Richard Bieck*, Valentina Wildfeuer, Viktor Kunz, Martin Sorge, Markus Pirlich, Max Rockstroh, Thomas Neumuth

Generation of surgical reports using keyword-augmented next sequence prediction

Abstract: The documentation of a surgical procedure remains a time-consuming task that surgeons must incorporate into their daily routine. However, since a surgical report should be produced immediately after the operation with all impressions of the procedure in mind, a means of automation assistance should be provided. We, therefore, propose a method that generates surgical reports based on keywords stated during the procedure. Our report generation is based on a sequence-to-sequence model that is trained on sentence pairs of two consecutive sentences in a surgical report. The known sentence is augmented with a keyword based on the following surgical action to be documented and is then passed into a language model to generate the next sentence. In this way, the complexity of predicting a vast number of possible surgical report phrasings is reduced to a next sentence prediction task. For the language model, an encoder-decoder structure was used with bidirectional 2-layer Long-Short Term Memory (LSTM) units for both components and an attention layer between input and output sentences. The training data consisted of 50 ear-, nose- and throat surgery (ENT) reports with 1500 sentences. The model training was performed in a k-fold cross-validation study with $k = 10$ and cross-entropy loss as the objective function. The generated reports were investigated using NIST, ROUGE, and METEOR metrics. Additionally, three medical experts identified the report content regarding plausibility and text errors. The trained models reached an accuracy of 0.82 for the next sentence predictions. The generated reports show consistent sentence structures and keyword correspondence for about 70 % of

provided keyword sequences. The NIST, ROUGE, and METEOR metrics reached 0.65, 0.71, and 0.64, respectively. The model underperformed for not yet known keyword sequences and shows signs of overfitting when keyword sequences deviate from the baseline of the training set. Our approach for the keyword-augmented generation of surgical reports shows the potential of reducing the text generation complexity by providing a sequence of anchor words. However, the automated generation of surgical reports remains a difficult task due to individual report phrasings and the high variance in keyword sequences.

Keywords: Natural Language Processing, Surgical Reports, Machine Learning, Next Sequence Prediction, Language Model, ENT.

<https://doi.org/10.1515/cdbme-2021-2098>

1 Introduction

The application of machine learning for work simplification and automation already plays a significant role in many areas of daily life. It is, therefore, no surprise that this trend from consumer electronics creates a demand from surgical experts for new automation functions in their daily clinical routine. There exists the enormous potential of using machine learning to relieve medical personnel of non-treatment-related and time-consuming tasks which opens up new possibilities of clinical workflow optimization, documentation, and operations efficiency [1], [2]. In particular, the process of surgical documentation could be further exploited and automated through the processing of intraoperative commentary and the use of enabling technology like automated speech recognition (ASR) [3].

In standard clinical practice, surgical reports have to be created after the procedure has been completed in the form of a free text. In this context, some physicians report defined masks that serve as a standardized template of OR reports and are

*Corresponding author: **Richard Bieck:** Innovation Center Computer Assisted Surgery, University Leipzig, Semmelweisstraße 14, Leipzig, Germany, richard.bieck@medizin.uni-leipzig.de

Valentina Wildfeuer, Viktor Kunz, Martin Sorge, Marcus Pirlich: Department for Ear, Nose and Throat Surgery, University Hospital Leipzig AöR, Leipzig, Germany

Max Rockstroh, Thomas Neumuth: Innovation Center Computer Assisted Surgery, University Leipzig, Leipzig, Germany,

presumably superior to conventional free-text [4]. Depending on the type and extent of the procedure, compiling a surgical report can be time-consuming and error-prone, since the exact course of the procedure must always be reproduced from memory.

We, therefore, propose a machine learning-based method to automatically generate surgical reports from a set of intra- or postoperatively dictated keywords. In contrast to automatic report generation in radiology [5], we implement a language model with an encoder-decoder architecture to generate a summarized surgical report without any additional image or patient information as model input.

2 Methods

2.1 Report Generation System

We hypothesize that a language model can generate a surgical report one sentence at a time in an autoregressive manner with procedure-specific keywords voiced intraoperatively by a surgeon. More specifically, we consider a sequence of uttered keywords to correlate with the logical topic structure of surgical reports and, therefore, allow a trained language model to create plausible documentation. For the keyword-augmented generation of surgical reports, we consider a system design that is centered around the commentary of the performing surgeon as a crucial information source (see fig 1). The surgeon is encouraged to describe procedure steps and observations that are noteworthy for documentation and the record of these descriptions is transcribed into a sequence of keywords known to the report generating system. For our approach, we assume that an ideal speech-to-text component, that is robust against any form of noise produced in the environment, exists in the operating room. In this paper, we

are focussing on the implementation of the sequence-to-sequence model that performs sentence generation.

