Enhancing Topical Relevance in Abstractive Summaries through Reweighting Logits of Topic-Relevant Tokens

Joschka Braun
joschka.braun@student.uni-tuebingen.de

Supervised by Seyed Ali Bahrainian

Abstract

In this project, a LogitsProcessor class is implemented to enhance topical relevance in abstractive summaries through reweighting logits of topic-relevant tokens. Three methods are explored to encourage topic-relevant tokens: Constant Shift, where logits are increased by a constant value; Factor Scaling, where logits are multiplied with a constant factor; and Threshold Selection, where logits are selectively encouraged above a certain probability threshold. These methods are evaluated on the NEWTS dataset against the baseline strategy of prompt engineering with topic-associated keywords.

1 Introduction

Abstractive topical summarization with autoregressive transformer-based language models (LMs) presents significant challenges, particularly when generating topic relevant summaries without extensive model retraining. To generate a high-quality summary of a given text and focussing on a selected topic usually requires in-context learning Dong et al. [2023], if the model was not previously fine-tuned to do so. Any model solely fine-tuned for summarization, or a medium-sized instruction-tuned LM will struggle to generate focused summaries based solely on in-context learning. Fine-tuning models for topical-summarization via methods such as Direct Preference Optimization (DPO) Rafailov et al. [2023] or Reinforcement Learning from Human Feedback (RLHF) Ouyang et al. [2022] while effective, are technically challenging, require extensive annotated data and frequently fail to maintain consistent effectiveness across different topics and articles. Therefore, the project aims to establish a more easily implementable, universally applicable and resource efficient method that improves the topical focus of summaries without extensive training or fine-tuning.

2 Experiment Setup

2.1 Datasets and Model

I evaluate the methods on the NEWTS dataset by Bahrainian et al. [2022], designed specifically for topical summarization. The NEWTS dataset is based on the CNN/DailyMail dataset Nallapati et al. [2016] and made up of 2400 training and 600 test samples. Each sample contains a source article and two reference topical summaries, each focussed on one of the two most prominent topics in the article (tid1 and tid2). The associated LDA model has 250 distinct topics. For the experiments, I use two state-of-the-art transformer-models: the 2 billion-parameter Gemma model by Google GemmaTeam [2024] and the 8 billion-parameter Llama3 model from Meta Llama3Team [2024]. Both models are pre-trained and have been fine-tuned with instructional data.

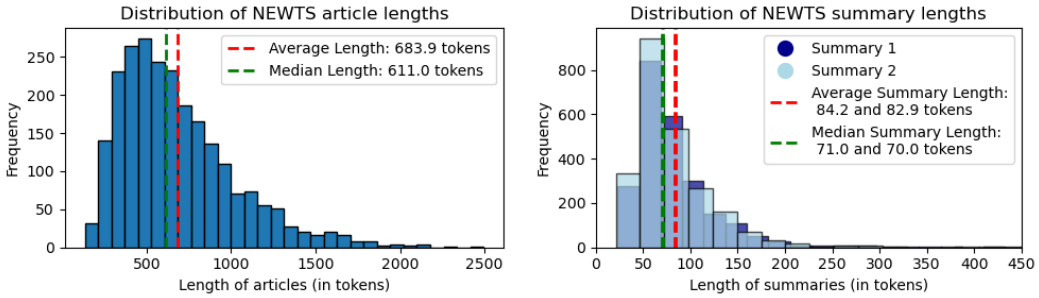

Figure 1: The summaries are roughly 12% as long as the original articles.

2.2 Hyperparameters

For all experiments conducted, the same hyperparameters are used to ensure comparability across different methods and models. For consistency, I selected articles in chronological order from the NEWTS training dataset. The number of articles varied between 25 or 50 based on the size of the model being tested or the use of beam search to accommodate computational constraints and model capabilities. The length of the generated summaries was limited between 80 and 90 tokens, which is roughly similar to the mean summary length of the human-written summaries of the NEWTS dataset. This length was chosen to make summaries comparable to the human-written summaries and balance detail and brevity. For context, the average article length in the dataset is approximately 680 tokens, so summaries are around 12% in length or the original article.

