# Text Simplification for Legal Domain: Insights and Challenges

**Aparna Garimella**[1*], **Abhilasha Sancheti**[1,2*], **Vinay Aggarwal**[3†],
**Ananya Ganesh**[4†], **Niyati Chhaya**[1], **Nandakishore Kambhatla**[1]
[1]Adobe Research [2]University of Maryland [3]Google [4]University of Colorado Boulder
{garimell,sancheti,nchhaya,nandakam}@adobe.com,
sancheti@umd.edu, vinayagg@google.com, ananya.ganesh@colorado.edu

## Abstract

Legal documents such as contracts contain complex and domain-specific jargons, long and nested sentences, and often present with several details that may be difficult to understand for laypeople without domain expertise. In this paper, we explore the problem of text simplification (TS) in legal domain. The main challenge to this is the lack of availability of complex-simple parallel datasets for the legal domain. We investigate some of the existing datasets, methods, and metrics in the TS literature for simplifying legal texts, and perform human evaluation to analyze the gaps.[1] We present some of the challenges involved, and outline a few open questions that need to be addressed for future research in this direction.

## 1 Introduction

Contracts are legal documents used in several business workflows. They consist of paragraphs of text (*clauses*) outlining the terms and conditions for the involved parties. Prior to signing a contract, the parties need to understand the clauses, to ensure that they are aware of what they are agreeing to.

Contract clauses are usually very long, domain-specific, and contain several complex phrases (Table 1). Table 2 shows a linguistic comparison of legal language from SEC[2] contract clauses (Tuggener et al., 2020) and simple English Wikipedia (Coster and Kauchak, 2011); the average number of tokens in legal clauses is 129.73, while that in Simple Wikipedia (Coster and Kauchak, 2011) is 18.16, and similarly, the average sentence length of the former is 3.5 times that of the latter. Readability metrics such as Flesch Kincaid (FK) (Kincaid et al., 1975) and Automatic Readability Index (ARI) (Senter and Smith, 1967), and the tree depth of the syntactic parse trees of legal sentences, also indicate that legal language is much more complex compared to simple language in Wikipedia, and may be particularly difficult to read for laypeople without much legal background (our target readers). Further, obtaining legal aid to help interpret and review such language may be expensive for such readers. We believe natural language processing (NLP) techniques for text simplification (TS) can be of particular utility to aid legal document understanding for laypeople without much legal knowledge, who are the target readers for this work.

| In the event that the Landlord shall deem it necessary or be required by any governmental authority to alter, repair, remove, reconstruct or improve any part of the demised premises or of the building in which the demised premises are located (unless the same result from Tenant's act, neglect, default or mode of operation in which event Tenant shall make all such repairs, alterations and improvements), then the same shall be made by the Landlord with reasonable dispatch, however, such obligation of Tenant shall not extend to maintenance, repairs or replacements necessitated by the intentional wrongdoing or gross negligence of Landlord. |
|---|

Table 1: Legal sentence from an SEC contract legal clause.

| DATA | # TOKENS | SENT. LEN. | FK | ARI | PARSE DEPTH |
|---|---|---|---|---|---|
| **Clauses** | 129.73 | 62.52 | 29.89 | 35.05 | 10.79 |
| **SimpleWiki** | 18.16 | 17.98 | 11.72 | 13.11 | 5.72 |

Table 2: Legal language vs. simple English Wikipedia.

The objective of TS is to provide simpler translations for complex input texts. TS is performed at different levels. Lexical simplification aims to replace complex words in a given text with simpler alternatives with equivalent meaning (Gooding and Kochmar, 2019; Qiang et al., 2020). Syntactic simplification typically involves splitting long sentences in shorter ones (Niklaus et al., 2019a,b). (Zhu et al., 2010; Wubben et al., 2012; Martin et al., 2020) use seq2seq-based supervised methods for TS, owing to the availability of parallel datasets (Xu et al., 2015; Niklaus et al., 2019b). There have also been recent advancements in unsupervised TS without the need for parallel datasets (Surya et al., 2019; Laban et al., 2021). However, we believe they may not be readily suited to legal TS, due to the extremely complex nature of legal text as op-

*Equal contribution.
†Work done while at Adobe Research.
[1]The model outputs and human ratings are available at https://bit.ly/3U3ddIl.
[2]Securities and Exchange Commission contracts.



*Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296 - 304

posed to the complex text seen in general news or Wikipedia-like datasets. While prior works on challenges in TS (Xu et al., 2015; Štajner, 2021) focus on the quality of the TS datasets and evaluation metrics, we focus on the generalizability of existing TS systems to legal domain, and challenges in using existing evaluation metrics for legal TS.

In this paper, we aim to address two main research questions. **(1)** How do existing simplification methods perform (in the absence of legal parallel datasets) on the task of legal TS? We specifically examine three types of simplification, namely lexical TS, sentence splitting, and end-to-end TS (split-and-rephrase). **(2)** What are the challenges, if any, in using existing automatic evaluation metrics for legal TS? To this end, we investigate three state-of-the-art (SoTA) unsupervised TS methods in the legal domain (§2.1): *(a)* a BERT-based method for lexical simplification (Qiang et al., 2020), *(b)* a rule-based discourse-aware sentence splitting framework (Niklaus et al., 2019a), and *(c)* a reward-based simplification method that learns to balance fluency, salience, and simplicity of output translations (Laban et al., 2021). We also investigate sequence-to-sequence-based supervised methods (Lewis et al., 2020) trained on three recently released parallel datasets for TS (§2.2). To address the second question, we use several reference-free automatic metrics in the TS literature for simplicity, meaning preservation, and fluency on the model outputs, and conduct human studies to analyze their effectiveness. Finally, we outline some of the challenges in adapting existing methods and metrics to the legal domain, and present a few preliminary research questions that need to be addressed for furthering the research in the space of legal TS.

## 2 Text Simplification for Legal Domain

We use several unsupervised and supervised methods. We briefly describe them below (please refer to Appendix B for further details).

### 2.1 Unsupervised Text Simplification

**Lexical simplification (LS)** aims to replace complex words in a given sentence with simpler words with equivalent meaning to make the resulting text more readable. We use a recent SoTA unsupervised LS method BERT-LS[3] (Qiang et al., 2020) that uses the pre-trained Transformer language model BERT (Devlin et al., 2019) to find simplification candidates for given complex words. Given a complex word $w$ in a sequence $S$, a new sequence $S'$ is constructed with $w$ masked. The original and new sequences are concatenated and fed into BERT to obtain the probability distribution of the vocabulary $p(\cdot|S, S'\backslash\{w\})$ corresponding to the masked word. The top 10 words from $p(\cdot|S, S'\backslash\{w\})$ are selected as simplification candidates, excluding any morphological derivations. The candidates are ranked based on features such as BERT prediction probability, semantic similarity with complex word, and the candidate with the highest average rank is selected as the replacement. We associate complexity of a word with its commonness in a large corpus (Biran et al., 2011; Glavaš and Štajner, 2015), and identify complex words based on their frequency ($<$10K) in normal Wikipedia (Coster and Kauchak, 2011). Further details are provided in Appendix B.

**Sentence splitting** involves the segmentation of a sentence into two or more shorter sentences that can be better processed by NLP systems. We use DISSIM, a discourse-aware syntactic TS framework, that breaks down a complex source sentence into a set of minimal propositions (Niklaus et al., 2019a).[4] Specifically, given a source sentence, it applies recursive transformations based on a set of 35 hand-crafted grammar rules based on syntactic and lexical patterns to split and rephrase the input sentence into structurally simplified sentences, and establish a semantic hierarchy among them.

**Sentence simplification.** We use a recent SoTA reward-based text simplification method KEEPIT-SIMPLE (KIS) (Laban et al., 2021) that uses a generative model GPT-2 (Radford et al., 2019) to transform a complex sentence into a simpler version, while balancing rewards for fluency, salience, and simplicity using reference-free scorers in a reinforcement learning setup.[5] For fluency, perplexity is used from GPT-2; for simplicity, the Fleish-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) and word frequency in a large corpus are used; and for saliency, a coverage model that uses the generated text to answer fill-in-the-blank questions about the input is used. (Laban et al., 2020). While this work can handle paragraphs as unit of text, we use it for sentence simplification, as legal sentences are much longer than typical sentences. Please refer to (Laban et al., 2021) for further details.

