renner, and Priyansh Trivedi. 2022. Anaphora resolution in dialogue: System description (CODI-CRAC 2022 shared task). In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 15–27, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Tatiana Anikina, Ivan Vykopal, Sebastian Kula, Ravi Kiran Chikkala, Natalia Skachkova, Jing Yang, Veronika Solopova, Vera Schmitt, and Simon Ostermann. 2025b. dfkinit2b at checkthat! 2025: Leveraging llms and ensemble of methods for multilingual claim normalization. In *CLEF 2025 Working Notes. Conference and Labs of the Evaluation Forum (CLEF-2025), Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, September 9-12, Madrid, Spain*. CEUR Workshop Proceedings.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Arne Binder, Tatiana Anikina, Leonhard Hennig, and Simon Ostermann. 2024. DFKI-MLST at DialAM-2024 shared task: System description. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 93–102, Bangkok, Thailand. Association for Computational Linguistics.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2024. Use random selection for now: Investigation of few-shot selection strategies in llm-based text augmentation for classification. *Preprint*, arXiv:2410.10756.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2025. LLMs vs established text augmentation techniques for classification: When do the benefits outweight the costs? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10476–10496, Albuquerque, New Mexico. Association for Computational Linguistics.

Yi-Ling Chung, Aurora Cobo, and Pablo Serna. 2025. Beyond translation: Llm-based data generation for multilingual fact-checking. *arXiv preprint arXiv:2502.15419*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *Preprint*, arXiv:2302.13007.

Yao Fu, Hao-Chun Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *ArXiv*, abs/2210.00720.

Aytuğ Onan. 2023. Srl-aco: A text augmentation framework based on semantic role labeling and ant colony optimization. *Journal of King Saud University - Computer and Information Sciences*, 35(7):101611.

Frédéric Piedboeuf and Philippe Langlais. 2023. Is ChatGPT the ultimate data augmentation algorithm? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615, Singapore. Association for Computational Linguistics.

Salsabila Zahirah Pranida, Rifo Ahmad Genadi, and Fajri Koto. 2025. Synthetic data generation for culturally nuanced commonsense reasoning in low-resource languages. *Preprint*, arXiv:2502.12932.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Natalia Skachkova, Tatiana Anikina, and Anna Mokhova. 2023. Multilingual coreference resolution: Adapt and generate. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 19–33, Singapore. Association for Computational Linguistics.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024. Self-taught evaluators. *CoRR*, abs/2408.02666.

Elena Zotova, Rodrigo Agerri, and German Rigau. 2021. Semi-automatic generation of multilingual datasets for stance detection in twitter. *Expert Systems with Applications*, 170:114547.