Nucleus sampling was implemented with a top-p parameter of 0.95 to restrict the model's selection to the most likely subsequent tokens. Additionally, the sampling was restricted by a top-k value of 50. In experiments without beam search, generation was solely based on this sampling strategy. For the experiment involving beam search, I utilized a configuration of 4 beams to diversify the exploration of potential summary outputs.

From the LDA model, I always used the top 25 words associated with the topic to represent the topic. To translate from words to tokens, I generated multiple variations of each word, including the lemmatized and stemmed versions, as well as variations in capitalization and spacing. So instead of 1–2 tokens, 3–5 tokens are associated with each word. This step is crucial for covering all possible tokens that should get sampled more likely when the model is generating a summary with enhanced topical focus on that specific word.

Overall, the experiment hyperparameters were chosen to generate high-quality summaries and make generated summaries comparable to the human-written summaries inf the NEWTS dataset. Keeping hyperparameters identical ensures that any observed differences in performance are attributable to the methodological variations rather than differences in experimental conditions.

2.3 Evaluation methods

To evaluate each generated summary for quality, each was compared to its respective reference summary from the NEWTS dataset. And to evaluate for topical focus respective to the chosen topic with respect to the topic of the LDA model. To rigorously assess the performance of the generated summaries, three metrics were employed for each the quality and topical focus of the summary. This dual assessment approach ensures a comprehensive evaluation of how well the summaries adhere to both the linguistic quality and relevance to the specified topic.

2.3.1 Summarization Quality

For summarization quality, I utilized the following metrics:

ROUGE-L Score: ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) Lin [2004] measures the longest common subsequence between the generated summary and a reference summary. It is particularly useful for evaluating the summary for fluency and coherence. I use the implementation from the **'rouge_scorer.RougeScorer'** class and active stemming.

MAUVE Score: MAUVE (Measure of Aggregated Unidirectional Validity and Entropy) Pillutla et al. [2021] quantifies the statistical gap between the human-written and machine-generated summaries using the Kullback-Leibler divergence between the two distributions in the embedding space. The MAUVE score correlates well with human evaluations for text quality. I use the official implementation of MAUVE via a wrapper from the **'evaluate'** library on Hugging Face.
BERTScore: BERTScore Zhang* et al. [2020] leverages the pre-trained contextual embeddings, originally from BERT Devlin et al. [2019] models, to compute the semantic similarity between the generated summary and the reference text. This metric is robust against paraphrasing and variations in word choice, thus offering a nuanced measure of semantic similarity. I employ the **'BERTScorer'** class with the microsoft/deberta-xlarge-mnli model He et al. [2021], as it currently correlates best with human evaluation.
Collectively, these three metrics offer a robust estimate of the summarization quality.

2.3.2 Summary Topical Focus

To evaluate the alignment of generated summaries with the intended topics, I utilize three methods to quantify topical focus:

Lemmatization-Based Evaluation: This method processes the generated text by lemmatizing words, converting them to their canonical form. Using the LDA model, it matches these lemmas against the lemmas of the top topic words identified for the relevant topic. The topical focus score is then calculated as the weighted presence of these lemmas in the summary, normalized by the total weight of topic-specific lemmas.
Tokenizing-Based Evaluation: In this approach, text is tokenized using a model-specific tokenizer, and tokens are matched against those generated from topic-specific words identified by the LDA model. The score is computed based on the proportion of topic tokens in the summary, offering a direct measure of topical vocabulary usage.
Dictionary-Based Evaluation: This method leverages a dictionary approach where each word in the summary is converted into its bag-of-words representation. The LDA model then provides a distribution over topics for these words. The score reflects the prevalence of the relevant topic's tokens within the summary, adjusted by the topic distribution provided by the LDA.