[3] https://github.com/qiang2100/BERT-LS

[4] https://github.com/Lambda-3/DiscourseSimplification

[5] https://github.com/tingofurro/keep_it_simple

| Model | Readability | | | Simplicity | Meaning Preservation | | | Hallucination | | Fluency |
|---|---|---|---|---|---|---|---|---|---|---|
| | FK ↓ | Smog ↓ | ARI ↓ | Depth ↓ | BS ↑ | Cov ↑ | Blanc ↑ | Entail ↑ | % Unseen ↓ | Ppl ↓ |
| Legal sent | 41.59 | 29.19 | 50.14 | 13.12 | **1.00** | **0.94** | **0.59** | **99.46** | - | 41.92 |
| BERT-LS | 41.09 | 28.38 | 49.46 | 12.85 | 0.98 | 0.78 | 0.51 | 96.76 | **0.11** | 43.29 |
| DISSIM | 18.69 | 18.55 | 20.83 | 6.42 | 0.92 | 0.89 | 0.54 | 94.59 | 0.25 | 688.12 |
| MWS | **14.71** | 17.52 | **15.96** | **5.92** | 0.93 | 0.84 | 0.50 | 96.76 | 0.19 | 601.04 |
| KIS | 14.98 | **15.87** | 19.35 | 7.83 | 0.85 | 0.45 | 0.14 | 10.27 | 0.23 | **32.91** |
| SBM-WIKI | 20.42 | 20.81 | 23.01 | 8.67 | 0.97 | 0.89 | 0.53 | 91.89 | 0.14 | 69.19 |
| SBM-CONT | 19.88 | 20.48 | 22.37 | 8.70 | 0.97 | 0.89 | 0.54 | 91.35 | 0.13 | 73.26 |
| CORREL | 0.16/0.29 | 0.08/0.36 | 0.20/0.29 | 0.10/0.22 | 0.76/0.75 | 0.05/0.08 | 0.27/0.24 | -0.08/0.72 | -/- | 0.34/0.00 |

Table 3: Results from automatic metrics with **best** and worst values in each column. Correlation between automatic metrics and human ratings are reported for each annotator (A1/A2) in the last row. Correlation for hallucination (fluency) aspect is computed with 1-Entail (1/ppl) and inverse of simplicity with readability and depth measures.

### 2.2 Supervised Text Simplification

We use BART, a denoising autoencoder for pre-training sequence-to-sequence models, for supervised TS (Lewis et al., 2020). It pre-trains a model combining bidirectional and auto-regressive Transformers, with pre-training tasks to corrupt text with noising functions and learning to reconstruct the original text. We fine-tune BART on three complex-simple datasets, one for sentence splitting, and two for split-and-rephrase task.[6]

**MINIWIKISPLIT (MWS)** is a sentence splitting corpus consisting of 203K complex-simple sentence pairs from Wikipedia edit histories (Niklaus et al., 2019b). It was created by running DISSIM (Niklaus et al., 2019a) over the complex input sentences from WIKISPLIT corpus (Botha et al., 2018) and filtering for grammatically incorrect sentences based on a set of dependency parse and part of speech tags.

For the task of split-and-rephrase, Zhang et al. (2020) proposed two benchmark datasets consisting of 500 complex-simple sentence pairs with significantly more diverse syntax in the Wikipedia and legal contracts domain. The data was collected by asking Amazon Mechanical Turk workers to split and rephrase the given complex sentences. We refer to them as **SMALL-BUT-MIGHTY (SBM)**.

## 3 Experiments

We train the KIS model on 67K legal text sentences selected randomly from LEDGAR dataset that do not occur in the test data (further implementation details in Appendix B). For evaluation, we use the LEDGAR dataset (Tuggener et al., 2020) consisting of Securities and Exchange Commission (SEC) contracts. We use 5K sentences randomly sampled from 100 most frequently occurring legal clauses in LEDGAR. Details on the types of clauses and sentence statistics are provided in Appendix A.

