Fine-tuning BERT Models for Summarizing German Radiology Findings

Siting Liang[1,*,§], Klaus Kades[2,3,*], Matthias A. Fink[4,5], Peter M. Full[2],
Tim F. Weber[4,5], Jens Kleesiek[6,7], Michael Strube[8], and Klaus Maier-Hein[2,9]

[1]German Research Center for Artificial Intelligence,
[2]Division of Medical Image Computing at German Cancer Research Center (DKFZ),
[3]Faculty of Mathematics and Computer Science, Heidelberg University,
[4]Clinic for Diagnostic and Interventional Radiology, University Hospital Heidelberg,
[5]Translational Lung Research Center (TLRC) Heidelberg,
[6]German Cancer Consortium (DKTK, Partner Sites Essen and Heidelberg),
[7]Institute for Artificial Intelligence in Medicine (IKIM), University Medicine Essen,
[8]Heidelberg Institute for Theoretical Studies gGmbH,
[9]Pattern Analysis and Learning Group, University Hospital Heidelberg,
Germany
siting.liang@dfki.de,k.kades@dkfz.de

Abstract

Writing the conclusion section of radiology reports is essential for communicating the radiology findings and its assessment to physician in a condensed form. In this work, we employ a transformer-based Seq2Seq model for generating the conclusion section of German radiology reports. The model is initialized with the pretrained parameters of a German BERT model and fine-tuned in our downstream task on our domain data. We proposed two strategies to improve the factual correctness of the model. In the first method, next to the abstractive learning objective, we introduce an extraction learning objective to train the decoder in the model to both generate one summary sequence and extract the key findings from the source input. The second approach is to integrate the pointer mechanism into the transformer-based Seq2Seq model. The pointer network helps the Seq2Seq model to choose between generating tokens from the vocabulary or copying parts from the source input during generation. The results of the automatic and human evaluations show that the enhanced Seq2Seq model is capable of generating human-like radiology conclusions and that the improved models effectively reduce the factual errors in the generations despite the small amount of training data.

1 Introduction

For patients with cancer, imaging findings are critical for primary diagnosis and treatment guidance during further disease progression. Depending on the tumor entity and stage, the results of imaging examinations may have a significant impact on the clinician's treatment decisions and strategies. Normally, imaging findings are communicated in clinical routine in the form of written radiology reports. However, it remains difficult to ensure the completeness and comprehensibility of relevant information in traditional written reports. Free-form narrative reports do not have standardized layout and uniform terminology, and key findings may be forgotten, which can lead to serious miscommunication (Weber et al., 2020).

Weber et al., 2020 implemented the application of Structured Oncology Reporting (SOR) to address the problems of traditional radiology reporting. The SOR, which structure is shown in Table 1, demonstrated superiority to the free-text format of radiology reports by providing disease-specific report templates and organizing the content in specific separate sections.

The main goal of this work is to automatically extract information relevant for treatment planning from standardized, real-life radiology reports. Expert validation is on the other hand still essential for this clinical routine application. For this purpose, we build a system that merges the information available in the general information and findings sections of the SOR radiology reports into a conclusion, which can be compared to conclusions generated by human experts.

Our main contributions in this work includes: (i) We tested the effectiveness of applying the generic

[*]Corresponding authors contributed equally.
[§]Work completed during master thesis at DKFZ.

Proceedings of the 4th Clinical Natural Language Processing Workshop, pages 30 - 40

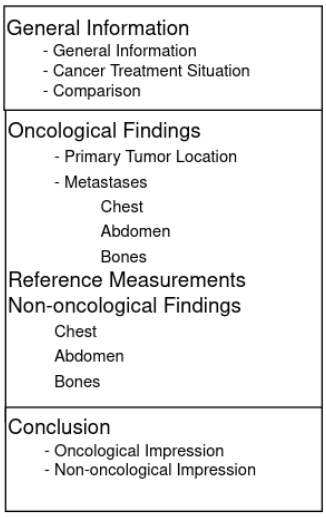

Figure 1: Standardised Layout of SOR (Weber et al., 2020). Each report has a uniform organization: the general section expresses background information on imaging and clinical data, the next section (Findings) describes oncology and non-oncology findings, and the Conclusion section gives oncological and non-oncological impressions.

pretrained German BERT model directly to the target task of generating conclusions of German radiology reports without domain-adaptive pretraining. (ii) Our system improves the factual correctness of the generated conclusions by combining extractive and abstractive learning objectives compared to the Seq2Seq baseline model. (iii) Our expert evaluation shows that the summarizations generated by our system are very close to the human reference. Since our work focuses on the application of NLP with pretrained language models to automated radiology documentation, the above contributions are limited to German SOR data. However, our experiments suggest that good results can also be obtained in low-resource domains by applying lightweight pretrained language models and minor modifications to standard architectures.

