Klexikon: A German Dataset for Joint Summarization and Simplification

Dennis Aumiller and Michael Gertz
Institute of Computer Science, Heidelberg University
Im Neuenheimer Feld 205, Heidelberg, Germany
{aumiller, gertz}@informatik.uni-heidelberg.de

Abstract

Traditionally, Text Simplification is treated as a monolingual translation task where sentences between source texts and their simplified counterparts are aligned for training. However, especially for longer input documents, summarizing the text (or dropping less relevant content altogether) plays an important role in the simplification process, which is currently not reflected in existing datasets. Simultaneously, resources for non-English languages are scarce in general and prohibitive for training new solutions. To tackle this problem, we pose core requirements for a system that can jointly summarize and simplify long source documents. We further describe the creation of a new dataset for joint Text Simplification and Summarization based on German Wikipedia and the German children’s encyclopedia “Klexikon”, consisting of almost 2,000 documents. We release a document-aligned version that particularly highlights the summarization aspect, and provide statistical evidence that this resource is well suited to simplification as well. Code and data are available on Github: https://github.com/dennlinger/klexikon

Keywords: Summarization, Text Simplification, German

1. Introduction

The goal of Text Simplification (TS) is to produce easily understandable texts that benefit disadvantaged readers such as children, dyslexic, or language learners. Simplifications are often generated by adapting a source text written for adult/native readers. However, recent work in simplification has mostly addressed TS as a monolingual translation task, where individual sentences are "translated" into a simplified version (Zhu et al., 2010; Coster and Kauchak, 2011; Hwang et al., 2015). The main focus is put on either lexicographic replacements, paraphrasing, sentence splitting, or the dropping of words within a single sentence (Amancio and Specia, 2014), which implies that the simplification of any input document will consist of roughly the same number of sentences. While this approach is appropriate for sufficiently short source documents, longer articles become strenuous for disadvantaged readers. As can be seen in Table 1, articles in different corpora come with varying lengths of their respective source texts. When simplifications are generated via manual sentence-by-sentence translations, the simplified texts tend to have more sentences than the source documents. When alignments are constructed from a source and simplification text on the same topic instead, they exhibit a drastic length disparity. Current simplification systems are, however, inherently limited in their ability to address the problem of joint simplification and summarization from much longer input documents. Sentence-level alignments were traditionally seen as one way to circumvent certain problems in TS, namely:

1. Human feedback for judging simplification quality is more consistent for sentences, compared to longer samples, such as entire documents.

In this work, we present remedies to the problem of missing document alignments, and argue that the inclusion of summarization into the broader context of Text Simplification is a necessary step towards end-to-end solutions for longer input texts. Specifically, it addresses the following problems:

1. Long-form documents can be compressed into significantly shorter summarized simplifications.
2. Document alignments provide context for models that are otherwise based on single sentence pairs.
3. The amount of accessible training data increases, which is especially important for languages other than English, where data is generally scarce.

| Resource          | Articles | Aligned | Avg. #Sentences Source | Simple |
|-------------------|----------|---------|------------------------|---------|
| Klexikon (Ours)   | 2,808    | 242.09  | 32.61                  |         |
| (Hewett and Stede, 2021) | 978    | 10.12   | 43.54                  |         |
| Battisti et al., 2020 | 378    | 45.29   | 55.75                  |         |
| (Kauchak, 2013)   | 59,775   | 64.52   | 8.46                   |         |
| (Xu et al., 2015)* | 1,130   | 49.59   | 51.27                  |         |

Table 1: Corpus statistics for datasets with document alignments in German (top) and English (bottom). * indicates resources created by simplifying articles sentence-by-sentence. For (Xu et al., 2015; Hewett and Stede, 2021), we refer to the respective simplified corpora with simplification level 1.

2. Metrics such as BLEU (Papineni et al., 2002) or SARI (Xu et al., 2016) rely on (aligned) reference texts for automated evaluation.
3. Prior alignment of sentences limits the length of input samples, which is essential for algorithms with non-linear runtime, or length constraints.