**Metrics.** We evaluate the legal sentences and model outputs on meaning preservation, syntactic simplicity, fluency, hallucination, and readability measures. For readability, we use Flesch Kincaid (**FK**) (Kincaid et al., 1975), **SMOG** (Mc Laughlin, 1969), and Automatic Readability Index (**ARI**) (Senter and Smith, 1967) to estimate the minimum age required to understand the given text. We compute syntactic simplicity as the average depth of dependency parse trees of the sentences. For meaning preservation, we use BertScore (**BS**) (Zhang et al., 2019) which is a similarity score for each token in the input sentence with each token in the simplified sentence, Coverage (**Cov**) (Laban et al., 2020) which is the accuracy of filling-in the masked tokens in the masked input sentence using the simplified sentence, and **BLANC** (Vasilyev et al., 2020). We measure hallucination as: (1) % of outputs entailed by the input (**Entail**) computed using SoTA RoBERTa-based (Liu et al., 2019) textual entailment model trained on MNLI (Williams et al., 2018), and (2) % of entities (found using spaCy library) in the output not present in the input (**%Unseen**) (Nan et al., 2021). We compute Fluency (**Ppl**) using perplexity score from GPT-2.

## 4 Results

Results are shown in Table 3. BERT-LS only replaces complex words (if they exist) in the legal sentences; thus it does not achieve much in terms of readability and syntactic simplicity. Since it retains rest other information, it has good meaning preservation and hallucination scores. However, the coverage scores drop, possibly due to

[6]Most of the existing TS datasets (Narayan et al., 2017; Botha et al., 2018; Niklaus et al., 2019b; Zhang et al., 2020; Kim et al., 2021) are for the task of split-and-rephrase; thus we study the splitting and split-and-rephrase tasks.

| Model | Translation |
|---|---|
| Legal sentences | In the event that the Landlord shall deem it necessary or be required by any governmental authority to alter, repair, remove, reconstruct or improve any part of the demised premises or of the building in which the demised premises are located (unless the same result from Tenant's act, neglect, default or mode of operation in which event Tenant shall make all such repairs, alterations and improvements), then the same shall be made by the Landlord with reasonable dispatch, however, such obligation of Tenant shall not extend to maintenance, repairs or replacements necessitated by the intentional wrongdoing or gross negligence of Landlord. |
| BERT-LS | In the event that the **owner** shall **find** it necessary or be required by any governmental authority to alter, repair, remove, rebuild or improve any part of the deceased premises or of the building in which the **deceased** premises are located (unless the same result from owner's act, neglect, default or mode of operation in which event owner shall make all such repairs, alterations and improvements), then the same shall be made by the owner with reasonable care, however, such obligation of owner shall not extend to maintenance, repairs or replacements caused by the intentional mistake or gross indifference of owner. |
| DISSIM | This is however. This is in the event. The Landlord shall deem it necessary. The Landlord shall be required by any governmental authority to alter, repair, remove, reconstruct or improve any part of the demised premises or of the building in which the demised premises are located (unless the same result from Tenant's act, neglect, default or mode of operation in which event Tenant shall make all such repairs, alterations and improvements). This was with reasonable dispatch. The same shall be made by the Landlord. Such obligation of Tenant shall not extend to maintenance, repairs or replacements. Maintenance, repairs or replacements were necessitated by the intentional wrongdoing or gross negligence of Landlord. |
| MWS | Such obligation of Tenant shall not extend to maintenance, repairs or replacements necessitated by the intentional wrongdoing or gross negligence of Landlord. In the event that the Landlord shall deem it necessary or be required by any governmental authority to alter, repair, remove, reconstruct or improve any part of the demised premises or of the building. This was however. The same result from Tenant's act shall make all such repairs, alterations and improvements. Tenant's act is neglect, default or mode of operation in which event Tenant. |
| KIS | The Landlord shall deem it necessary or be required by any governmental authority, to alter, repair, or improve, any part of the demised premises, in which he or she is to retain as a condition of his or her employment. This notice will be sent to Tenant via e-mail, to inform her of the changes that are to be made to the structure of the service. When the service is offered, the holder will advise him or her of his or her choice. |
| SBM-WIKI | In the event that the Landlord shall deem it necessary or be required by any governmental authority to alter, repair, remove, reconstruct or improve any part of the demised premises or of the building in which the Demised premises are located (unless the same result from Tenant's act, neglect, default or mode of operation in which event Tenant shall make all such repairs, alterations and improvements), then the same shall be made by Landlord with reasonable dispatch. Such obligation of Tenant may not extend to maintenance, repairs or replacements necessitated by the intentional wrongdoing or gross negligence of Landlord. |
| SBM-CONTRACTS | In the event that the Landlord shall deem it necessary or be required by any governmental authority to alter, repair, remove, reconstruct or improve any part of the demised premises or of the building in which the Demised premises are located. Such obligation of Tenant shall not extend to maintenance, repairs or replacements necessitated by the intentional wrongdoing or gross negligence of Landlord. This is unless the same result from Tenant's act, neglect, default or mode of operation. In this case, then the same shall be made by the Landlords with reasonable dispatch. This shall be in the event Tenant makes all such repairs, alterations and improvements. |

