Patient-friendly Clinical Notes: Towards a new Text Simplification Dataset

Jan Trienes†  Jörg Schlötterer†  Hans-Ulrich Schildhaus‡  Christin Seifert†
†University of Duisburg-Essen, Germany  ‡University Hospital Essen, Germany

{jan.trienes, joerg.schloetterer, christin.seifert}@uni-due.de  hans-ulrich.schildhaus@dls.com

Abstract

Automatic text simplification can help patients to better understand their own clinical notes. A major hurdle for the development of clinical text simplification methods is the lack of high quality resources. We report ongoing efforts in creating a parallel dataset of professionally simplified clinical notes. Currently, this corpus consists of 851 document-level simplifications of German pathology reports. We highlight characteristics of this dataset and establish first baselines for paragraph-level simplification.

1 Introduction

Many hospitals worldwide give patients access to their own clinical notes with the goal to strengthen patient autonomy and increase transparency of the care process (Delbanco et al., 2012). Yet, clinical notes are seldomly written with the patient in mind. Being a communication tool for doctors, clinical notes must use a precise and unambiguous medical vocabulary. With limited health literacy, these notes are therefore practically inaccessible to most patients (Sørensen et al., 2015). The urgency of making clinical notes accessible to patients is underlined by initiatives like “What’s my diagnosis?” where medical doctors and students volunteer to translate patient notes into a simple language (Bittner et al., 2015). Approaches to automatic text simplification (TS) have the potential to assist with this time consuming manual process (Shardlow, 2014; Alva-Manchego et al., 2020).

However, there is a lack of resources to develop TS methods for clinical notes. Most commonly used resources for TS include, on the one hand, professionally simplified news articles such as Newsla (Xu et al., 2015) and OneStopEnglish (Vajjala and Lučić, 2018), and on the other hand, large scale but potentially noisy alignments of Wikipedia (Zhu et al., 2010; Jiang et al., 2020). In the medical domain, datasets cover consumer health lexicons (Cao et al., 2020), laymen summaries of scientific articles (Devaraj et al., 2021) and medical subsets of Wikipedia (Grabar and Cardon, 2018; van den Bercken et al., 2019; Van et al., 2020). In addition, there is a lack of parallel document-level TS datasets (with the notable exception of Newsla and OneStopEnglish). This makes it difficult to study document-level phenomena of TS such as sentence reordering, deletion and explanation generation (Alva-Manchego et al., 2019b; Zhong et al., 2020; Srikanth and Li, 2021).

Contributions. As a response to the lack of resources, we report initial results towards a parallel dataset of patient-friendly clinical notes. Currently, the dataset covers 851 German surgical pathology reports with document-level simplifications by medical professionals. We describe a lightweight simplification protocol that encourages an intuitive simplification style and medical correctness, while not demanding linguistic training on the side of the editors. Furthermore, we conduct first analyses of the dataset and establish baselines for automatic simplification. Overall, we find that the source text is dense, with partly incomplete sentences and a highly specialized vocabulary. We observe a variety of simplification operations in the references including lexical substitution, syntactic simplification, explanations and content selection. We believe that this dataset can be an interesting testbed for research on document-level TS.

Our goal with this dataset is to increase the diversity of resources that TS research has available in terms of the text types, domains and languages. Sharing this dataset with the research community while respecting ethical, legal and privacy standards is currently underway.  

1We share the experiment code and provide updates on the dataset at github.com/jantrienes/simple-patho.
2 Dataset Creation and Analysis

We describe our design decisions for the creation of a parallel corpus of clinical notes. An example report is given in Table 1.

2.1 Data Selection

We decided to focus on pathology reports of sarcoma patients since clinicians noted particularly high amounts of questions concerning these reports. Sarcomas are a rare type of cancer with many subtypes which can affect people of all ages. The pathology report describes an analysis of tumor tissue and establishes the main diagnosis.