2 Related Work

Existing text summarization models can be broadly classified into three categories: extractive, abstractive and hybrid. Early extractive approaches relied on human-designed features extracted from texts to identify key sentences. Deep learning methods show good performance in various of NLP tasks. The data-driven approaches are able to learn features representations automatically. Extractive models have the advantage of producing semantically and syntactically correct summaries. Abstractive models employing an encoder-decoder framework with attentive recurrent neural networks, e.g. on news article corpus, became a standard architecture in abstractive summarization, which translates the original source content to a concise expression about the main content of the source input (Nallapati et al., 2016a; See et al., 2017; Gu et al., 2016; Kryściński et al., 2018; Chopra et al., 2016). In order to improve the faithfulness of the generated summarization given the facts in the source input, abstractive models are usually enhanced to replicate facts from the source combining extractive and abstractive approaches. Nallapati et al., 2016b incorporated a pointer network (Vinyals et al., 2015) that selects a word from a predefined vocabulary to replace an unknown word predicted by a RNN-based encoder-decoder model. Our work aims to combine both benefits of extractive and abstractive summarization with a transformer-based model.

See et al., 2017 used the pointer network Nallapati et al.2016b as a soft switch to either produce a word from the vocabulary distribution or to select a word from a copy distribution provided by a target-source attention distribution. Chen and Bansal; Kryściński et al., 2018; 2018 also applied the copy mechanism to the RNN-based model, but decomposed the decoder into a first-stage extraction model and a second-stage generator. In the first stage, the encoders in both works processed sequential document representation and provided sentence-level representations to the extractor for selection. In the second stage, Kryściński et al., 2018 used the language model to rewrite the selected sentences into the summary. Chen and Bansal, 2018 trained the decoder from scratch by using ROUGE (Lin, 2004) scores as a reward strategy for reinforcement learning to generate summaries based on the selected sentences. In our work, we integrate the pointer network to a transformer-based encoder-decoder model.

Summarizing radiology findings with neural Seq2Seq learning of Zhang et al.; Zhang et al. is very closely related to our work. Zhang et al., 2018 collected a large set of domain-specific training data to train the RNN-based pointer-generator (See et al., 2017). Because there are usually two sections in radiology reports: background and findings, to provide relevant information for the summary, Zhang et al., 2018 incorporated an extra encoder for encoding the background information and findings separately. In contrast, we feed the combination of sequences of the background and findings

section as one input and into one encoder. Zhang et al., 2019b improved the radiology summarization model by optimizing the factual correctness of the summaries via policy learning. In order to combine extraction and abstraction in one model, we propose two target sequences paired with an input sequence. One target sequence is the reference summary and the other is a sequence consisting of key sentences extracted from the input. Our goal with the dual target sequences is to encourage the encoder-decoder model to retain some of the input while generating new phrases for the summaries.

Pretrained language models have advanced the state-of-the-art when fine-tuned in various NLP tasks, as well as in automatic text summarization (Miller, 2019; Liu and Lapata, 2019; Zhang et al., 2019a). Rothe et al., 2019 demonstrated the efficacy of warm-starting the encoder and decoder from checkpoints of publicly available large language models, including BERT(Devlin et al., 2018) and GPT-2(Radford et al., 2019), for text generation task such as machine translation and text summarization. Depending on different initialization combinations, they investigated variants of the Seq2Seq model, such as BERT2Random, BERT2BERT, BERT2GPT, etc. Warm-starting the Seq2Seq model leveraging these pretrained language models checkpoints can reduce computational resources and time by orders of magnitude, while improving the sequence generation performance. We adopt the warm-starting idea and initialize both the encoder and decoder with a generic pretrained German BERT model (deepset.ai, 2019). We fine-tune the model with our German radiology report data and enhance the model by combining extractive and abstractive objectives.

3 Models

The main task of summarizing radiology findings is to transform the salient and clinically significant findings from a source of words and phrases $X = \{x_1, x_2, ..., x_T\}$, to a sequence of concise expressions $Y = \{y_1, y_2, ..., y'_T\}$. Background information in the radiology report conveys important information for short-term or long-term examination of each patient in the clinical routine, which is why abstractive models needs to incorporate background information into the summary generation (Zhang et al., 2018). The content of the source sequence X contains the background information and imaging findings. These findings convey the information about the location of the primary tumour, the presence of metastases at different body regions, and other non-oncological findings. Y is the conclusion of the radiology report, which on the one hand assesses the patient's condition according to the detailed findings and on the other hand concisely summarizes the significant findings from the source sequence X. We use a collection of aligned X and Y pairs to train Transformer-based Seq2Seq models to generate Y.

Baseline Model Warm-starting the Seq2Seq model leveraging pretrained checkpoints can reduce computational resources and time by orders of magnitude, while improving the sequence generation performance (Rothe et al., 2019). We utilize the **BERT2BERT** model defined in Rothe et al., 2019, as our abstractive summarization baseline model.

The encoder and decoder of the model are initialized from a public available BERT checkpoint (deepset.ai, 2019), except the encoder-decoder attention layers in the decoder. Taking advantage of the Transformer architecture and pretrained language models, among the 221 millions trainable parameters in the **BERT2BERT** model, only 26 millions parameters in the encoder-decoder attention layers are initialized randomly, and 195 millions are loaded from the pretrained BERT model. The reduction of randomly initialized, trainable parameters, allows for fewer fine-tuning steps, and the model's ability to perform well on small training data sets.