Table 4: Example model outputs. Phrases that may be factually inconsistent with the input sentence are highlighted in red Phrases or sentences that are not grammatical or fluent are highlighted in blue.

the lexical replacements being less legal-like, thus making it difficult to reconstruct the original sentence while using this metric. Perplexity increases slightly, perhaps due to some not very meaningful replacements. A few examples illustrating this are provided in Table 8 (Appendix D).

Both the unsupervised and supervised sentence splitting methods (DISSIM, MWS) result in significantly better readability scores and dependency depth, indicating splitting of longer legal sentences. Their meaning preservation and entailment scores are also high. However, they have very high perplexity scores, due to the abrupt sentence breaks.

KIS achieves good readability and fluency scores; however, meaning preservation and entailment scores decrease significantly, also indicated by the generation of a few factually inconsistent phrases in the output (Table 4). This may be due to unsupervised nature of generation and the particularly complex nature of legal text as opposed to general news-like text. It is interesting that BART model trained on both SMB-Wiki (out-of-domain) and SMB-Contracts (in-domain) result in similar meaning preservation and entailment scores, where the out-of-domain effect is not seen. On closer examination, we note that in most cases, they just copy the sentences from input, with occasional sentence splitting or phrase deletions, that sometimes results in not very grammatical sentences and increased perplexity. A few qualitative examples are shown in Table 4.

**Human evaluation.** Due to the domain-specific nature of legal texts, we conduct human studies with two legal experts (A1 and A2) on Upwork. Since legal experts will be able to better comprehend legal text, we choose them for our human evaluation as opposed to laypeople (who form our target group of readers for legal TS). We provide them 150 sentences randomly selected from the test data along with corresponding model outputs, and instruct to rate the legal sentences for simplicity, and model outputs for simplicity, meaning preservation, fluency, and hallucinations on a scale of 1 (very complex, low meaning preservation, least fluent, or less hallucinated) to 5 (simple, high meaning preservation, most fluent, or highly hallucinated).[7] The task description and guideline are provided in Appendix C.

[7]For simplicity, we instruct them to rate the examples as per how they would explain to their clients (laypeople).

| Model | Simp.↑ | MP↑ | Hall.↓ | Flu.↑ |
|---|---|---|---|---|
| Legal sentences | 2.62/2.26 | - | - | - |
| BERT-LS | 2.77/3.14 | 4.94/4.66 | **1.00**/1.21 | **4.75/4.74** |
| DISSIM | 2.24/2.69 | **4.95/4.93** | 1.10/1.58 | 3.52/3.10 |
| MWS | 2.70/2.93 | 4.71/4.45 | 1.49/1.70 | 3.94/3.23 |
| KIS | 2.07/**3.82** | 1.30/1.31 | 4.24/4.68 | 1.16/3.23 |
| SBM-CONT | **2.79**/2.86 | 4.92/4.75 | 1.57/**1.07** | 4.67/4.50 |
| $\alpha$(A1, A2) | -0.06 | 0.90 | 0.70 | 0.41 |

Table 5: Human ratings (A1/A2) with **best** and worst values in each column, with Krippendorff $\alpha$ between the ratings.