By integrating these three measures, I can estimate the effectiveness of the logits reweighting methodologies in producing topically focused summaries.

3 Abstractive Topical Summarization via Prompt Engineering

3.1 Method

To establish a baseline, I employ prompt engineering to explicitly direct the model's attention to specific topics. The prompt begins with the instruction to generate a summary, which sets the task context for the language model. To direct the model's focus towards the desired topic, the prompt incorporates a specific instruction to concentrate on the topic, delineated by the top 25 words associated with it. These topic-associated words are integrated into the prompt to ensure that the model focuses on the specific topic during the summary generation.

3.2 Results

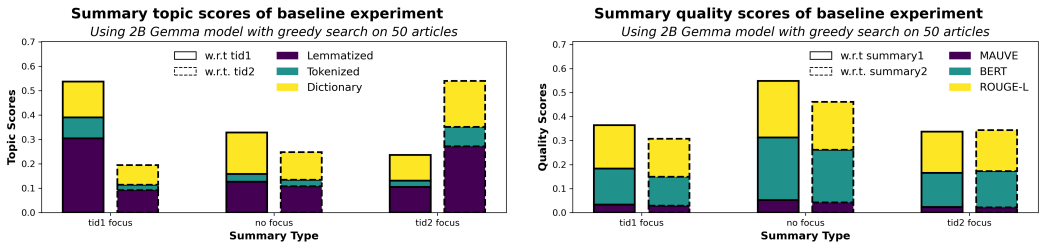

Figure 2: Summary scores for the baseline experiment using Gemma 2B with greedy decoding.

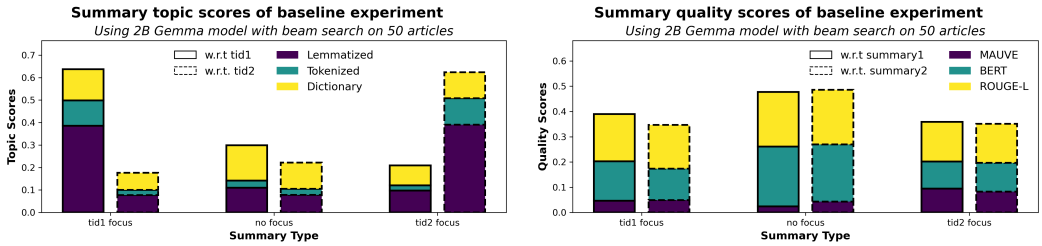

Figure 3: Summary scores for the baseline experiment using Gemma 2B with beam search.

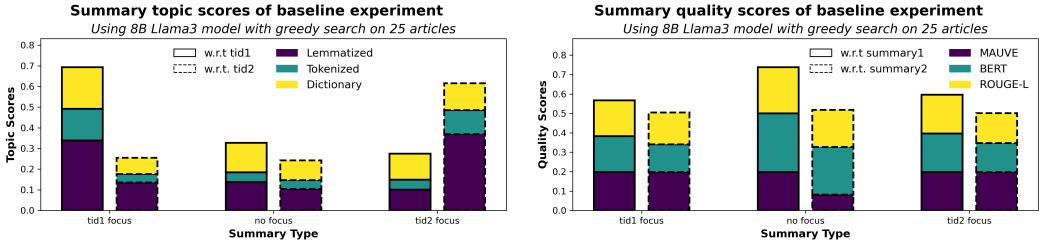

Figure 4: Summary scores for the baseline experiment using Llama3 8B with greedy decoding.