BERT2BERT + Extraction Most abstractive systems suffer from the problem of creating spurious facts due to their ability to paraphrase. Hybrid systems that combine extraction and abstraction are expected to improve the correctness of the generated facts by using more criteria to extract the original facts from the source (Kryscinski et al., 2019; Cao et al., 2017; Zhang et al., 2019b; Chawla et al., 2019; Falke et al., 2019). Different to previous works, which incorporated separate extraction and abstraction stages (Hsu et al., 2018; Li et al., 2018; Chen and Bansal, 2018), we propose a new learning scenario with little modification to the architecture of the **BERT2BERT** model by adding an extraction learning objective (**BERT2BERT+Ext**). Therefore, during training, we optimize the following combined loss:

$$Loss = loss_{abstraction} + loss_{extraction} \quad (1)$$

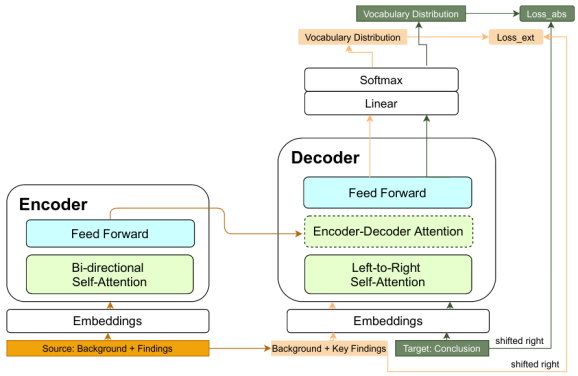

Figure 2: BERT2BERT model adding Extraction Loss. In order to train the decoder to extract the key findings through generation, we supply an additional target sequence ("Background + Key Findings"), which consists of the key findings selected from the source sequence.

The setup is illustrated in Figure 2. Through the extraction objective, the model is trained to reconstruct the key sentences in the generation.

In the original setting of **BERT2BERT**, we only train the model using our source and target sequence pairs (X, Y). As showed in Figure 2, X symbolizes the source input and contains "Background + Findings" and Y is the target input "Conclusion". During training, the decoder of **BERT2BERT+Ext** is fed with additional target sequences ("Background + Key Findings") including the general section and key sentences from the findings sections as input. Section 4.3 explains how to extract these key findings from the finding section from our training data. Extractive loss encourages the model to reconstruct key phrases from the source input. Abstractive loss prompts the model to generate new formulations that are not from the source sequence.

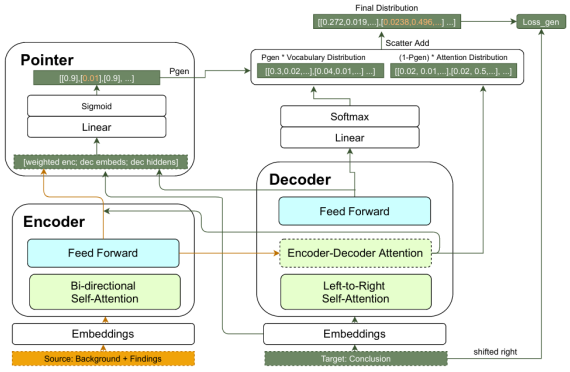

Figure 3: BERT2BERT model incorporating the Pointer Mechanism.

BERT2BERT + Pointer Pointer networks allow the model to copy words from the source sequence through an alignment between the target sequence and the source sequence (See et al., 2017). The benefits of incorporating the pointer to the generation procedure are not only to reduce the number of tokens, which are not known to BERT, but also to ensure factual correctness while generating new phrases. Pointer networks have been used for abstract summaries of Seq2Seq models based on RNNs as a standard architecture. However, to the best of our knowledge, there has been little exploration of incorporating pointer networks into the Transformer encoder-decoder model for summarization tasks. Figure 3 illustrates the combination of **BERT2BERT** and the pointer mechanism (**BERT2BERT+Ptr**). The pointer network consists of one linear layer followed by a sigmoid function which generates a pseudo-probability p_{gen} in the range of $[0, 1]$. In the original function of See et al., 2017, p_{gen} is given by:

$$p_{gen} = sigm(w_{ptr}^T[h_t^x; y_t; s_t] + b_{ptr}) \quad (2)$$

where w_{ptr}^T and b_{ptr} are learnable parameters. p_{gen} is determined by the concatenated representation containing the word embeddings of the input token y_t, the decoder hidden state s_t and the weighted encoder hidden representations h_t^x, at each decoding step t.