Table 5 shows the ratings from the annotators. It is very interesting to note that the inter-rater agreement using Krippendorff's $\alpha$ (Krippendorff, 1970) between their ratings for simplicity is $-0.06$, indicating disagreement between the way they perceive simplicity of legal text. However, they have high agreement for meaning preservation and hallucinations, possibly due to their good understanding of legal text, and a moderate agreement for fluency. From a few simplifications (Table 7 in Appendix C) the annotators provided (as per their selection process), we note that A1 simplifies colloquially , and sometimes chooses to exclude some details that may not concern an average layperson. Whereas, A2's language is less colloquial, with most of the details included, in a simpler language (with considerable paraphrasing, fewer nestings, and fewer legal jargons). We suspect this disagreement may be due to the legal experts' varying notions of simplicity in the manner in which they explain legal contract clauses to their clients;[8] further studies are needed to examine the simplicity of model outputs from laypeople's perspective—simplification datasets need to be curated based on whether the target audience prefers all the details or the most important content, colloquial or more formal simplifications, to develop TS models for legal domain.

Overall, both the annotators rate the KIS model poorly in terms of meaning preservation and hallucination; in terms of simplicity, A2 rates KIS highest, while A1 rates it lowest, possibly due to the amount of hallucinations in the outputs.

**Correlation with automatic metrics.** Table 3 (last row) shows the Pearson correlation coefficients of human ratings with automatic metrics for the 150 legal sentences and their model outputs. Since lower values are better for depth and readability metrics, we compute the correlation of inverse of human ratings with them. Tree depth and readability have weak (A1) to moderate (A2) correlations with annotators' simplicity ratings, indicating that these may not be appropriate metrics to measure simplicity of legal texts (Tanprasert and Kauchak, 2021). While splitting methods such as DISSIM and MWS are rated well for readability and depth using the automatic metrics, the annotators rate them lower for simplicity (Table 5), as these methods do not rephrase complex phrases into simpler ones. For meaning preservation, BertScore has good correlation with both the annotators' ratings; however, coverage and Blanc metrics have weak correlations, indicating that they may not fully capture the meaning preservation in legal texts. For hallucination, entailment score captures to a significant degree any factually inconsistent information (A2), though A1's ratings indicate no correlation. Similarly for fluency, A1's ratings are moderately correlated with the inverse of perplexity, while A2's ratings show no correlation. Further investigation is needed to concretely understand these metrics before using them for this task.

[8]Note that the clients of these legal experts form our target group of readers, and not the legal experts themselves.

## 5 Conclusions

While legal text is complex and domain-specific, thus making it a very interesting domain for TS, it is still in a nascent stage in NLP literature. We investigate and compare some of SoTA methods for lexical simplification, sentence splitting, and seq2seq sentence simplification, either unsupervised, or trained on closely related parallel datasets, using automatic metrics and human ratings. We conclude that lexical simplification methods will benefit from having a legal lexicon as they still sometimes generate replacements that do not fit the legal context. Seq2seq methods perform only surface-level transformations by either directly copying input sentences, or deleting a few phrases to make the sentences shorter, without much paraphrasing. While sentence splitting methods make the long nested sentences much shorter, they do so by sacrificing fluency. Reward-based generation method achieves transformations to an extent, but does so at the cost of meaning preservation. Legal TS can be particularly challenging, as even expert annotators have varied views of how to simplify legal sentences for laypeople. Understanding whether every detail is needed to be conveyed or providing a high-level overview suffices can aid in curating parallel datasets for furthering research in this space.

## 6 Ethical statement

We are committed to ethical practices and protecting the anonymity and privacy of individuals who have contributed. We ensure that the privacy of the annotators is protected. For annotations, $\$15 - 20$/hr was paid per task.

## References


Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA.

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium.

William Coster and David Kauchak. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China.

Sian Gooding and Ekaterina Kochmar. 2019. Recursive context-aware lexical simplification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4853–4863, Hong Kong, China. Association for Computational Linguistics.

Joongwon Kim, Mounica Maddela, Reno Kriz, Wei Xu, and Chris Callison-Burch. 2021. BiSECT: Learning to split and rephrase sentences with bitexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6193–6209, Online and Punta Cana, Dominican Republic.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, Online.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378, Online.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical*

*Methods in Natural Language Processing*, pages 606–616, Copenhagen, Denmark. Association for Computational Linguistics.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019a. DisSim: A discourse-aware syntactic text simplification framework for English and German. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 504–507, Tokyo, Japan. Association for Computational Linguistics.

