Sequential Transfer Learning in NLP for German Text Summarization

Pascal Fecht
inovex GmbH
76131 Karlsruhe
pfecht@inovex.de

Sebastian Blank
inovex GmbH
76131 Karlsruhe
sblank@inovex.de

Hans-Peter Zorn
inovex GmbH
76131 Karlsruhe
hzorn@inovex.de

Abstract

This work examines the impact of sequential transfer learning on abstractive machine summarization. A current trend in Natural Language Processing (NLP) is to pre-train extensive language models and then adapt these models to solve various target tasks. Since these techniques have rarely been investigated in the context of text summarization, this work develops an approach to integrate and evaluate pre-trained language models in abstractive text summarization. Our experiments suggest that pre-trained language models can improve summarizing texts. We find that using multilingual BERT (Devlin et al., 2018) as contextual embeddings lifts our model by about 9 points of ROUGE-1 and ROUGE-2 on a German summarization task.

1 Introduction

Summarizing is the ability to write a brief abstract of the essential content given in a text. Two types of approaches for automatic summarization systems can be distinguished. *Extractive methods* aim to identify the crucial information of a written text and solely copy these parts as summary (Conroy and O'leary, 2001; Shen et al., 2007). On the other hand, *abstractive methods* aim to express the summaries as coherent and fluent texts (Rush et al., 2015; Nallapati et al., 2016). This work focuses on abstractive methods with deep neural networks.

A summarization system, however, is optimized for the objective of a *single* task only. In order to be able to reuse previously learned knowledge, *transfer learning* methods share beneficial information across multiple tasks. Recently, various approaches (Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2018) in sequential transfer learning (Ruder, 2019) have lead to improvements in a wide range of tasks in NLP by extensively pre-training a language model (LM) and adapting the model for specific tasks.

Hence, this work develops an approach based on a deep neural model for abstractive summarization that applies recent advances for the task of text summarization. Therefore, our model is evaluated on a German dataset extracted from 100,000 German Wikipedia articles.

2 Related work

In sequential transfer learning (Ruder, 2019), two arbitrary tasks are learned in sequence. During pre-training on the source task, the objective is commonly very generic with large data and high computational costs. An established approach is the adaptation of pre-trained word embeddings (Mikolov et al., 2013; Pennington et al., 2014) to several target tasks. However, one shortcoming of these embeddings is that they are context-free, meaning that their representation of words are identical in any context.

An early approach with deep neural networks incorporates context into embeddings by using the encoder of a machine translation system with shallow RNNs (McCann et al., 2017). ELMo (Peters et al., 2018) generalizes this approach by pre-training a language model (LM) and extracting its features as contextual embeddings. Subsequent contributions like GPT (Radford et al., 2018), BERT (Devlin et al., 2018) or GPT-2 (Radford et al., 2019) replace the shallow RNNs in LMs with Transformers (Vaswani et al., 2017) resulting in deep representations. Further, these approaches do not only extract the features of language models but fine-tune the entire model for several classification tasks.

Recent work in abstractive text summarization is commonly based on encoder-decoder models with RNNs and additional attention (Nallapati et al., 2016; See et al., 2017). Furthermore, pointer-generator networks (Gu et al., 2016; See et al., 2017) copy tokens from the source document to generated summaries. This addresses the problem of summarization systems which tend to produce many out-of-vocabulary (OOV) words during inference. Another known issue of summarization systems is the repetition of words and sequences of words in generated summaries. The *coverage vector* (Tu et al., 2016) addresses this by tracking and controlling the covered and uncovered parts of the source document (See et al., 2017). Finally, Paulus et al. (2017) apply *policy-gradient learning* (Rennie et al., 2016) in order to use the ROUGE as auxiliary learning objective to dedicatedly measure the quality of generated summaries.

3 Summarization model

Our abstractive summarization system is designed as an encoder-decoder model with attention (Bahdanau et al., 2014) and integrates a copy mechanism (Gu et al., 2016) to reduce OOV words in the generated summaries.