See et al., 2017 recycled attention scores directly from the encoder-decoder attention layer. However, in the **BERT2BERT** model, we not only have multiple encoders and decoders, but also multiple heads of the encoder-decoder attention. We can solve the dimension of multiple heads in the attention distribution using the mean of the multi-head attentions (Deaton, 2019). These hidden states from the final encoder are used as context vectors passed to each decoding step. Each decoder state s_t used for predicting the next token is also from the last decoder, as well as the multi-head encoder-decoder attention scores a_t. h_t^x in Equation 3 represents the hidden output from the final encoder weighted by the sum of the heads of the encoder-decoder attention layers at each decoder step from the last decoder, analogous to the RNN-based context vector. h_t^x is given by:

$$h_t^x = \sum_j^{T_x} \sum_i^{N_{heads}} a_t \cdot h_j^x \quad (3)$$

where i is the index of the attention head, j is the position of the source sequence and T_x is the total length of the source sequence. The formula for computing the final distribution $P_{final}(w)$ is as follows:

$$P_{final}(w) = p_{gen} \cdot P_{vocab}(w) + (1 - p_{gen}) \cdot \sum_{i:w_i=w} a_i^t \quad (4)$$

$P_{vocab}(w)$ has the dimension of the size of the vocabulary. a^t contains the values for each token in the source sequence, and each value has a corresponding index i in the vocabulary dimension. The encoder and decoder of **BERT2BERT** share the same vocabulary. Hence, we can sum the values from a^t and P_{vocab} at the same indices.

4 Experiments

4.1 Datasets for Training and Testing

The concept of structured oncology reports (SOR) has been implemented to generate high-quality radiology reports for the general follow-up assessment of cancer patients in the clinical routine at the University Hospital Heidelberg (UKHD) in Germany by Weber et al., 2020. The design and application of SOR can be accessed using the internet link: `http://www.targetedreporting.com/sor/`. For our experiments, we use a collection of 10,514 structured reports from the years 2018 and 2019 from the radiology department of the UKHD. The HIPAA-compliant retrospective study was approved by the Institutional Review Board (S-083/2018), and informed consent was waived. The reports are divided into a training set (80%), a validation set (10%), and a test set (10%).

	training (8410)	valid (1052)	test (1052)
general	2.0	2.0	2.0
findings	21.1 ± 8.2	20.5 ± 7.5	21.7 ± 7.5
conclusion	3.1 ± 2.0	3.4 ± 2.0	3.5 ± 2.0

Table 1: The average number of sentences after segmentation in each section. The general section contains 2 sentences of the background information. The number of sentences in the findings section averages about 22 sentences, with a variation of 7-8 sentences. The conclusion consists of approximately 3-6 sentences.

Sentence Segmentation Each section of the SOR report contains documentation in a tabulated form. Different sections have different table blocks. We need to customize different methods to segment sentences from different sections. In the general section, there are normally two sentences expressing the treatment situation and previous examinations. In the finding sections, we have notes organized in different blocks and free-text content. There are four main blocks: primary tumour location, metastases, reference measurements and non-oncology findings.

The first step is to detect the boundaries of the blocks. After that, we apply a tailor-made regular expression segmenter to split the text in these blocks into sentences. In report texts, periods are usually used to mark the end of sentences and can be used to split text into sentences. However, applying this rule to the findings and conclusion sections requires consideration of several cases, such as abbreviations, dates, and serial numbers, where the period is part of the tokens. We customize the regular expressions to handle the above exceptions. The average number of sentences in each section calculated for each split set can be found in Table 1.

Patient Degree Categories Weber et al., 2020 used a uniform terminology to ensure the formalities of the content in the conclusion section as assessments of patient responses. These terminologies are shown in Table 2.

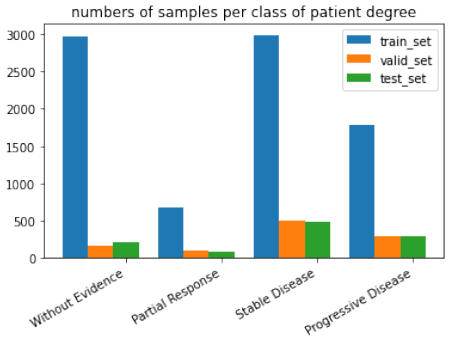

Figure 4: Number of reports for the three data partitions after matching to patient degree categories. We have significantly more reports in the Without Evidence and Stable Disease categories than in the other two categories, and the fewest reports are found in the Partial Response category.

The reports from different patient degree categories challenge our model to varying degrees. For example, a report that contains findings indicating progressive disease is much more complex than a report that does not show findings regarding tumour burden. It would be more appropriate to judge the performance of the model based on the patient degree class of the report. As shown in Figure 4, after dividing the reports into four patient categories, the

Patient Degree	SOR Category	German Template
Without Evidence (WE)	no tumour burden evidence	Oncological regular findings without evidence of recrudesce or metastasis (Onkologisch regelrechter Befund ohne Nachweis von Rezidiv oder Metastasierung)
Partial Response (PR)	significant decrease of tumour burden	Oncological improvement of findings; constancy of findings with a tendency to decrease (Onkologisch Befundverbesserung; Befundkonstanz mit tendenzieller Abnahme)
Stable Disease (SD)	no significant change of tumour burden	Oncological constancy of findings (Onkologisch Befundkonstanz)
Progressive Disease (PD)	significant increase of tumour burden	Oncological worsening of findings; constancy of findings with a tendency to increase (Onkologisch Befundverschlechterung; Befundkonstanz mit tendenzieller Zunahme)

Table 2: Patient degree categories and corresponding uniform terminology in conclusion. The SOR categories are defined by threshold criteria for tumour burden development in the implementation. For example, if there is a significant decrease of tumour burden (more than 30%), the patient degree is defined as Partial Response.