In this work, we present remedies to the problem of missing document alignments, and argue that the inclusion of summarization into the broader context of Text Simplification is a necessary step towards end-to-end solutions for longer input texts. Specifically, it addresses the following problems:
Simultaneously, TS offers interesting challenges to the summarization community, which hopefully facilitates exchange between the two fields: On existing summarization datasets, simply taking the leading three sentences offers strikingly good results (Nallapati et al., 2017), which may lead to systems learning specific extractive strategies instead of generalizing to broader textual relevance. Preliminary experiments show that our dataset poses a harder challenge for summarization systems, due to the additional simplification aspect. Our proposed resource was obtained from semi-automated alignments between the German Wikipedia and the children’s encyclopedia "Klexikon" (Schulte and van Dijk, 2015), written for children aged 8 to 13 years. With almost 2,900 articles, it is the largest non-English resource with document alignments.

2. Related Work

Related work can broadly be categorized into relevant simplification work, and associated works on resources for (German) summarization datasets.

2.1. Text Simplification

Previously mentioned work frequently deals with data aligned based on Simple Wikipedia (Zhu et al., 2010; Coster and Kauchak, 2011; Hwang et al., 2015). The main differences between these approaches lie in their alignment strategies and underlying simplification model. The only work on Simple Wikipedia that specifically introduces a document-aligned version is (Kauchak, 2013), who investigates performance gains from supplementing language models with additional (non-simplified) texts. Importantly, it is not explicitly used for learning simplification. (Hancke et al., 2012) introduced a first German resource containing simplified texts based on unaligned articles from GEO and GEOlino, a German magazine similar to National Geographic, and its edition specifically for children. They build a classification system that is able to classify between normal and simplified texts for several article categories. A larger and improved version from the same source was collected by (Weiß and Meurers, 2018), who also introduce a resource based on transcripts from German TV broadcasts (Tagesschau/Logo!), again without any alignment. The first mention of an aligned corpus for German can be found in (Klaper et al., 2013), who automatically align websites with their corresponding versions in accessible language. Their corpus contains a total of about 270 articles.

Most recently, (Battisti et al., 2020) collected a larger corpus, where 378 texts contain document alignments. Arguably, unaligned resources might still be helpful to facilitate pre-training of models. In an attempt to circumvent data scarcity, (Mallinson et al., 2020) employ multi-lingual pre-training, which they tested with a small, manually labeled German evaluation set. To our knowledge, (Hewett and Stede, 2021) were the first to utilize alignments between Wikipedia and Klexikon, with an additional extension to MiniKlexikon, a secondary simplification level. Due to the further required alignments, the overall size of their data is about 10% of our presented corpus. To avoid problems stemming from extreme length discrepancies, they also only extract introduction and abstracts for Wikipedia articles, which is something we explicitly encourage in our version. This also explains the different lengths while using the same document sources, as reported in Table 1.

2.2. Summarization

(Parmanto et al., 2005) are the first to explicitly explore summarization and simplification in a common context, albeit for the task of website accessibility. Further work models summarization itself as a simplification technique, e.g., (Margarido et al., 2008) investigated extractive summarization approaches and how they help disadvantaged readers. A similar experiment was conducted by (Smith and Jönsson, 2011) for Swedish texts, who find summarized texts to be more readable as well. Also dealing with extractive summarizers, (Finegan-Dollak and Radev, 2016) look at simplifications in the biomedical and legal domain, but their findings indicate that altered sentences lead to fewer correctly answered questions by domain experts. Simplification has also been suggested for multi-document summarization: (Siddharthan et al., 2004) select relevance exclusively over syntactically simplified sentences, whereas other works use simplification as an alternative to regular sentence selection (Vandewende et al., 2007; Yih et al., 2007).