Consistent trends can be observed across both the Llama3 8B and Gemma 2B model with and without beam search. When prompted to focus on a specific topic, the topical scores increased roughly twofold in all three setups, especially the lemmatization and tokenization-based topic scores. When prompting the model to focus on the respective topic 1, the topic score approximately doubled for topic 1 and decreased noticeably for topic 2 compared to the baseline. The same is also true for summaries focussed on topic 2. Changes in topical scores are larger when using beam search compared to greedy decoding, or using the larger Llama3 8B model compared to the smaller Gemma 2B model. In terms of summary quality, prompting the model to focus on a particular topic came at the expense of overall summary quality. Summaries generated without a specific topical focus consistently measured higher scores. When expecting the generated summaries, this was likely caused due to the model reiterating the instructions at the start of the generated summary. This included reiterating words related to the designated topic, a characteristic behavior of instruction-tuned models, which could not effectively mitigated by further prompt engineering. In terms of summary quality, both beam search and a larger model size increased absolute summary quality. These findings underscore the challenge of balancing topical focus with summary quality in instruction-tuned models, and highlight the limitations of prompt engineering in effectively addressing this issue.

4 Abstractive Topical Summarization via Logits Reweighting

4.1 Methods

To manipulate the logits predicted by the model, I have implemented a custom LogitsProcessor class within the Hugging Face transformer framework. This class enables the altering the logits of topic-specific tokens during text generation. Each of the three method modifies the logits in a distinct manner, allowing for experimental comparisons of their effects on the generated summaries.

4.1.1 Constant Shift

The Constant Shift method adds a constant value c to all logits of predefined topic-relevant tokens. This shift uniformly increases or decreases the probability of selecting such tokens, irrespective of their original logits. The modification is implemented as:

$$\text{scores}_{\text{modified}}[i] = \text{scores}[i] + c$$

where i indexes into topic-relevant tokens. This method ensures consistent manipulation of specified tokens across different contexts.

4.1.2 Factor Scaling

Factor Scaling alters the logits of topic-specific tokens by multiplying them by a predetermined scaling factor α. The method can be expressed mathematically as:

$$\text{scores}_{\text{modified}}[i] = \text{scores}[i] \times \alpha$$

where i denotes topic-relevant tokens. This scaling affects logits proportionally, magnifying or diminishing their original values and thus, altering their likelihood of selection.

4.1.3 Threshold Selection

Threshold Selection involves selectively adjusting logits based on a predefined probability threshold θ. Logits are first converted to probabilities using the softmax function. If a topic-relevant token's probability exceeds θ, its logit is increased to the maximum logit in the current logit distribution plus an additional encouragement factor β. This is expressed as:

$$\text{scores}_{\text{modified}}[i] = \begin{cases} \max(\text{scores}) + \beta & \text{if softmax}[i] \geq \theta \\ \text{scores}[i] & \text{otherwise} \end{cases}$$

where softmax[i] is the calculated probability of topic-relevant token i. Threshold Selection only changes token logits that are already likely under the model's current context, enhancing their prominence without affecting less relevant tokens. These methods provide distinct strategies for changing topic relevance during the summary generation process.

4.2 Results

4.2.1 Constant Shift

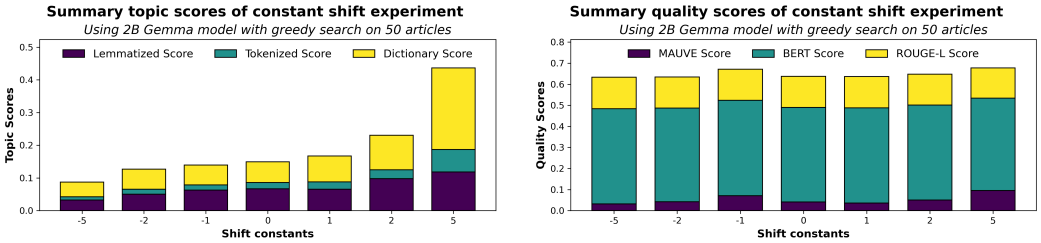

Figure 5: Summary scores for the constant shift experiment using Gemma 2B with greedy decoding.