2.2 Data Selection & Preparation

We retrospectively gathered 50 anonymized surgical reports of performed panendoscopic procedures. The reports had a length of between 10 and 45 sentences and were authored by 10 different surgeons with 5 to 15 years of experience. Overall, a database of 1500 sentences was compiled with a mean sentence length of 8 words. In the first preprocessing step, a vocabulary of 150 keywords was defined together with surgical experts that reflect the main topics addressed in report sentences. The goal was to establish a set of anchor words that would likely be used in surgical commentary during the procedure and could be incorporated in the language model training. Subsequently, the keywords were assigned to all report sentences according to topics determined by two surgical experts. In this way, between one and four keywords were assigned to each sentence. Every two consecutive sentences of a surgical report were then paired to compile individual training samples consisting of a source and target sentence and the keyword of the target sentence attached to the source sentence.

The prepared sentence pairs were then tokenized to obtain a numeric representation for the model training. We chose to train an unsupervised word tokenizer using SentencePiece¹ for this purpose as it offered several advantages for the report generation task. The tokenizer uses a fixed vocabulary size which is predetermined before the neural model training using sub-word regularization from Kudo [6]. This allows the trained language model to be independent of language-specific issues such as open-endedness of vocabularies. Additionally, it introduces a virtual augmentation step that increases model robustness in the context of a vocabulary with imbalanced

Figure 1: Concept for the automated generation of surgical reports using intraoperatively recorded keywords spoken by the performing surgeon.

¹ <https://github.com/google/sentencepiece>

word occurrences. We considered the vocabulary size for our tokenizer as a training hyperparameter and determined a size of $s = 500$ to be the optimum through trial-based elimination.

2.3 Model Design & Training

For the language model, we chose an encoder-decoder-architecture consisting of bidirectional 2-layer LSTM units for both coders. The encoder maps the tokenized input sentences to a continuous intermediate representation known as latent features. The decoder then generates an output sentence one word at a time based on the previous word and the latent features of the input sentence. The decoder model is autoregressive and uses the last output word as input for the generation of the subsequent word to maximize the conditional probability

$$y = \operatorname{argmax} \prod_{i=0}^n p(y_i|x) \quad (1)$$

for all possible output words y_i . Our report sentences follow an intrinsic word distribution specific for ENT surgical reports of that department and the decoder training maximizes the likelihood of achieving the right word selection for new report sentences to mimic this word distribution. Additionally, we adopted the attention layer from Chan et al. [7] that calculates a weighting context between the latent features of the encoder and the already generated output words for each word generation step. This enables the decoder to better train single word-specific dependencies between the input and output sentence. Both the encoder and decoder have an embedding layer with embedding size $n=512$ that translates our numeric word representations into a vector form that is optimized for GPU-based model training.

For the model training, a k-fold cross-validation study with $k = 10$ was set up for our 50 report datasets. The sentences of left-in datasets were compiled into one dataset, shuffled, and split into training and validation sets with an 85:15 - ratio. At the beginning of the training pipeline, we applied data augmentation using a sequence of random swapping and deletion of up to two words. We did not apply model-pretraining since there currently does not exist any large corpus of German medical reports for a potential baseline language model training.

The language model was trained using an Adam optimizer with an initial learning rate of $l_r = 0.5 \cdot e^{-3}$ that was decayed every 4 epochs by a factor of 0.5, a batch size of $b = 12$, and cross-entropy loss as the objective function. The optimization

process was controlled using the f1-score (f1). The process was terminated when changes of either metric fell below 1 % between epochs. In this way, models were trained over 28 epochs on average (range 27-30) with an average model computation time of 22.5 minutes. The stated hyperparameter values represent the final training configuration after semi-automatic parameter search for batch size, learning rate, and optimization step size. All data processing and model computations were performed using Pytorch v1.7.0² with Python v3.8 and on an NVIDIA RTX 2070S with 8GB RAM.

2.4 Validation Study

The trained language models were then used to generate surgical reports from the left-out keyword sequences. Overall, 1500 sentence pair translations were performed. Using a starting sentence “*Start of procedure*” in combination with the individual keywords the following sentences were generated one word at a time. For the word candidate search, we used beam searching with a beam size of $n=32$. The quality of the generated report sentences was then measured using the NIST, ROUGE, and METEOR metrics. The NIST score is a sentence-level metric and calculates the arithmetic mean distance between ground truth and predicted sentences whereas ROUGE and METEOR can be calculated on a word level. For sentences where a metric score of 0.1 or lower was observed, the content was explicitly compared with the corresponding keyword to investigate whether the model still performed in a reasonable form. Additionally, three medical experts identified 3 randomly chosen reports and were tasked to correct the reports and appraise them in terms of plausibility.