Closest to a unified framework is the work by (Ma and Sun, 2017), who use the same neural encoder-decoder architecture for separate simplification and summarization tasks, which highlights the shared similarities in terms of shared model architectures and training.

To our knowledge, there exist few resources for German single-document summarization. (Nitsche, 2019) mention a (private) resource, provided by the German Press Agency (dpa), which uses headlines as target summaries. (Freid, 2020) generate a corpus based on German Wikipedia articles, and treat the overview paragraph at the beginning as the summary of the article. A similar approach including cross-lingual alignments between English and German has also been recently published (Fatima and Strube, 2021).

3. Text Simplification with Joint Summarization

As previous work has shown, summarization in itself can already be considered a weaker form of simplification (Margarido et al., 2008; Smith and Jönsson, 2011), although existing work never formalizes TS as a
summarization problem. Several points have to be addressed by both simplification and summarization components for a full end-to-end solution. In this section, we outline suggestions for a unified system design.

3.1. Considerations for Simplification

As previously stated, current simplification systems cannot generate significantly shorter output texts when simplifying individual sentences. This is mainly due to the sentence-aligned training setup instead of training with the entire input document. Further, this drops a sizable portion of the source text from training, since sentences are only considered when they align directly with a simplified part. Several resources also lack a document alignment altogether, which completely precludes them from being used as a training resource for end-to-end systems.

Importantly, relevance of individual segments (sentences or paragraphs) has to be computed without knowledge about the output corpus. This can, for example, be achieved by pre-training strategies on monolingual corpora (Mallinson et al., 2020), but could otherwise be learned as an intermediate step in neural architectures. This has been previously shown to work well for multi-document summarization (Liu and Lapata, 2019), where paragraph relevance was learned across several documents.

Further, existing manually annotated corpora are frequently generating simplifications of short texts by "translating" sentence-by-sentence. This reinforces the bias towards equally long documents, which cannot be observed in post-aligned resources (i.e., where existing simplified texts were written independently on the same topic, cf. Table 1). An amended assumption is that simplifications may only be up to a certain length, due to varying attention spans of the target groups. This then requires additional "simplification" based on the length of the source document. This could also be used as a parameter to model levels of difficulty, which is available for some resources, see the Newsela corpus (Xu et al., 2015).

Lastly, existing evaluation metrics strictly focus on sentence-level references (Xu et al., 2016). Extending system evaluations to document-level simplifications poses challenges that need to be overcome in order to collect both manual and automated feedback on the simplification quality.

3.2. Considerations for Summarization

For summarization, TS offers additional challenges not considered by current works. Given a high enough compression rate, simplification can be seen as a special case of summarization. However, existing metrics, such as ROUGE (Lin, 2004), rely on the re-appearance of n-grams in the target summary (in our case, the simplification). This is not guaranteed, given that the simplification can appear in the form of lexicographic replacements. It is thus unclear whether simplification should be considered a separate criterion or jointly modeled for the evaluation of summaries, specifically when considering other input factors as well (ter Hoeve et al., 2020). Additionally, the varying vocabulary and sentence structure pose a challenge to summarization systems, especially extractive approaches. See Section 4.3 for experiments on our Klexikon corpus.

Previous work in that direction has mostly dealt with sentence-level lexicographic simplifications (Sididanthan et al., 2004), yet there are several other simplification operations to be considered (Amancio and Specia, 2014) in a joint end-to-end system.

4. Klexikon Dataset

We introduce a new dataset, loosely inspired in its construction by English Simple Wikipedia, to facilitate future research in joint simplification and summarization. Specifically, we use the German children’s encyclopedia "Klexikon" to obtain simplifications, and align them with reference articles from the German Wikipedia. Compared to Simple Wikipedia, which can be freely edited, Klexikon specifically targets children between roughly the age of 8-13 as readers, and follows a strict reviewing procedure for individual articles, resulting in higher quality texts. We only consider Wikipedia articles with a minimum length of 15 paragraphs, which helps to filter out disambiguation pages or stubs'. Additionally, this results in a clear contrast
in overall article length between source and simplified texts (cf. Table 1 and Figure 1). The final dataset consists of 2,898 article pairs, with Wikipedia documents having on average 8.94 times more sentences compared to their Klexikon counterparts.