BERT2BERT	baseline
BERT2BERT+Ext	adding extraction learning objective
BERT2BERT+Ptr	integrating pointer network
BERT2BERT+Ext+Ptr	combining extraction and pointer

Table 3: The abstractive models are warm-started with the checkpoints from the **German BERT** (deepset.ai, 2019).

number of reports is imbalance across patient categories, however, is kept similar across the three data splits. The number of training samples is an important factor in the performance of the model. Given uneven quantity and the varying complexity of reports across categories, we expect inconsistent performance of the models across the four patient degree categories.

4.2 Experimental Setup

In our experiments, we evaluate the efficacy of the proposed **BERT2BERT** baseline and its enhancements, shown in Table 3. The implementation of all BERT-based models is based on the open source library **HuggingFace Transformers** by Wolf et al., which is dedicated to supporting state-of-the-art Transformer architectures and to collecting and supplying pretrained models for the community. The models are fine-tuned on 8410 reports and validated on 1052 samples during the training. The maximum number of training epochs is 10 with an early stopping setting according to the validation loss metric: when the validation loss is no longer decreasing within 3 epochs, the training process is terminated. All fine-tuning processes are conducted using one single GPU of 32GB memory and completed in no longer than 6 hours.

Input Sequences We combine the sentences from the background and finding sections in one input sequence and feed them into the encoder of the model. We adopt the idea from Liu and Lapata, 2019 of inserting "[CLS]" tokens between the sentences to construct structured sequences. Since BERT is not a generative model and does not learn an end of text token like GPT-2 does, we use the "[SEP]" token to make the end of the whole sequence, so that the decoder in **BERT2BERT** stops the generation when it sees this special token.

Evaluation Metrics For quantitative evaluation, we firstly apply the ROUGE metric (Lin, 2004) and report the F_1 scores for ROUGE-1 and ROUGE-L about the tokens overlaps between the system-generated summaries against the reference conclusions. Secondly, we propose the patient degree matching metric, evaluating whether the assessments generated by the abstractive models can be categorized to the same patient degree category as their reference. After that, we conduct a human evaluation with two domain experts in which the annotators are asked to score the system-generated conclusions as well as the reference based on three criteria: comprehensibility, oncology and non-oncology correctness.

4.3 Extracting Key Sentences

We propose the **BERT2BERT+Ext** model in Section 3 to improve the extraction ability of the decoder during generation, however, we lack key sentences for training. For finding the most effective way to extract the key sentences, we evaluate several non-neural, automatic extractive methods on the test data:

1. **Longest-k**. This method simply extracts the k longest sentences from the findings. We hypothesize the longer a sentence of findings is, the more information it may communicate in the summary.

2. **Tfidf-Ex**. This approach is built on the scores of TF-IDF (Jones, 1972). TF-IDF produces a vocabulary based on the collection of documents and outputs a TF-IDF vector of vocabulary breadth. We can set a threshold to extract

the top keywords from the TF-IDF vector. The sentences are ranked based on the scores by summing up the TF-IDF of all the keywords found in the sentence of a document. Top k sentences are extracted as the salient sentences.

3. **TextRank** (Mihalcea and Tarau, 2004) algorithm scores sentences based on the graph theory. In the algorithm, a graph is constructed with each sentence in the document as a vertex, and the score of edges between sentences are determined based on the number of overlapping tokens indicating the similarity between sentences.

All the extractive approaches assign an importance score to each sentence from the findings section and rank the sentences according to the scores. We compare the results of the different methods in Table 4. All the three extractive methods return comparable results. The **Longest-k** method is the simplest extractive method for which no computation is required and indicates that phrases from the longest sentences in the finding sections are usually included in the human-written summaries.

	Longest-k	Tfidf-Ex	TextRank
ROUGE-1	**41.9**	40.6	40.4
ROUGE-L	**40.8**	38.7	39.6

Table 4: The recall scores of ROUGE metrics for the different extractive approaches. The scores imply how much of overlaps between the key sentences and the reference is found. The 2 sentences from the general section are always included in the extraction. Because the average number of sentences in the reference summaries does not exceed 6, we evaluate the key sentences given k=4, i.e, 4 key sentences from the findings section.