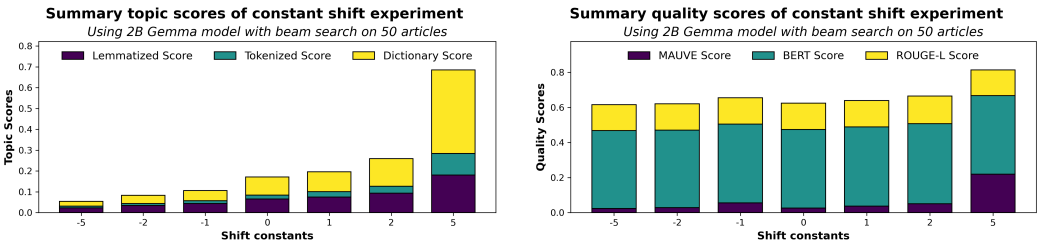

Figure 6: Summary scores for the constant shift experiment using Gemma 2B with beam search.

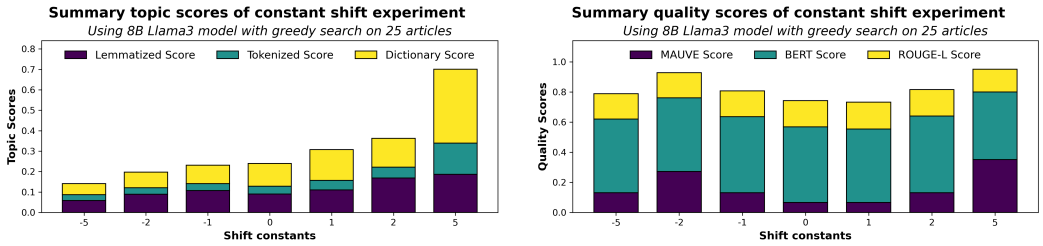

Figure 7: Summary scores for the constant shift experiment using Llama 8B with greedy decoding.

In the Constant Shift experiment, a consistent pattern emerged across both the Llama and Gemma model with greedy decoding and beam search. Adjusting topic-relevant token logits by adding constant values significantly impacted topical scores measured by lemmatization, tokenization, and dictionary-based evaluations. Specifically, negative constants decreased topical scores, while positive values increased them across all metrics. This effect was more pronounced with beam search and in the larger Llama 3 8B model compared to the Gemma 2B model. ROUGE-L and BERT scores remained stable despite these modifications; however, MAUVE scores showed considerable variability, peaking notably at a shift constant of 5. This variance underscores the sensitivity of the MAUVE metric to changes in topical emphasis. Despite this, the Llama model consistently demonstrated superior summary quality. As tested in other experiments, if the shift constant exceeds 10, summaries are made up almost exclusively out of topic-relevant tokens, severely compromising summary quality in terms of fluency and coherence. This underscores, that while topical encouragement via constant shift is successful, the shift constant should not be chosen too large.

4.2.2 Factor Scaling

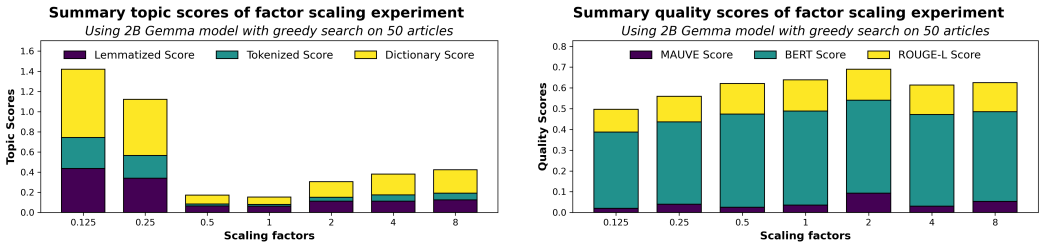

Figure 8: Summary scores for the factor scaling experiment using Gemma 2B with greedy decoding.