3 Results

The model training and validation study were performed as planned. The models trained in the cross-validation study reached a mean f1-score of 0.56 with an accuracy of 0.82 and a recall of 0.43. The NIST, ROUGE, and METEOR metrics reached 0.65, 0.71, and 0.64, respectively. About 25 % of the generated sentences had a ROUGE and METEOR metric of under 0.25, with about 10 % having a metric below 0.1. For sentences with the lowest scores the comparison with the respective keywords still showed plausible semantic relation to the keyword. For the examination of generated reports by

² <https://pytorch.org/>

medical experts, the number of mean corrections per report was 5.4 consisting of 3.4 word deletions and 2.0 word additions. Most sentences were valued as plausible while faulty sentences were being described as redundant or superfluous leading to longer reports with no additional content.

4 Discussion

We successfully implemented a language model that generates surgical reports of panendoscopic procedures using a sequence of previously acquired keywords. Despite the implementation of a customized tokenizer and the application of augmentation functions, the model shows signs of overfitting when keyword sequences deviate from the baseline of the training set. However, our approach for the keyword-augmented generation of surgical reports shows the potential of reducing the text generation complexity to an autoregressive task by providing a sequence of keywords. The automated generation of surgical reports remains a difficult task due to individual report phrasings and the high variance in keyword sequences. Additionally, the assumed Markov property of two consecutive sentences may be too restricting. We expect a performance increase with the scaling of training data to 100 or more surgical reports and also believe additional validation metrics are needed to take semantic overlap of medical statements into account. Approaches like word rotator distance proposed by Yokoi et al. [8] that consider word vector norms between sentences may be more suitable for the assessment of generated reports.

5 Conclusion

We presented a method to automatically generate surgical reports from a sequence of keywords that represent procedure steps suitable for documentation. Further improvements need to focus on the extension of the database as well as the performance metrics for the comparison of generated sentences to the ground truth. To our knowledge, this is the first approach that attempts the automatic generation of surgical reports without additional image or patient data.

Author Statement

Research funding: Funding is provided by the German Federal Ministry of Education and Research (BMBF) in the strategic innovation program “KMU Innovativ – Mensch-Technik-Interaktion” under the grant number 16SV8017. **Conflict of interest:** Authors state no conflict of interest. **Informed consent:** Informed consent has been obtained from all individuals included in this study. **Ethical approval:** The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the author’s local ethics committee (455/20-ek).

References

- [1] M. Willis, P. Duckworth, A. Coulter, E. T. Meyer, and M. Osborne, “Qualitative and quantitative approach to assess of the potential for automating administrative tasks in general practice,” *BMJ Open*, vol. 10, no. 6, p. e032412, Jun. 2020, doi: 10.1136/bmjopen-2019-032412.
- [2] L. Letourneau-Guillon, D. Camirand, F. Guilbert, and R. Forghani, “Artificial Intelligence Applications for Workflow, Process Optimization and Predictive Analytics,” *Neuroimaging Clinics of North America*, vol. 30, no. 4, pp. e1–e15, Nov. 2020, doi: 10.1016/j.nic.2020.08.008.
- [3] H. Feussner et al., “Surgery 4.0,” in *Health 4.0: How Virtualization and Big Data are Revolutionizing Healthcare*, C. Thuemmler and C. Bai, Eds. Cham: Springer International Publishing, 2017, pp. 91–107. doi: 10.1007/978-3-319-47617-9_5.
- [4] Ö. Eryigit, F. W. van de Graaf, and J. F. Lange, “A Systematic Review on the Synoptic Operative Report Versus the Narrative Operative Report in Surgery,” *World J Surg*, vol. 43, no. 9, pp. 2175–2185, Sep. 2019, doi: 10.1007/s00268-019-05017-8.
- [5] M. M. A. Monshi, J. Poon, and V. Chung, “Deep learning in generating radiology reports: A survey,” *Artificial Intelligence in Medicine*, vol. 106, p. 101878, Jun. 2020, doi: 10.1016/j.artmed.2020.101878.
- [6] T. Kudo, “Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates,” *arXiv:1804.10959 [cs]*, Apr. 2018, Accessed: Jul. 05, 2021. [Online]. Available: <http://arxiv.org/abs/1804.10959>
- [7] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, Attend and Spell,” *arXiv:1508.01211 [cs, stat]*, Aug. 2015, Accessed: Jul. 31, 2020. [Online]. Available: <http://arxiv.org/abs/1508.01211>
- [8] S. Yokoi, R. Takahashi, R. Akama, J. Suzuki, and K. Inui, “Word Rotator’s Distance,” *arXiv:2004.15003 [cs]*, Nov. 2020, Accessed: Jul. 05, 2021. [Online]. Available: <http://arxiv.org/abs/2004.15003>