4.4 Human Evaluation

Since the ROUGE metric only assesses the similarity between the system-generated conclusions and the references, we conduct an expertise evaluation with two domain annotators (one radiologist and one final year medical student) to understand the clinical validity of the conclusions generated by the abstractive models. According to the radiologist, there are two important criteria to judge the clinical validity of the conclusions, namely the degree of correctness of the oncological and non-oncological impressions based on the patient's condition. In addition, we ask the annotators to score the comprehensibility of system-generated and referenced findings with expert judgment to investigate whether the abstractive models could produce medical terms that are as comprehensible as those written by specialists. In the evaluation, we first create a pool of samples, where each sample has scored higher ROUGE-1 scores than the average in the entire test set for all four abstractive models. Next, we randomly select five examples from the pool for each patient degree category, totalling twenty samples. We present the general information, findings secions and the four system-generated conclusions as well as the reference conclusion of each sample to the annotators in a random order. They are asked to score the conclusions on a likert scale from 0 to 5, indicating oncological and non-oncological correctness degrees as well as comprehensibility from very poor to very good. A score of 3 indicates satisfaction. The annotators have no prior knowledge of the models nor the reference. The annotator instructions are given in Appedix A. The annotation was performed with the open source text annotation tool doccano (Nakayama et al., 2018).

5 Results and Discussion

	whole	WE	PR	SD	PD
BERT2BERT	36.15	55.27	30.86	32.09	30.93
BERT2BERT+Ext	42.13	**58.99**	38.19	38.17	36.68
BERT2BERT+Ext(random)	37.27	57.22	31.43	32.71	31.32
BERT2BERT+Ptr	42.25	55.9	38.66	**39.88**	**39.04**
BERT2BERT+Ext+Ptr	**43.32**	57.91	**40.15**	39.39	38.65
BERT2BERT+Ext+Ptr(random)	42.10	57.37	38.71	39.41	37.81

Table 5: ROUGE-1 F_1 scores of BERT2BERT-based Models on the whole test set and different partitions of four Patient Degree Classes. BERT2BERT+Ext(random) has random selected sentences targets. When the target sentences to be extracted are replaced with randomly selected sentences, no significant improvement is found in BERT2BERT+Ext models.

Table 5 shows the F_1 scores of ROUGE-1 metric across the different settings of the abstractive model overall reports and according to the patient degree categories. The hybrid models outperform the **BERT2BERT** model by nearly 6 points. Integrating extraction or pointer mechanism yields comparable results. According to the metrics, the last hybrid model combining the two facilities achieves only a small improvement compared to enhancing the model only with extraction training or pointer network. One SOR report and the generations of the abstractive models are shown in Appendix B.

Both the baseline model and the hybrid model have less difficulty in generating summaries for the *WE* class. We hypothesize that this is because in this category, there are many training samples (almost one-third of the reports), uniform templates,

and barely important information can be extracted from the findings. In the templates of SD, there is only one statement about the findings: "oncological constancy". For the PR and PD classes, there are cases in which an oncological constancy is described, however, with a tendency to an improvement or deterioration, which increases the difficulties of the generation task for the models.

In the case of evaluating oncology facts, their correctness requires more expertise to assess. Hence, we need to present examples of system-generated conclusions to domain experts to assess clinical validity. The results of the expertise assessment are presented in Section 5.

Validation of the Extraction Learning Since we do not have human-annotated labels in the radiology reports indicating the important sentences, we apply the **Longest-k** method explained in 4.3 to extract the key sentences used as target for training **BERT2BERT+Ext**. In **BERT2BERT+Ext(random)** we replaced the target sentences with random ones. The results in Table 5 show that, the performance of **BERT2BERT+Ext(random)** drops in comparison to **BERT2BERT+Ext**. This verifies the importance of target sentences for improving the extraction ability of the **BERT2BERT+Ext** model. In **BERT2BERT+Ext+Ptr(random)**, the scores obtained by integrating the pointer mechanism are not significantly affected when the decoder is trained to extract sentences that include irrelevant sentences. From the ROUGE scores, we can conclude that the hybrid models achieve better results than the baseline model.

Results of Expert Evaluation Figure 5 presents the correctness results of oncological and non-oncological impressions as well as the comprehensibility of the impressions. A score of zero indicates unacceptable generation given the facts in the source input, while a score of five means that the facts in the generation are completely correct.

The results shown in the bar charts are the average scores of the two annotators, normalized by number of the examples in each category. In terms of correctness regarding oncological and non-oncological impressions in the WE patient degree, all conclusions generated by the abstractive models are scored close to 5. In SD category, the generated conclusions are almost as good as the human-written conclusions, except for the baseline model. Summarizing the findings for the PR and PD categories is more challenging for the models due to the complexity of the findings and the small number of training examples. The hybrid models perform better than the baseline, but the correctness of their generations are rated very differently in these two categories. The **BERT2BERT+Ext+Ptr** model performs best in ensuring correctness across patient degree categories in general. Figure 5 shows that, the abstractive models are capable of generating good comprehensible radiology conclusions, except for the baseline model in the PD category. Although the PR class has the fewest training instances, the abstractive models also achieve results above 3.