Christina Niklaus, André Freitas, and Siegfried Handschuh. 2019b. MinWikiSplit: A sentence splitting corpus with minimal propositions. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 118–123, Tokyo, Japan. Association for Computational Linguistics.

Jipeng Qiang, Yun Li, Zhu Yi, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. *Thirty-Fourth AAAI Conference on Artificial Intelligence*, page 8649–8656.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Cincinnati Univ OH.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. Ledgar: a large-scale multi-label corpus for text classification of legal provisions in contracts. In *12th Language Resources and Evaluation Conference (LREC) 2020*, pages 1228–1234. European Language Resources Association.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. The multi-genre nli corpus.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Li Zhang, Huaiyu Zhu, Siddhartha Brahma, and Yunyao Li. 2020. Small but mighty: New benchmarks for split and rephrase. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1198–1205, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

## A Dataset Statistics

We use 5K sentences randomly sampled from 100 most frequently occurring legal clause types from SEC contracts from the LEDGAR dataset (Tuggener et al., 2020) for evaluation. Some of these clause types include *amendments, base salary, benefits, duties, employment, entire agreements, expenses, governing laws, notices, positions, severability, terms, vacations*, *waivers*, and so on.

## B Implementation Details

BERT-LS. BERT-LS requires identification of complex words in a sentence; we identify complex words in a given test sentence based on their frequency ($<$ 10K) in normal Wikipedia (Coster and Kauchak, 2011) which is essentially the unsimplified text from Wikipedia. Its vocabulary is of size 594K tokens. In the test sentences, we consider a token potentially complex (or specific to legal domain) if it is less likely seen in normal Wikipedia

In this project, you are given a few sentences. For each sentence, there are at most 6 translations obtained using automatic AI models or human translations. Your task is to rate the sentence along with the translations on their simplicity. In addition, for each of the translations, you are required to rate the content preserved in them, their fluency, and any hallucinations that may have been introduced in them.

Simplicity: This refers to how simple of plain English-like the given sentence or translation is. When we say simplicity, we are referring to how plain English-like a given translation is looking. For pointers on plain English versions of SEC contracts, this resource gives very nice examples in Chapter 6: https://www.sec.gov/pdf/handbook.pdf.
1: very complex; 5: very simple and easily understandable for laypeople without much legal background.

Content preserved in a translation: This refers to the amount of information from the given sentence that is retained in the translation.
1: Almost every detail is missed; 5: Every detail is covered in the translation.

Fluency of a translation: Fluency refers to how natural and grammatical a sentence/translation is.
**Example of fluent sentence**: In addition, it is impractical to make such a law.
**Example of non-fluent sentence**: It is unfair to release a law only point to the genetic disorder.
1: Not fluent or unnatural or grammatically incorrect translation; 5: Very fluent, natural, and grammatically correct translation.

Hallucination in a translation: The refers to the degree of incorrect or redundant information included in the translation compared to given sentence.
1: No redundant or incorrect information is present in the translation, every detail in it is taken from the given sentence; 5: Lot of redundant or incorrect information present in the translation compared to given sentence.

Table 6: Instructions for human studies.

(frequency $< 10$K)[9]. This results in a total of $2,708$ complex tokens, which include *misconduct*, *acquisitions*, and *obligors*.

DISSIM outputs a graph-like structure of the input. To get a sentence from the graph-like structure, we traverse it from left to right and construct an output using the leaf nodes. If DISSIM fails to generate any graphs, we copy the input as output without any transformations. It uses a set of hand-crafted transformation rules to recursively transform an input sentence into a two-layered hierarchical representation in the form of core sentences and accompanying contexts that are linked via rhetorical relations (such as list, elaboration). For further details on the specific rule, we refer the readers to Niklaus et al. (2019a). We train the KIS model on 67K legal text sentences selected randomly from LEDGAR dataset that do not occur in the test data. We train the KIS model using the same GPT-2 medium checkpoint and other hyperparameters as in (Laban et al., 2021). We use huggingface's transformers library (Wolf et al., 2019) to fine-tune BART models for 3 epochs using Adam optimizer with batch size of 8 and maximum sequence length of 256.