4.1. Corpus Creation
All manual steps during corpus creation were performed by the first author of this work. We begin the extraction based on the list of all available articles from the Klexikon overview page in April 2021. At the time of experimentation, this returned 3,150 Klexikon articles, although more articles have been added since.

4.1.1. Document Alignment Strategy
For the identification of matching articles between German Wikipedia and Klexikon, the following steps were performed:

1. Querying the MediaWiki Search API with the title of the Klexikon article. 2,861 articles, or around 90%, have an entry with a directly matching heading on Wikipedia. However, this may include disambiguation pages or stubs.
2. All remaining 289 unmatched articles are manually matched against the top five suggestions by the Wikimedia Search API. If no candidate article is appropriate, the entry is dropped from the corpus.
3. Wikipedia articles with less than 15 paragraphs (108 articles) are again flagged and manually reviewed. Short Wikipedia entries may correspond to disambiguation pages (see next step), or are otherwise dropped because of their short length.
4. Disambiguation pages are replaced with a specific Wikipedia page, if it topically matches at least 66% of the Klexikon paragraphs.

4.1.2. Text Extraction
The Klexikon website runs on the Wiki software, which makes text extraction across platforms very similar. For both websites, we extract all direct children elements of the main content block (div-class: mw-parser-output). Of those, we only use text within <p> tags as the main paragraph content, and heading elements <h1>-<h5>. This simultaneously discards non-textual contents, e.g., images, as well as malformed text elements, such as image captions or lists. We note that the removal of lists can also remove valid content, but frequently suffers from inconsistent grammatical correctness; while some bullet lists...
ment strategies exist (Paetzold et al., 2017; Jiang et al., 2020), but lack compatibility with German texts. We instead experimented with alignments based on sentence embeddings from sentence-transformers (Reimers and Gurevych, 2019) and selecting the most similar source sentence (or pair of sentences) for each Klexikon sentence. However, sentence splitting and merging are impossible to model with this naive alignment strategy, but were frequently found to be the issue of sub-par alignments in a manual review of preliminary results. In particular, we also note that there were both cases of several relevant Wikipedia sentences for a single Klexikon sentence (highlighting the importance of a notion of "relevance"), as well as instances of long sentences from Wikipedia splitting into several (non-consecutive) sentences in the Klexikon text.

4.2. Comparison to Existing Resources

The only other two German datasets with document alignments are the recent resource by Battisti et al. (2020), as well as a smaller version of Klexikon data by Battisti et al. (2020) compiled documents from accessibility options on websites. Compared to our dataset, they potentially cover a more heterogeneous set of topics, but only provide alignments for a subset of articles. As mentioned before, Battisti et al. (2020) provide additional alignments to MiniKlexikon, and otherwise limit the maximum length of articles, which reduces the number of available alignments between all three resources to 295 documents. Even when considering only the equivalent Klexikon-Wikipedia alignments, there are less than 1,000 documents, with additional constraints to the completeness of the Wikipedia texts.

Concerns raised about the quality of Wikipedia as a resource (Xu et al., 2015) mention the problems with sentence alignment, inadequate simplifications, and poor generalization. Our version of the Klexikon dataset partially alleviates these issues:

1. We provide document and (automated) sentence alignments, which allows focusing on both summarization and simplification in a joint manner.
2. Articles for Klexikon are written following stricter guidelines both in their content structure, and we include stricter pre-processing criteria for the Wikipedia articles, resulting in a high-quality collection of text documents.
3. We provide sufficient training samples for potential neural approaches, by increasing the Klexikon-based resource to almost 2,900 articles.