Encoder Given s words $u_1, \cdots, u_s$ in an input document, the words are embedded as $x_1, \cdots, x_s$ in the first layer. Subsequently, the encoder processes each embedding x_i at timestep i to a hidden state $\bar{h}_i$. More specifically, the encoder is a multi-layer multi-head-attention Transformer (Vaswani et al., 2017). The set of all encoder hidden states is referred to as the memory $\mathcal{M} = \{\bar{h}_1, \cdots, \bar{h}_s\}$ and accessed during decoding. In a similar notion to fully LSTM-based encoder-decoder models, we use the final encoder hidden state $\bar{h}_s$ to initialize the decoder. We did not investigate if separating the concerns of pooling the encoder's memory to a fixed-length context representation and encoding the last word of a sequence influences the performance.

Decoder The decoder distinguishes between two modes. The generation mode computes the probability $P_{gen}(\bullet)$ to generate a word from a predefined vocabulary. Following a similar idea of pointer generator networks (Paulus et al., 2017), a second copy mode outputs the probability of copying a word from the source document $P_{copy}(\bullet)$. Both probabilities are combined to approximate the output probabilities for the next word y_i as

$$P(y_i \mid h_i, y_{i-1}, c_i, \mathcal{M}) = \quad (1)$$
$$P_{gen}(y_i, \mathbf{g} \mid h_i, y_{i-1}, c_i, \mathcal{M}) + \quad (2)$$
$$P_{copy}(y_i, \mathbf{c} \mid h_i, y_{i-1}, c_i, \mathcal{M}) \quad (3)$$

where h_i is the current state of the decoder, y_{i-1} the last decoded word and $\mathbf{g}$ refers to the generation and $\mathbf{c}$ to the copy mode (Gu et al., 2016).

On the one hand, $P_{gen}(\bullet)$ uses the additive attention function (Bahdanau et al., 2014) of the encoder-decoder model. On the other hand, the scoring function for copying the j-th input word x_j with the encoder state $\bar{h}_j$ is

$$f(y_i = x_j) = \tanh(\bar{h}_j^\top W_c) h_i \quad (4)$$

where W_c is a learned parameter. These probabilities are jointly optimized with backpropagation during training by minimizing the negative log-likelihood.

4 Approach and Implementation

Our approach embeds learned knowledge of pre-trained language models to improve the language understanding of documents for abstractive text summarization. Let e denote the word embedding and c the contextual embedding (see Section 2) of an input word u. Following recent work (McCann et al., 2017; Peters et al., 2018), the final embedding of words is the concatenation of word embedding and contextual embedding $x = [e; c]$. To keep track of positional information in the Transformer encoder (see Section 3), we use relative position encodings (Vaswani et al., 2017).

In our implementation, the word embeddings are *pre-trained German GloVe embeddings*[1] of dimension 300. The contextual embeddings are extracted from the *multilingual BERT model*[2] of dimension 768. The concatenated embeddings of dimension $d_x = 1068$ are passed to the stacked self-attention encoder with $N = 4$ layers, $h = 8$ attention heads and a hidden dimensionality of 256. Furthermore, our decoder is a single-layer LSTM of dimensionality $d_{dec} = d_{enc}$.

[1] https://deepset.ai/german-word-embeddings

[2] https://github.com/google-research/bert/blob/master/multilingual.md

In order to avoid catastrophic forgetting (Howard and Ruder, 2018), contextual embeddings are fixed parameters and not optimized during training. On top of this, recent work (Peters et al., 2019) suggests that feature extraction with frozen parameters is favorable if the target task is very different from the source task and requires many learned parameters.

5 Dataset

We use an unreleased dataset[3] consisting of 100,000 samples extracted from German Wikipedia articles. For the best of our knowledge, a summary is the first section of the Wikipedia article and the document represents the subsequent sections. Documents consist of 602.81 words and summaries have 35.79 words, on average.