6 Conclusion

In this work, we experiment and demonstrate the efficacy of the **BERT2BERT**-based abstractive models on summarizing German radiology findings in structured reports. We propose two strategies to improve the **BERT2BERT** model with the aim of optimizing the factual correctness in the conclusions generated by the system, **BERT2BERT+Ext** and **BERT2BERT+Ptr**. Both **BERT2BERT+Ext** and **BERT2BERT+Ptr** models have very few modifications to the baseline model and improve the performance of the model. In **BERT2BERT+Ext**, we train the model to generate summaries, encouraging the model to reconstruct key sentences based on the source text in the training process. **BERT2BERT+Ptr** incorporates the pointer mechanism to modify the decoder's prediction by copying the salient segments directly from the source sequence. Despite the limitations of the models and the imbalanced training data, the issue of unfaithful facts in the conclusions generated by the baseline model is greatly improved by these hybrid models. One pressing issue in the future work is to investigate the potential advantages of these models on free-text radiology data or data in other domains.

Acknowledgments

This research was supported by the German Cancer Consortium (DKTK, Strategic Initiative *Joint Imaging Platform*) and carried out during a master thesis at German Cancer Research Center. This work is further supported by the pAItient project (BMG, 2520DAT0P2).

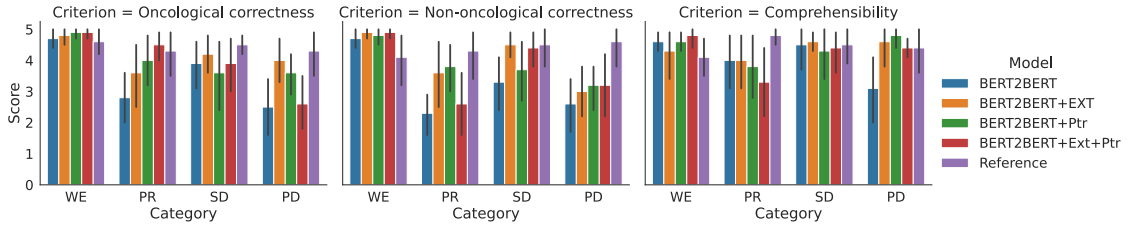

Figure 5: Average scores with standard deviation for the three criteria: Oncological correctness, non-oncological correctness and comprehensibility.

References

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the original: Fact aware neural abstractive summarization. *CoRR*, abs/1711.04434.

Kushal Chawla, Kundan Krishna, and Balaji Vasan Srinivasan. 2019. Improving generation quality of pointer networks via guided attention. *CoRR*, abs/1901.11492.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

John Deaton. 2019. Transformers and pointer-generator networks for abstractive summarization.

deepset.ai. 2019. Open sourcing german bert.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *CoRR*, abs/1603.06393.

Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. *CoRR*, abs/1805.06266.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *CoRR*, abs/1910.12840.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. 2018. Improving neural abstractive document summarization with explicit information selection modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Brussels, Belgium. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *CoRR*, abs/1908.08345.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Derek Miller. 2019. Leveraging BERT for extractive text summarization on lectures. *CoRR*, abs/1906.04165.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016a. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.

Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016b. Classify or select: Neural architectures for extractive document summarization. *CoRR*, abs/1611.04244.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2019. Leveraging pre-trained checkpoints for sequence generation tasks. *CoRR*, abs/1907.12461.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks.

TF Weber, M Spurny, FC Hasse, O Sedlaczek, GM Haag, C Springfeld, T Mokry, D Jäger, HU Kauczor, and AK Berger. 2020. Improving radiologic communication in oncology: a single-centre experience with structured reporting for cancer patients. *Insights Imaging*, 11(1):106.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019a. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. *CoRR*, abs/1809.04698.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis P. Langlotz. 2019b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *CoRR*, abs/1911.02541.

A Annotator Instructions

We have discussed the criteria with domain experts for judging the correctness of system-generated conclusions. We define an annotation task for grading generated assessments according to certain criteria. The following instructions are presented to annotators for evaluating the generated summaries. are a senior medical student and a radiologist.

Evaluation Criteria We consider four summary models and the reference conclusion. Each model generates a radiological summary (assessment) under the specification of a source text (general examination information and radiological findings). You will be presented with the source text, the four generations from the models, and the assessment written by the physician. Please rate the generations of each model and the reference according to the following criteria: Oncolgical correctness, nononcolgical correctness, and readability:

- Oncological correctness: is the summary and the details about metastases (none, new, proliferation or regressive) correct? (0) not assessable; (1) not at all correct ; (2) correct to a small extent ; (3) half correct; (4) correct to a large extent; (5) everything correct.

- Nononcological correctness: is the general date, organ, and other information correct? (0) not assessable; (1) not at all correct ; (2) correct to a small extent ; (3) half correct; (4) correct to a large extent; (5) everything correct.

- Readability: is the generation easy to understand, without broken expressions or unknown words? (0) not assessable; (1) many unknown words, difficult to read and comprehend; (2) several unknown words and aborted expressions, not fluent; (3) several unknown words; (4) fluent and coherent, but some unknown words; (5) correct words and expressions, fluent and coherent.