4.3. Baseline Performance

To quantify the quality of our automatically generated alignments, we investigate the dataset from both a summarization and simplification perspective.

4.3.1. Summarization

To verify the suitability of our corpus for summarization purposes, we computed several baselines and compared them to the Klexikon articles as a presumable gold standard summary:

1. **Lead-3**: A baseline frequently used in news article summarization, which consists of the first three sentences. In our case, this corresponds to the first three sentences of the Wikipedia article.
2. **Lead-k**: A related baseline, taking all sentences of the overview section in the Wikipedia article.
3. **Full article**: The full Wikipedia article as a reference for the maximum possible vocabulary overlap (this corresponds to ROUGE-1 recall).
4. **ROUGE-2 oracle**: As an approximation of the upper limit for extractive summaries on this dataset, we select the sentence maximizing ROUGE-2 F1 scores for each sentence in the Klexikon article and
5. **Luhn**: A simple unsupervised baseline for extractive summaries can be generated by Luhn’s algorithm (Luhn, 1958). We use a target of 25 extracted sentences for each generated summary, which corresponds roughly to the median number of sentences in the Klexikon articles.
6. **LexRank S-T**: As a more sophisticated baseline, this approach applies LexRank (Erkan and Radev, 2004) with embeddings extracted by sentence-transformers (Reimers and Gurevych, 2019). The length is similarly limited to at most 25 extracted sentences.

We use ROUGE (Lin, 2004) to gauge summarization quality, which evaluates n-gram overlap between system outputs and gold references. In particular, we report F1 scores for ROUGE-1, ROUGE-2 and ROUGE-L. Results in Table 3 indicate that our dataset poses a significantly harder challenge compared to performance of baselines on standard summarization corpora, such as CNN/DailyMail (Nallapati et al., 2017), where simple lead-3 baselines obtain extremely high ROUGE scores due to an overly pronounced lead bias. On our dataset, lead-3 likely struggles with the very different output lengths and comparatively low recall scores; the opposite is true for the full article baseline, which does not summarize at all, and therefore scores poorly in terms of precision. However, the full article baseline obtains a recall score of 77.3% ROUGE-1, implying there is still a sizable vocabulary overlap between the Klexikon and Wikipedia articles. With proper summarization methods, it is therefore possible to produce decent ROUGE scores, and another indicator of the corpus’ suitability to summarization. Best-suited as a baseline is lead-k, which is a decent approximation of the actual target article length. Even so, lead-k is shorter than the corresponding Klexikon articles. Based on these results, coupled with varying
compression levels between articles (cf. Figure 1), a high sensitivity to the overall input length seems to be required in order to generate appropriate summaries. From the extractive summaries generated by unsupervised methods, it becomes obvious that content from sections outside the overview paragraph is beneficial in terms of ROUGE scores, which is a promising distinction from other summarization datasets, especially in German. Finally, the ROUGE-2 oracle gives insights into the limitations of extractive summarization methods on this dataset. In particular, the differing expressiveness and vocabulary impacts the achievable ROUGE-2 and ROUGE-L scores. It should be noted, however, that the determination of output lengths seems to play a crucial role in the overall balance between precision and recall scores. Given that both unsupervised baselines work with informed choices of the expected summary length, their results should also be taken within the correct context.

4.3.2. Simplification

We further provide different metrics to estimate the level of simplification present in the available documents. For this, we compute Flesch reading-ease scores (Flesch, 1948), specifically an adjusted variation for German (Amstad, 1978). In addition, we hypothesize that the average sentence length (in tokens), as well as the average number of characters per words are suitable proxies for simplification. The latter is especially important for German, which is famous for its long compound words. In particular, we limit the word length calculation to “content word classes”, i.e., nouns, verbs, adjectives, and adverbs only.