6 Experiments and Results

We hypothesize that contextual embeddings benefit the generation of German summaries. In order to test this hypothesis, we train with multilingual BERT embeddings and German GloVe embeddings (Section 4) and compare the results to two different baselines (Table 1). First, a plain model has randomly initialized embeddings of dimension 300. Secondly, embeddings of the same dimension are initialized with pre-trained German GloVe embeddings which reveals the actual impact of contextual BERT embeddings. In all experiments, word embeddings are fine-tuned during training.

Experimental Setup We train the model for a maximum of 25 epochs with early stopping and a patience of 5. Following recent work (See et al., 2017; Gehrmann et al., 2018), the models are optimized with Adagrad, a learning rate of $\eta = 0.15$ and an initial accumulator value of 0.1. The vocabulary is pre-defined and contains the 50,000 most frequent German words of the training dataset. The input documents are clipped to a length of 400 words and the target summaries to a length of 100 words. The 100,000 samples are randomly partitioned into three subsets of 80% training, 10% validation and 10% testing data. During inference, the model uses beam search with a beam size of 3. The subsequent results are obtained from a single run on the test dataset of the German Wikipedia dataset (Section 5).

[3] https://drive.switch.ch/index.php/s/YoyW9S8yml7wVhN

6.1 Lexical word similarity

We evaluate the lexical word similarity between generated summaries and the given reference summaries with ROUGE-F1 (Lin, 2004). Despite the fact that measuring lexical overlap is counter-intuitive to the concept of abstraction, our approach outperforms both extractive baselines, Lead-3 and TextRank (Mihalcea and Tarau, 2004), by a large margin (Table 1).

	R-1	R-2	R-L
GloVe + BERT	38.48	23.39	35.67
GloVe	29.16	14.33	36.44
Plain	27.39	12.96	24.66
TextRank	20.79	5.30	14.60
Lead-3	20.66	5.40	15.22

Table 1: ROUGE-F1 scores of our three different approaches on the German Wikipedia dataset. Lead-3 refers to a baseline extracted from the first three sentences of the document. TextRank is limited to 40 words.

Further, we find a significant improvement of additional multilingual BERT embeddings over pre-trained GloVe embeddings and learning embeddings from scratch. This supports our hypothesis that contextual embeddings are beneficial to the generation of summaries.

6.2 Level of abstraction

Abstractive summaries aim to express content in different words instead of merely copying sequences of words (Section 1). However, the ROUGE scores do not indicate the level of abstraction in generated summaries. For this reason, the copy rate (Nallapati et al., 2016) measures the average percentage of copied unigrams (words) from the given document.

The copy rate of the reference summaries is 72.52%, which highlights the need for abstraction in this dataset (Table 2). Both extractive baselines are not able to paraphrase and are therefore not fully capable to meet the requirements of the task. In contrast to this, all of our models generate summaries with an evident degree of abstraction. Although, evaluating the quality of abstraction still requires human assessment.

	Length	Copy	OOV	RR-4
B+G*	25.86	78.91%	3.58%	0.12
GloVe	18.79	74.11%	12.73%	0.04
Plain	22.33	73.53%	11.02%	0.06
T-Rank	40.19	100%	0%	0
Lead-3	52.47	100%	0%	0
Ref.	35.79	72.52%	0	0

Table 2: Average length (length), copy rate (copy), number of out-of-vocabulary words (OOV), and repetition rate with $n = 4$ (RR-4) on the German Wikipedia dataset for the approaches from Table 1. The references (ref.) refer to the gold summaries from the dataset. * multilingual BERT and German GloVe embeddings

6.3 Out-of-vocabulary words (OOV)

The copy mechanism of the CopyNet model (Section 4) encounters the shortcoming of OOV words during inference (Section 2). However, the results demonstrate that generated summaries still contain unknown words (Table 2). In comparison to other languages and datasets (Gu et al., 2016), this emphasizes that the model on the German Wikipedia dataset requires greater weights on the generation mode. Hence, this suggests that the CopyNet model has a trade-off between the level of abstraction and the number of unkown tokens.