If the generation is not assessable, select 0 - not assessable. Otherwise, the scale are grades from 1 to 5 and must be assigned for each criterion.

B SOR Report Example

We present one SOR example from (Weber et al., 2020) in Table 6 along with the generations of the

General Section	Untersuchungsregion Thorax (CT), Abdomen (CT) Behandlungssituation Ausgangsbefund. Vergleich Letzte Vergleichsuntersuchung: 17.11.2017.
Findings Section	Primärtumor / Lokalrezidiv Soweit messtechnisch erschwert erfassbar progrediente diffus infiltrierende Raumforderung des Pankreaskopfs mit Gangstau im Pankreasschwanz und vollständiger Ummauerung des Truncus coeliacus, mindenstens 180° Ummauerung der A. liniealis . Bekannter kompletter Verschluss der extrahepatischen Pfortader und V. mesenteria superior mit ausgeprägten Kollateralen . Regionäre Lymphknoten Gering prominenterer vermehrter Lymphknotenbesatz mesenterial, exemplarisch mit einem KAD von 7 mm, zuvor 5 mm (8-137) . Metastasen Lunge und Pleura: Keine . Thorakale Lymphknoten und Weichteile: Keine . Leber: Keine . Abdominale Lymphknoten und Weichteile: Keine . Peritoneum: Kein eindeutiger Nachweis einer Peritonealkarzinose, jedoch Infiltration der Mesenterialwurzel durch den Primarius und Nachweis geringer freier Flüssigkeit im kleinen Becken . Skelett: Keine . Referenzmessungen Keine Referenzmessungen . Thorax Portsystem in situ . Kein katheterassoziierter Thrombus . Abdomen Unveränderte Nebenmilze . Konstante geringe Cholestase Skelett Im Übrigen ohne besondere Auffälligkeiten .
Conclusion	Onkologische Beurteilung Onkologisch Befundkonstanz mit tendenzieller Zunahme der Tumorlast gegenüber 17.11.2017. Progredient diffus infiltrierende Raumforderung des Pankreaskopfs. Andere Befunde Keine Angabe.(Progressive diffusely infiltrating mass of the pancreatic head.)
BERT2BERT	Onkologische Beurteilung Onkologisch Befundverschlechterung gegenüber dem Ausgangsbefund vom 11.04.2018 und 12.05.2019. In diesem Zeitraum zunehmende Peritonealkarzinose mit Ummauerung der VMS, sowie Ummauerung des TCs bei bekanntem Magenkarzinom. Andere Befunde Keine Angabe.(During this period, increasing peritoneal carcinomatosis with walling of the VMS, as well as walling of the TC in known gastric carcinoma.)
BERT2BERT +Ext	Onkologische Beurteilung Onkologisch Befundkonstanz gegenüber 17.11.2017. In diesem Zeitraum keine wesentliche Änderung des Pankreasschwanzes mit Infiltration der Pfortader und Ummauerung der Truncus coeliacus, AMS / VMS sowie Arteria hepatica communis durch die AMS - suspekte Lymphknotenmetastasen. Andere Befunde Keine Angabe.(During this period, no significant change in the tail of the pancreas with infiltration of the portal vein and walled truncus coeliacus, AMS / VMS and arteria hepatica communis by the AMS - suspicious lymph node metastases.)
BERT2BERT +Ptr	Onkologische Beurteilung Onkologische Befundverschlechterung zum 17.11.2017. In diesem Zeitraum neu aufgetretene Lymphknotenmetastase im Oberbauch mit ausgeprägter Infiltration der Pfortader und der V. Pankreatitis. Unveränderte Darstellung der bekannten Peritonealkarzinose. Andere Befunde Keine Angabe.(Newly appeared lymph node metastasis in the upper abdomen during this period with marked infiltration of the portal vein and pancreatic vein. Unchanged presentation of known peritoneal carcinomatosis.)
BERT2BERT +Ext+Ptr	Onkologische Beurteilung Onkologisch Befundkonstanz mit tendenzieller Zunahme der Tumorlast gegenüber 17.11.2017. In diesem Zeitraum moderate Zunahme des Primärtumors im Pankreasschwanz. Bekannter Verschluss der extrahepatischen Pfortader und Ummauerung der A. gastrica sinistra und Infiltration der Pfortader. Andere Befunde Keine Angabe.(Moderate increase in primary tumor in the pancreatic tail during this period. Known occlusion of the extrahepatic portal vein and walling of the gastrica sinistra artery and infiltration of the portal vein.)

Table 6: In conclusion is a human-written summary reference. Followings are the generations by the Seq2Seq models given the input text combining general and finding sections. In this example, words in red are unfaithful generations comparing to the input and extracted information (highlighted in green) that appears in the source sequence.

abstractive models given the input from the general and findings sections in the report. The date of the previous radiology examination is very important information for short-term or long-term response assessments. The baseline BERT2BERT model tends to predict more new phrases and always generate a spurious date. While the other hybrid models are able to address this issue and more constraint to the original phrases from the source input.