To cover lexicographic peculiarities in the data, we estimate the underlying vocabulary. Notably, all texts are quite different in lengths, so an absolute count of distinct tokens would heavily bias the results on Wikipedia. Instead, we approximate this problem by looking at corpus-specific lemma coverage. By computing a corpus-specific list of the 1000 most frequently occurring lemmas, we are then able to compute what fraction of all used lemmas is contained in this top-1000 list. A higher percentage likely points to fewer rare words used, and greater reliance on commonly understood words or an overall smaller vocabulary.

Indeed, we find a consistent pattern in our data (cf. Table 4), where Klexikon data indicates simpler language on all our metrics, which confirms the suitability of our dataset for simplification tasks. We would like to point out the general consensus of the field that heuristics are only scratching the surface of representative readability judgments (Chall, 1958), but still offer a chance for initial exploratory analysis of data suitability.

5. Conclusion and Future Work

In this work, we laid out basic requirements for a unified Text Simplification and Summarization framework. Specifically, we also provided a document-aligned resource of German texts to facilitate future research in this area, and provide quantitative evidence of the suitability of our dataset. We see the following points as the most critical issues for successful joint models: i) Learned sentence relevance and simplification in a joint setting. This can be potentially achieved by modeling sentence alignments similar to existing methods (Stajner et al., 2018; Jiang et al., 2020), but already during the training of an end-to-end system, instead of a separate pre-processing step. ii) Implementation of automated evaluation metrics that align both with human judgments of appropriateness for the summary, as well as simplification steps taken. ROUGE, based on n-grams, potentially suffers similar shortcomings to BLEU as an evaluation metric, since it fails to capture lexicographic simplifications. Existing simplification metrics, however, are unable to quantize the quality based on much longer source documents. iii) Extension of current abstractive summarization systems towards lexicographic simplification, potentially in the form of regularization during training.

6. Bibliographical References

Amancio, M. and Specia, L. (2014). An analysis of crowdsourced text simplifications. In Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), pages 123–130, Gothenburg, Sweden, April. Association for Computational Linguistics.

Amstad, T. (1978). Wie verständlich sind unsere Zeitungen? Studenten-Schreib-Service.

Battisti, A., Pfütze, D., Säuberli, A., Kostrzewa, M., and Ebling, S. (2020). A corpus for automatic read-

Table 3: Average ROUGE F1 for simple extractive baselines. 95% confidence intervals for all scores differ by less than one point.

| Method             | R-1  | R-2  | R-L  |
|--------------------|------|------|------|
| Lead-3             | 16.95| 3.77 | 9.81 |
| Lead-k             | 24.87| 5.10 | 12.01|
| Full article       | 16.81| 4.23 | 6.95 |
| ROUGE-2 oracle     | 41.85| 10.68| 16.00|
| Luhn               | 31.86| 5.55 | 11.57|
| LexRank S-T        | 33.90| 6.11 | 12.86|

Table 4: Indicators of simplified target texts: averages for Flesch complexity scores (between 0 to 100; higher scores indicate simpler texts); average sentence length in tokens; average word length in characters (nouns, verbs, adjectives, adverbs); percentage share of occurrences of the top-1000 corpus-specific lemmas.

| Method             | Wikipedia | Klexikon |
|--------------------|-----------|----------|
| Avg. Flesch score  | 40.1 ± 7.3| 66.7 ± 6.0|
| Avg. sentence length| 22.7 ± 2.6| 13.5 ± 1.5|
| Avg. word length   | 8.7 ± 4.0 | 6.9 ± 3.0 |
| Share of top 1000 lemmas | 68.8% | 82.3% |
ability assessment and text simplification of German.
In Proceedings of the 12th Language Resources and Evaluation Conference, pages 3302–3311, Marseille, France, May. European Language Resources Association.

Chall, J. S. (1958). Readability: An appraisal of research and application. Number 34. Ohio State University.

Coster, W. and Kauchak, D. (2011). Simple English Wikipedia: A new text simplification task. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 665–669, Portland, Oregon, USA, June. Association for Computational Linguistics.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22:457–479.