We sample reports from the electronic health records of the University Hospital Essen, a large research hospital in Germany. Each year, about 60,000 pathology reports are written by the pathology department. We identify suitable reports based on clinical codings (ICD-O-M). A query for the period of January 2019 until August 2021 yielded 1,644 reports on sarcoma patients. All reports were fully anonymized and we received ethics approval from our institutional review board.2

2.2 Simplification Protocol

To create a parallel corpus of original and simplified clinical notes, we ask medical experts how they would intuitively explain a given report to a patient. We take a decidedly inductive approach here: while guidelines for simplified language exist,3 it is not clear to what extent these are suitable for clinical notes, and if annotators without formal linguistic training could operationalize them. In the terminology of Allen (2009), we use an intuitive rather than a structural simplification process.

It is commonly accepted that a good simplification depends on the target audience (Xu et al., 2015; Bingel et al., 2018; Gooding, 2022). To better define the audience and ensure a common simplification goal among editors, we developed

---

2University of Duisburg-Essen; Reference: 21-10198-BO
3For example Basic English (Ogden, 1930); we refer to Saggion (2017) and Štujner (2021) for more examples.
a *patient persona*. A persona is a rich description of a prototypical user of a software system, a tool often used in human-computer interaction research (Cooper, 1999). With this persona at hand, editors were asked “What questions would this patient have about the report?” Additionally, we provided following simplification guidelines to further increase consistency across editors: (i) preserve the section structure of the reports, (ii) use the same tense as the original report, and (iii) do not add any interpretations that go beyond the stated facts.

We hired a team of 9 medical students in their fourth year of studies. A senior pathologist provided guidance on clinical questions during regular meetings and through email. All reports were simplified by one editor at the document-level using a plain text editor with grammar and spellchecking functionality. We implemented several quality gates for consistency and medical correctness of simplifications. First, we used an initial trial period of 10 reports to refine the guidelines and to allow editors to get familiar with the task. Second, we held monthly meetings to discuss simplification challenges and examples. A chat platform was setup to resolve urgent questions in a timely manner. Over the span of one year, we simplified 851 reports with a total effort of 812 hours (median 50 min./report). The students were compensated for their work with 10.5€ per hour corresponding to the usual rate for student assistants in Germany.

### 2.3 Preprocessing

For studying the characteristics of our TS corpus, we apply minimal pre-processing. We segment each document into sentences and tokens using NLTK. To establish a reliable vocabulary size, we lemmatize the text using spaCy and replace tokens that only consist of digits, punctuation or combinations thereof with a special token.4

We found that most reports consist of four core sections: background, microscopy, macroscopy, and conclusion. Therefore, we also compile a section-aligned version of the dataset where we keep reports that have a one-to-one alignment for all core sections (820 out of 851 reports). This makes our dataset also amenable for paragraph-level simplification (Devaraj et al., 2021) in addition to document-level simplification.

#### 2.4 Dataset Characteristics

To better characterize the dataset, we analyze several surface-level properties (see Table 2 for an overview). We focus on measures that were commonly reported in prior work (Xu et al., 2015; Dmitrieva and Tiedemann, 2021) including the number of sentences and tokens, the vocabulary size (types), the length of documents and sentences, and the n-gram novelty. The type-token ratio (TTR) is used as a measure of lexical diversity and the Flesch Reading-Ease (FRE, Flesch, 1948) serves as a first indication of changes in readability.5

Simplifications are on average 41% longer than the original text (Table 2). Through manual inspection, we identified two potential reasons. First, the original reports tend to use a brief writing style with partly incomplete sentences. These were expanded to full sentences by the editors. Second, editors often added contextual information and explanations (e.g., why an examination was done, and what the result mean to a patient). The most striking difference in length can be observed for the background section (Figure 1). We assume that simplifications are “setting the scene” in this section by simplifying terminology and explaining concepts which do not have to be repeated again in the remainder of the report.