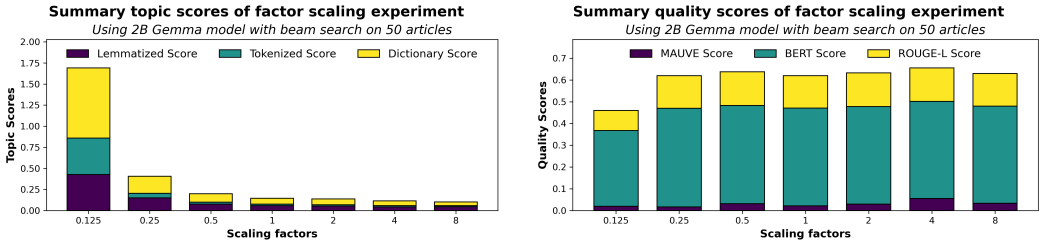

Figure 9: Summary scores for the factor scaling experiment using Gemma 2B with beam search.

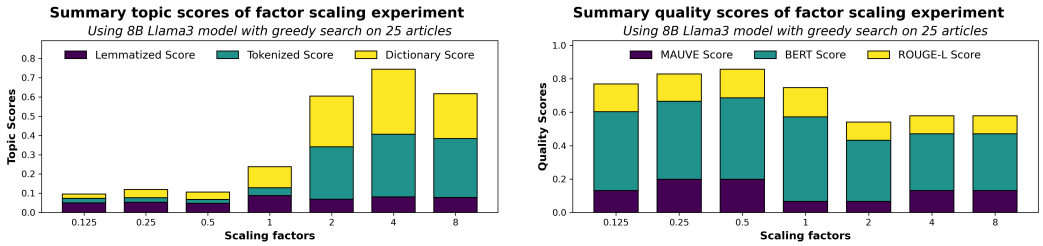

Figure 10: Summary scores for the factor scaling experiment using Llama 8B greedy decoding.

Factor Scaling exhibited varying effects across the Llama and Gemma model, influenced by the sign of their logits. For the Gemma model, where logits are predominantly negative, scaling factors below 1 effectively increased the logits, making them less negative and thus more likely to be chosen. Conversely, for positive logits in the Llama model, factors greater than 1 increased logits, enhancing their probability of selection. The use of beam search notably enhanced topical scores only when factors were below 1. Across both models, however, increased topical scores generally correlated with a reduction in summary quality as measured by ROUGE-L, BERTScore, and MAUVE. These results suggest that the impact of scaling factors is contingent on the initial sign of the logits, pointing to the potential benefit of dynamically adjusting factors based on the logit sign for more consistent outcomes across different models.

4.2.3 Threshold Selection

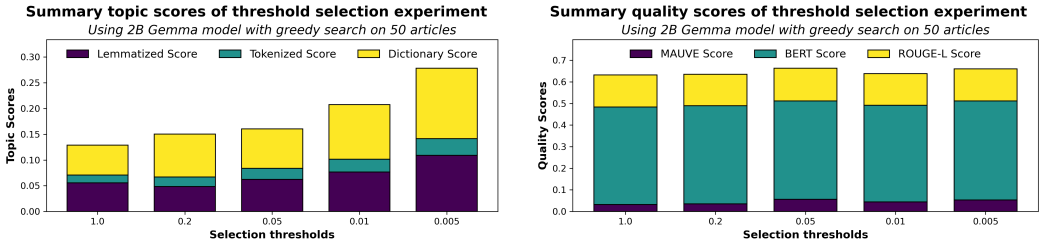

Figure 11: Summary scores for the threshold selection experiment using Gemma 2B with greedy decoding.