## C Human Evaluations

Table 6 shows the instructions used to guide the Upwork annotators for rating the legal sentences and model outputs for their simplicity, meaning preservation, hallucinations, and fluency. We conducted interviews by first giving a few legal sentences from SEC contracts and instructing them to explain the information conveyed in them in easy-to-understand language. Based on further discussions, we selected two annotators for this task. The two annotators are paid $\$15$ and $\$20$ per hour respectively. Table 7 shows a few simplifications that the annotators provided during the interviews.

## D Qualitative Results

[9]We use 10K as threshold based on manual observation of resulting words. The maximum frequency of any token in Wikipedia is 173M.

| **Legal sentence** | **Annotator-1** | **Annotator-2** |
|---|---|---|
| In the event that the Landlord shall deem it necessary or be required by any governmental authority to alter, repair, remove, reconstruct or improve any part of the demised premises or of the building in which the demised premises are located (unless the same result from Tenant's act, neglect, default or mode of operation in which event Tenant shall make all such repairs, alterations and improvements), then the same shall be made by the Landlord with reasonable dispatch, however, such obligation of Tenant shall not extend to maintenance, repairs or replacements necessitated by the intentional wrongdoing or gross negligence of Landlord. | The landlord will repair, remove, reconstruct or improve the leased property if it is required by any governmental authority. However, the landlord is not entitled to do so if it is the tenant's fault. The Landlord will make the repairs or replacements as soon as possible. Also, it is not the tenant's duty to maintain or repair the property if the damages were caused by the Landlord's negligence. | Where the landlord feels necessary or where it is required by any government authority to repair, remove or reconstruct any part or the building which is used by a Tenant under lease agreement. The landlord will make reasonable efforts to repair or reconstruct such part or building leased. As an exception, where such damage to the leased part or building is the result from the Tenant's act, default or mode of operating the area in such case the Tenant will make all such repairs. This obligation of Tenant will not extend to repairs if such damage is the result of intention carelessness on the part of the landlord. |
| Any termination of Executive's employment by the Company without Cause (and not due to Executive's death or Permanent Disability) shall be made by the provision of at least fourteen (14) days' prior written notice to Executive in accordance with Section 4.2 ; provided , however , that the Company may, in its sole discretion, elect to pay Executive for all or any part of the notice period in lieu of providing prior written notice, calculated based on the annualized rate of Executive's Effective Base Salary at the time of termination. | A written notice of fourteen days must be given by the company to the employee if the employee is terminated without any cause and not due to death. However, the company can pay an employee for the notice period as per their annual base salary. | As per Section 4.2, for terminating the Executive without cause (and not due to Executive's Death or Permanent Disability) the Company will provide a prior written notice of 14 days to the Executive. In this case, the Company at its own discretion can choose to pay to the Executive all or any part of the amount against such notice period. The calculation of such amount will be based on annual base salary of the Executive at the time of termination. |

Table 7: Sample simplifications from legal experts.

| LEGAL SENTENCE | **Lexical Simplification** (BERT-LS) | BERTSCORE | COVERAGE |
|---|---|---|---|
| The Stockholder hereby ratifies and confirms all that such irrevocable proxy may lawfully do or cause to be done by virtue hereof. | The company now **agrees** and agrees all that such **a** proxy may **illegally** do or cause to be done by virtue of. | 0.92 | 0.18 |
| There are no strikes, lockouts or other material labor disputes or grievances against the Borrower or any of its Subsidiaries, or, to the Borrower's knowledge, threatened against or affecting the Borrower or any of its Subsidiaries, and no significant unfair labor practice charges or grievances are pending against the Borrower or any of its Subsidiaries, or, to the Borrower's knowledge, threatened against any of them before any Governmental Authority. | There are no strikes, **strikes** or other material labor disputes or **claims** against the company or any of its **branches**, or, to the **company's** knowledge, threatened against or affecting the **company** or any of its **branches**, and no significant unfair labor practice charges or **claims** are pending against the **company** or any of its **branches**, or, to the **company's** knowledge, threatened against any of them before any Governmental Authority. | 0.95 | 0.38 |

Table 8: Example BERT-LS outputs for lexical simplification to illustrate low coverage cases.