Simplifications select and summarize content. While simplifications are longer than their original counterpart, we also note a form of summarization (see example in Table 1). In some cases, particularly technical concepts were not included in the simplification, presumably because there is no simple explanation or because an explanation

---

4[nlk.org](https://www.nltk.org) and [spaCy.io](https://spacy.io)

5We use constants adapted to German text (Amstadt, 1978). Implementation in [github.com/textstat/textstat](https://github.com/textstat/textstat).
would not help a user to better understand the report. This is in line with prior work which argues that document-level TS also requires summarization (Zhong et al., 2020; Aumiller and Gertz, 2022).

Simplifications have a different and more constrained vocabulary. While the simplified corpus is substantially larger in the number of tokens, the vocabulary size has only slightly increased. This is reflected in the lexical diversity measure (TTR: 0.47 → 0.42, Table 2). A decrease in TTR indicates that simplifications use a more constrained vocabulary which might translate to better readability. Furthermore, we observe a high average rate of unigram novelty (around 63%), which signals that large parts of the vocabulary are not shared.

Simplifications have a slightly higher readability. We observe a small increase in the readability measure (FRE: 32.9 → 40.3, Table 2). However, the overall readability is low according to this measure. By means of comparison, Aumiller and Gertz (2022) reported FRE values of 40 for the original and 67 for the simplified parts of a document-level TS dataset collected from German Wikipedia. There are inherent limitations with readability measures like FRE, so this finding has to be interpreted with care (Tanprasert and Kauchak, 2021).

3 Simplification Baselines

We next establish a first baseline for pathology report simplification using paragraph-level sequence-to-sequence methods (Devaraj et al., 2021).

3.1 Modeling Considerations

As discussed in Section 2.4, our dataset features multiple simplification operations including lexical simplification, paraphrasing, summarization and explanation generation. Therefore, we focus on monolingual neural machine translation models which can learn these operations simultane-ously (Nisioi et al., 2017). Prior work on medical text investigated lexical simplification (Abrahamsson et al., 2014; Kloehn et al., 2018) or hybrid systems that combine pre-trained translation models with domain-specific phrase tables (Shardlow and Nawaz, 2019). With our parallel dataset, fine-tuning large general-purpose language models becomes a realistic option (Rothe et al., 2020).

Inspired by Devaraj et al. (2021), we train a paragraph-level simplification model. Compared with sentence-level methods, a paragraph-level model has the benefit that we do not need sentence alignments (Štajner et al., 2018) and that we can capture simplification phenomena like syntactic simplification and summarization (Alva-Manchego et al., 2019b). Our dataset has a natural paragraph-level alignment in the form of four core sections, so we consider this a suitable first baseline.

Methods. We experiment with four instantiations of paragraph-level methods. (1) Identity: A simple baseline which outputs the original text as simplification. (2) Bert2Bert: A transformer-based encoder-decoder where both parts are initialized with BERT (Devlin et al., 2019; Rothe et al., 2020). (3) Bert2Share: Same as Bert2Bert, but weights of the encoder and decoder are shared. (4) mBART: A sequence-to-sequence transformer, pre-trained on a sentence reconstruction objective (Liu et al., 2020). We include hyperparameters and replication details in Appendix A.

Evaluation. We report the standard TS metrics SARI (Xu et al., 2016), BLEU (Papineni et al., 2002) and ROUGE F1 (Lin, 2004) for unigram (R-1) and bigram (R-2) matches, and the longest common subsequence between the reference and system output (R-L). To calculate SARI and BLEU, we use the implementation in EASSE (Alva-Manchego et al., 2019a) with default settings. For ROUGE,
Table 3: Automatic simplification results on paragraph-aligned data. The identity baseline simply returns the input as simplification. For the reference simplification, the average length (Len.) is 132 tokens and the average unigram novelty (Nov.) is 70% (cf. Table 2).