Fatima, M. and Strube, M. (2021). A novel Wikipedia based dataset for monolingual and cross-lingual summarization. In Proceedings of the Third Workshop on New Frontiers in Summarization, pages 39–50, Online and in Dominican Republic, November. Association for Computational Linguistics.

Finegan-Dollak, C. and Radev, D. R. (2016). Sentence simplification, compression, and disaggregation for summarization of sophisticated documents. Journal of the Association for Information Science and Technology, 67(10):2437–2453.

Flesch, R. (1948). A new readability yardstick. Journal of applied psychology, 32(3):221.

Frefel, D. (2020). Summarization corpora of Wikipedia articles. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 6651–6655, Marseille, France, May. European Language Resources Association.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. In Proceedings of COLING 2012, pages 1063–1080, Mumbai, India, December. The COLING 2012 Organizing Committee.

Hewett, F. and Stede, M. (2021). Automatically evaluating the conceptual complexity of German texts. In Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021), pages 228–234, Düsseldorf, Germany, 6–9 September. KONVENS 2021 Organizers.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning sentences from standard Wikipedia to Simple Wikipedia. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 211–217, Denver, Colorado, May–June. Association for Computational Linguistics.

Jiang, C., Maddela, M., Lan, W., Zhong, Y., and Xu, W. (2020). Neural CRF model for sentence alignment in text simplification. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7943–7960, Online, July. Association for Computational Linguistics.

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1537–1546, Sofia, Bulgaria, August. Association for Computational Linguistics.

Klaper, D., Ebling, S., and Volk, M. (2013). Building a German/simple German parallel corpus for automatic text simplification. In Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations, pages 11–19, Sofia, Bulgaria, August. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, Y. and Lapata, M. (2019). Hierarchical transformers for multi-document summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5070–5081, Florence, Italy, July. Association for Computational Linguistics.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2):159–165.

Ma, S. and Sun, X. (2017). A semantic relevance based neural network for text summarization and text simplification. arXiv preprint arXiv:1710.02318.

Mallinson, J., Sennrich, R., and Lapata, M. (2020). Zero-shot crosslingual sentence simplification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5109–5126, Online, November. Association for Computational Linguistics.

Margarido, P., Pardo, T., Antonio, G., Fuentes, V., Aires, R., Aluisio, S., and Fortes, R. (2008). Automatic summarization for text simplification: Evaluating text understanding by poor readers. In Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, pages 310–315.

Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31.

Nitsche, M. (2019). Towards German Abstractive Text Summarization using Deep Learning. Master’s thesis, HAW Hamburg.

Paetzold, G., Alva-Manchego, F., and Specia, L. (2017). MASSAlign: Alignment and annotation of comparable documents. In Proceedings of the IJCNLP 2017, System Demonstrations, pages 1–4.
A. Experimental Resources and Parameters

For the evaluation of ROUGE scores, we used the Python implementation provided by Google Research.\footnote{https://github.com/google-research/rouge} We replace the original stemmer with Cistem (\textcite{weissweiler2017}) to account for appropriate treatment of German tokens. Flesch complexity scores were computed with the \texttt{textstat} library\footnote{https://github.com/shivam5992/textstat} using the function for German. Sentence length in tokens was derived from the tokenization mentioned in the main article.

B. Data Split

We additionally present a stratified data split for the corpus, with an approximate 80/10/10 split for training, validation and testing. For stratification, we represent each pair of source/simplification documents by their respective lengths in number of sentences. We then divide the coordinate system into a rectangular grid (steps of 100 for Wikipedia article length, step size 10 for Klexikon), and proceed to sample from each grid block according to our pre-defined split (10% of grid samples are selected for validation, 10% for testing, and 80% for training).
the remaining 80% for training). When fewer than ten samples are within a block, all samples are added to the training set. This results in a final split of 2350 training pairs, and 274 samples each for validation and testing. The split is available through the previously mentioned Github repository.