Nevertheless, contextual BERT embeddings significantly drop the number of OOVs compared to our other approaches. This further justifies the aforementioned copy rate which decreases as the number of unknown words increases since these are not part of the source document.

6.4 Repetition

To measure the issue of repetition in text summarization models (Section 2), we use the repetition rate (Cettolo et al., 2014) which scores a summary by the number of repeated n-grams. More specifically, the repetition rate RR-$n(s)$ of a candidate summary s is

$$\text{RR-}n(s) = \left(\prod_{k=1}^{n} \frac{||f_{ng}(s,k) - f_{ng}(s,k,1)||}{||f_{ng}(s,k)||} \right)^{(\frac{1}{n})} \quad (5)$$

where n is the maximum number of considered n-grams, $f_{ng}(s,k)$ is a function creating a list of k-grams of s and $f_{ng}(s,k,1)$ consists of *unique* k-grams of s. Furthermore, $||\bullet||$ is the number of words in a set.

In our work, the generated summaries of the approach including contextual BERT embeddings create much higher repetition than other approaches (Table 2). However, this work focuses on transfer learning for text summarization and thus neglects further improvements to reduce repetition (Section 2).

6.5 Factual Incorrectness

As human observations suggest, summaries may contain false facts (Table 3) and yet achieve good results across several metrics. These factual errors are particularly difficult to detect and resemble with content-based measures since the lexical overlap can still be very high. Moreover, these summaries appear to be fluid and, at first sight, coherent. Thus, these issues are critical and remain an unsolved problem.

Generated summary	Reference summary
Miroslav Lazo ist ein Slowakischer eishockeyspieler, **der seit 2010 bei Awtomobilist Jekaterinburg** in der neugegrundete Champions League unter vertrag steht .	Miroslav Lazo ist ein slowakischer Eishockeyspieler , **der seit 2011 bei den Malmo Redhawks** in der schwedischen HockeyAllsvenskan unter Vertrag steht .

Table 3: Example of factual incorrectness in a generated summary with the BERT+GloVe approach.

7 Conclusion

Sequential transfer learning with pre-trained language models has shown to improve the performance for many tasks in NP. While previous research focussed on tasks like e.g. text classification or question answering (Devlin et al., 2018; Radford et al., 2018), this work investigates on the impact of pre-trained language models on abstractive summarization. Our experiments show that leveraging contextual embeddings extracted from multilingual BERT (Devlin et al., 2018) improves performance on a large summarization dataset in German language.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473* https://arxiv.org/abs/1409.0473.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2014. The Repetition Rate of Text as a Predictor of the Effectiveness of Machine Translation Adaptation. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*. pages 166–179.

John M Conroy and Dianne P O'leary. 2001. Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 406–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* https://arxiv.org/abs/1810.04805.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pages 4098–4109.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O K Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. *arXiv:1603.06393v3* https://arxiv.org/abs/1603.06393v3.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. *arXiv:1801.06146* https://arxiv.org/abs/1801.06146.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. *Text Summarization Branches Out* pages 74–81.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. *arXiv:1708.00107* https://arxiv.org/abs/1708.00107.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A Deep Reinforced Model for Abstractive Summarization. *arXiv:1705.04304* https://arxiv.org/abs/1705.04304.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

Matthew Peters, Sebastian Ruder, and Noah A Smith. 2019. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. *arXiv:1903.05987v2* https://arxiv.org/abs/arXiv:1903.05987v2.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv:1802.05365* https://arxiv.org/abs/1802.05365.

Alec Radford, Karthik Narasimhan, Salimans. Tim, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016. Self-critical Sequence Training for Image Captioning. *arXiv:1612.00563* https://arxiv.org/abs/1612.00563.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. *arXiv:1509.00685* https://arxiv.org/abs/1509.00685.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *arXiv:1704.04368* https://arxiv.org/abs/1704.04368.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *IJCAI*. pages 2862–2867.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling Coverage for Neural Machine Translation. *arXiv:1601.04811* https://arxiv.org/abs/1601.04811.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762* https://arxiv.org/abs/1706.03762.