Table 4: Evaluation by report section. Micro averaged metrics over all sections are reproduced from Table 3.

we use the rouge-score package with stemming disabled. We randomly split reports into training/validation/test sets with an 80/10/10 ratio.

3.2 Results and Discussion

Quantitative Results. According to automatic metrics, the generated simplifications have a substantially higher simplicity (SARI) but only slightly higher adequacy (ROUGE and BLEU) than an identity baseline (Table 3). mBART provides best results with an average simplification length and novelty close to the reference (129 vs. 132 tokens, and 65% vs. 70% novelty, Table 3). While not directly comparable, metrics are in a similar range as the paragraph-level simplification results on English medical abstracts by Devaraj et al. (2021).

For a better intuition of where the models can be improved, we report metrics by section type in Table 4. We see that the background section is most difficult to simplify. The low BLEU score of the identity baseline (0.1 in Table 4) indicates that there is little overlap between the original and simplified vocabulary. We hypothesize that simplifications for the background section include explanations and contextual domain knowledge which are difficult to generate with sequence-to-sequence methods (Srikanth and Li, 2021).

Qualitative Observations. By manual inspection, we found that system outputs are mostly fluent, grammatical and subjectively easier to read (Table 1). Furthermore, we observe that models generate elaborations and perform a certain degree of content selection. We also found factual errors in the automatically generated simplifications. In the example in Table 5, a clinical result was reported as positive in the original report but negative in the generated simplification (STAT6 positive vs. negative). The subsequently generated sentence (“This combination of tumor markers is suggestive of GIST”) is a clinically conceivable statement, but in the context of this report wrong. We anticipate that factual correctness will be of high importance for any practical deployment of a TS system for clinical notes and consider the evaluation of factual correctness as a significant avenue for future work on this dataset (Devaraj et al., 2022).

4 Conclusion and Future Work

We present ongoing work towards a dataset of professionally simplified clinical notes. Currently, the corpus consists of 851 parallel documents totaling close to 790k tokens. Quantitative and qualitative analyses show potential challenges for paragraph-level and document-level TS research. Despite a moderately sized training set, fine-tuning general language models led to promising results.

In future work, we will increase the size of the dataset and conduct a formal analysis of the simplification operations in the data to better understand the challenges for TS on clinical notes. Human evaluations with a focus on factual correctness, as well as user studies with end-users such as patients and patient advocacy groups are also envisioned.

Acknowledgements

We thank Celina Bandowski, Theresa Bernsmann, Monika Coers, Lisa Gödde, Hannah Göke, Lisa Meißner, Ral Merjanah, Justin Roschlak and Chiara Wedekind for writing the simplifications. We also thank Andrea Tonk for helping to translate the example report into English.

References


Alan Cooper. 1999. *The inmates are running the asylum*. In *Software-Ergonomie ’99*, volume 53 of *Berichte des German Chapter of the ACM*.


Nicholas Kloehn, Gondy Leroy, David Kauchak, Yang Gu, Sonia Colina, Nicole P. Yuan, and Debra Revere. 2018. Improving consumer understanding of medical text: Development and validation of a new SubSimplify algorithm to automatically generate term explanations in English and Spanish. *Journal of Medical Internet Research*, 20(8).


Charles Kay Ogden. 1930. Basic English: A general introduction with rules and grammar.


A Implementation Details

Hyperparameters. All simplification models were trained for 25 epochs using the AdamW optimizer with an initial learning rate of $3e^{-5}$ (Kingma and Ba, 2015; Loshchilov and Hutter, 2019). We use a learning rate schedule with an initial warmup period of 10% of the training steps and a linear decay afterwards. Checkpoints are taken every epoch and the checkpoint with lowest validation loss is kept. For Bert2Bert and Bert2Share, we set the batch size to 16 and for mBART to 4. During inference, we use beam search decoding with 5 beams. Generation ends when an end-of-sequence token is generated. We did not perform any manual or automatic hyperparameter tuning.