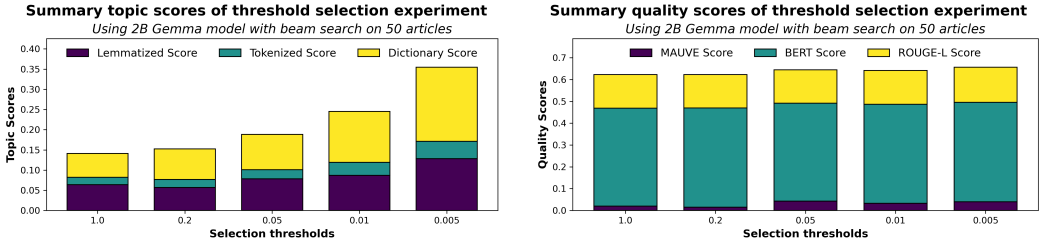

Figure 12: Summary scores for the threshold selection experiment using Gemma 2B with beam search.

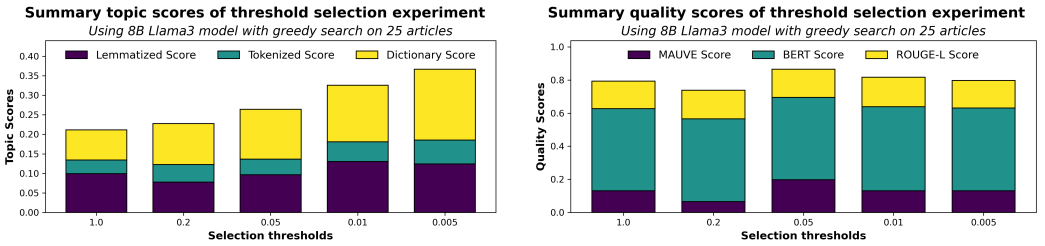

Figure 13: Summary scores for the threshold selection experiment using Llama 8B with greedy decoding.

Consistent trends were observed in the Threshold Selection experiments across both models and decoding strategies, similar to those seen in the Constant Shift setup. Lowering the probability threshold consistently led to higher topic scores, an effect amplified by beam search because it favors beams with a higher concentration of topic-relevant tokens. The Llama model consistently outperformed the Gemma model in achieving higher summary topic and quality scores. Unlike the Factor Scaling method, increased topical focus through threshold adjustment did not adversely affect summary quality across ROUGE-L, BERTScore, and MAUVE metrics. This indicates that the method of setting topical tokens to the maximum logit value within the current distribution, combined with controlled sampling (top-p and top-k), allows for substantial topical emphasis without sacrificing summary quality.

5 Conclusions

In this project, I investigated the effectiveness of logits reweighting techniques - Constant Shift, Factor Scaling, and Threshold Selection - in generating topical abstractive summaries with transformer-based language models. The experiments conducted on the NEWTS dataset show that reweighting logits is a viable method to focus a summary on a desired topic without the need for model retraining.

Among the tested methods, Threshold Selection emerged as particularly suitable because its probability selection threshold is relatively independent of the logit distribution of the model. This is unlike the Constant Shift and Factor Scaling method, whose hyperparameters require more tuning based on the specific logit distribution of each model. This makes Threshold Selection a preferable choice, as it offers a more universally applicable and less model-dependent approach. Additionally, the use of beam search with Threshold Selection further amplifies its effectiveness, suggesting that this combination may be optimal for improving topical focus in abstractive summaries.

Looking forward, experimenting with different settings of the encouragement factor β and integrating temperature scaling with Threshold Selection could potentially refine the method and yield even better results. These adjustments could help achieve even stronger topical focus and without degrading summary quality.

Overall, my findings underscore the potential of logits reweighting as a resource-efficient alternative to more complex model fine-tuning methods, providing a practical solution for improving the topical focus of generated abstractive summaries.

References

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=HPuSIXJaa9`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton,

Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.42. URL https://aclanthology.org/2022.findings-acl.42.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Stefan Riezler and Yoav Goldberg, editors, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL https://aclanthology.org/K16-1028.

GemmaTeam. Gemma: Open models based on gemini research and technology, 2024.

Llama3Team. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/, April 2024. Accessed: 2024-04-22.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Tqx7nJp7PR.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.