
Generation of German Text Controlled by Sentiment and Keywords

Paulina Aleksandra Żal

Master's Thesis

Master of Science in Applied Information and Data Science

School of Business

Lucerne University of Applied Sciences and Arts

Wettingen AG, 22nd of December 2023

Author:	Paulina Aleksandra Żal paulina.zal@stud.hslu.ch
Supervisor:	Dr. Guang Lu guang.lu@hslu.ch
Co-Supervisor:	Dr. Nianlong Gu nianlong@ini.ethz.ch

Management Summary

The introduction of transformer models into Natural Language Processing (NLP) led to Pre-trained Language Models (PLM), which are constantly becoming more attractive for a variety of today’s use cases. Due to their power, these models can be employed in Controllable Text Generation (CTG) to exert precise control over attributes like sentiment or topic, advancing text customization in various applications. While proficient, the probabilistic nature of those models may lead to outputs not precisely aligned with user intent. Emerging methods such as Fine-Tuning and Plug-and-Play Models (PPLMs) empower users to steer text generation, aiming for more tailored outputs, particularly in marketing and content creation domains.

This thesis contributes by refining text generation techniques, ensuring outputs adhere to predefined guidelines related to both topic and sentiment. The aim is to modify a pre-trained GPT-2 model in German, enabling it to produce paragraphs derived from an input sequence, a sentiment token, and a specific keyword set.

The pipeline consists of Supervised Fine-Tuning (SFT), optimization with Reinforcement Learning (RL), and a Logit Modification mechanism. SFT is utilized to adapt the GPT-2 model to generate text concerning the specified sentiment token. RL optimizes the model and improves the text generation in terms of sentiment. The model is rewarded by scores from a Sentiment Discriminator that was previously created. In the last step a logit modification mechanism from Pascual et al. (2020) is adapted and tested with different decoding strategies.

SFT and RL show promising results in generating sentiment-based text. However, implementing keyword control mechanisms reduces the performance in terms of keyword utilization compared to the GPT-2. The human assessment indicates that texts generated with sentiment tokens exhibit moderate fluency and coherence. The introduction of keyword control negatively impacts fluency and coherence. Using the method may save time and effort during the creative process of content creation. However, the proposed pipeline is not yet suitable for practice, as it faces some limitations in generating coherent and fluent texts with a specific set of keywords. The author believes that keyword control could be improved within the SFT and RL processes to achieve a balanced emphasis on both sentiment and keywords in Fine-Tuning models.

Keywords. Natural Language Processing, Controllable Text Generation, Pre-trained Language Models, Supervised Fine-Tuning, Reinforcement Learning

Contents

List of Figures	6
List of Tables	7
List of Code	9
List of Abbreviations	10
Acknowledgements	11
1 Introduction	12
1.1 Background	12
1.2 Topic Definition and Thesis Objective	13
1.3 Related Work	14
1.4 Research Questions	15
1.5 Structure of this Thesis	16
2 Theoretical Fundamentals	16
2.1 Transformers Architecture	16
2.1.1 Attention Mechanism	18
2.1.2 Encoder	19
2.1.3 Decoder	19
2.2 Pre-trained Language Models	20
2.2.1 GPT	20
2.2.2 BERT	20
2.3 Text Generation	21
2.3.1 Controllable Text Generation	22
2.3.2 Plug-and-Play Language Models	23
2.4 Decoding Strategies	24
2.4.1 Greedy Search	24
2.4.2 Top- k	24
2.4.3 Top- p	24
2.4.4 Beam Search	25
2.5 Reinforcement Learning	26
3 Research Design and Data	27
3.1 Workflow	27
3.2 Data	28
3.2.1 Dataset Description	28

3.2.2	Data Cleaning	29
3.3	Automated Evaluation Metrics	30
3.3.1	Classification Metrics	30
3.3.2	Perplexity	31
3.3.3	SLOR	32
3.3.4	Flesch Reading Ease	32
3.3.5	Coherence Score	32
3.3.6	Success Rate	33
3.4	Sentiment Classification	33
3.5	Fine-Tuning	34
3.6	Proximal Policy Optimization	34
3.7	Logits Modification Mechanism	36
3.8	Human Evaluation	37
4	Experiments	38
4.1	Data Combination and Preprocessing	39
4.2	Sentiment Discriminator	39
4.2.1	Data Processing	39
4.2.2	BERT model with Convolutional Neural Network	40
4.2.3	BERT model with Classification Head	41
4.3	Sentiment Control	42
4.3.1	Supervised Fine-Tuning	42
4.3.2	Reinforcement Learning Optimization	43
4.4	Keyword Control	46
4.4.1	Logit Modification Mechanism	46
4.4.2	Decoding Strategies	47
4.5	Survey Design	48
5	Results	50
5.1	Performance of Sentiment Discriminator	51
5.2	Evaluation of Sentiment Control	51
5.2.1	Evaluation of Supervised Fine-Tuning	52
5.2.2	Results of Reinforcement Learning	54
5.2.3	Improvement of Sentiment Control During Training	56
5.3	Performance of Keyword Control	56
5.3.1	Evaluation Based on Nouns	56
5.3.2	Evaluation Based on Sentiment-Carrying Adjectives	58
5.3.3	Evaluation of Different Control Inputs	59

5.4	Analysis of Survey Results	60
5.4.1	Selection of Models	60
5.4.2	Demographics	60
5.4.3	Inter-Annotator Agreement	61
5.4.4	Evaluation of Cronbach Alpha	61
5.4.5	Evaluation of Mutual Influence of Keyword and Sentiment Control	61
5.4.6	Influence of Fine-Tuning on Sentiment Control	63
5.4.7	Influence of Decoding Strategy and Keyword Control	64
6	Discussion	65
6.1	Research Question 1	66
6.2	Research Question 2	67
6.3	Research Question 3	68
6.4	Research Question 4	69
7	Conclusion and Outlook	70
7.1	Conclusion	70
7.2	Outlook	71
	References	73
	Appendix A Function for Logit Modification	80
	Appendix B Beam Search	82
	Appendix C Examples of Generated Texts	84
	C.1 Texts Generated by Different Models with Negative Sentiment Token .	84
	C.2 Texts Generated by Different Models with Positive Sentiment Token .	85
	Appendix D Survey	87
	Appendix E Texts Evaluated in the Survey	90
	Appendix F Likert Scale Interpretation, Categories and Items	94
	Appendix G Functions for SLOR and Coherence Score	96
	Appendix H Evaluation of Survey's Items	98
	Appendix I Influence of Human Intervention on Text Quality	102
	Declaration of Originality	103
	Declaration of the use of Generative AI	104