Implementation. To adapt mBART for the task of monolingual translation, we follow recommendations by Rios et al. (2021) and add a special language token for the original text and for the simplified text. We implement the models using the Transformers library (Wolf et al., 2020). Models are initialized with the bert-base-multilingual-cased and facebook/mbart-large-cc25 checkpoints.

Computation Cost. All models are trained on a single NVIDIA RTX A6000 GPU with 48GB of memory. Training duration is 2:45h for mBART, 1:30h for Bert2Bert and 1:16h for Bert2Share.
3. Mikroskopie


Translation: Human Reference: After being processed into narrow slices and further preprocessing (e.g. staining), the tissue samples can be viewed under the microscope. The samples can be viewed under the microscope, parts of a tumor consisting of spindle-shaped cells can be seen. The tumor itself shows a rather homogeneous appearance. The blood vessels grow in a particular pattern. Many small, branched vessels are visible. The tumor cells show enlargement, slightly vesicular nuclei. Nuclei are in a place in a cell where genetic material is stored as DNA. Mitotic figures are chromosome structures that appear during cell division and are visible under the microscope.

3. Microscopy

Original: Microscopically, the biopsy shows portions of a spindle-cell shaped tumor. The tumor has a relatively monomorphic appearance, with a hemangioepitelytoma-like vascular pattern. The tumor cells have enlarged, slightly vesicular nuclei. Mitotic figures are barely visible (1/10 HPF). The stroma is relatively fine and contains single collagen fibers. Necroses are not detectable. Additionally, immunohistochemical examinations were conducted. The tumor shows strong positivity for CD34 and strong nuclear expression of STAT6. The following antigens are not expressed by the tumor: Actin, caldesmon, pancytokeratin (CKplus), desmin, EMA, MUC4, S100, SOX10, and TLE1.

4. Epikritische Stellungnahme

Original: PE (proximaler Oberschenkel ventral links) mit einem solitären fibrösen Tumor (SFT), klassischer Typ. Zur Validierung des histologischen Befundes wurden ergänzende immunhistochemische Untersuchungen (Fusionspanel) eingeleitet. Diesbezüglich folgt ein Nachbericht.

Translation: System Output (mBART): The prepared tissue samples were examined microscopically to elucidate the type of tissue and the spread of the tumor. Additionally, antibody examinations (immunohistochemical examinations) were performed to further determine the tumor. A microscopic examination revealed portions of a tumor consisting of spindle-shaped cells. The tumor cells look altered compared to healthy cells and deviate from the normal. Only a small number of cells are in the process of cell division (1 mitotic figure/ 10HPF), suggesting that the tumor grows slowly. In addition, immunohistochemical examinations were performed. In these examinations, antibodies are used to detect cellular components that are typically found in tumors and thus provide information about the type and origin of the tumor. Of the tumor markers tested (CD34, STAT6, actin, caldesmon, pancytokeratin (CKplus), desmin, EMA, MUC4, S100, SOX10 and TLE1), CD34 was positive and STAT6 was negative. This combination of tumor markers is suggestive of the presence of a gastrointestinal stromal tumor (GIST) or a metastasis of GIST.

R: 1/2/L: 30.92/9.02/27.93, BLEU: 5.12, SARI: 35.94

4. Conclusion

Original: BX (proximal thigh ventral left) with a solitary fibrous tumor (SFT), classic type. A supplementary molecular pathological examination (fusion panel) was initiated to validate the findings. There will be a follow-up report on this topic.

Translation: System Output (mBART): By means of histological examinations the diagnosis was confirmed and it is about a solitary fibrous tumor (SFT), classic type. A supplementary molecular pathological examination (fusion panel) was initiated to validate the findings. There will be a follow-up report on this topic.

Table 5: Example report (continued from Table 1).