Document-Level Text Simplification: Dataset, Criteria and Baseline

Renliang Sun, Hanqi Jin, Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
Center for Data Science, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
sunrenliangpku@gmail.com
{jinhanqi, wanxiaojun}@pku.edu.cn

Abstract

Text simplification is a valuable technique. However, current research is limited to sentence simplification. In this paper, we define and investigate a new task of document-level text simplification, which aims to simplify a document consisting of multiple sentences. Based on Wikipedia dumps, we first construct a large-scale dataset named D-Wikipedia and perform analysis and human evaluation on it to show that the dataset is reliable. Then, we propose a new automatic evaluation metric called D-SARI that is more suitable for the document-level simplification task. Finally, we select several representative models as baseline models for this task and perform automatic evaluation and human evaluation. We analyze the results and point out the shortcomings of the baseline models.

1 Introduction

Text simplification is a valuable technique that deserves to be studied in depth (Woodsend and Lapata, 2011). One definition of text simplification is to simplify the original text to a more understandable text, while keeping the main meaning of the original text unchanged (Štajner and Saggion, 2018; Maddela et al., 2020). It can provide convenience for non-native speakers (Petersen and Ostendorf, 2007; Glavaš and Štajner, 2015; Paetzold and Specia, 2016), non-expert readers (Elhadad and Sutaria, 2007; Siddharthan and Katsos, 2010) and children (De Belder and Moens, 2010; Kajiwara et al., 2013).

1.1 Why it is Valuable to Study Document-level Text Simplification

Currently, researches on text simplification focus on sentence simplification, and the existing common text simplification datasets such as Wikilarge, Wikismall, and Newsela are also designed for sentence simplification. However, various complex applications in the real world often require document-level simplification rather than sentence-level simplification. Imagining that if you want to simplify an article in Time magazine for children to read, it is very inefficient to simplify the sentences separately. Besides, sentences that are obscure and have little relation to the subject should be deleted instead of simplified. Therefore, studying document-level text simplification may be more meaningful than studying sentence-level text simplification alone. Unfortunately, the research on document-level text simplification is still scarce: there is no formal definition, no suitable dataset, and evaluation criteria.

1.2 Similarities and Differences with Text Summarization

Other tasks that may be related to document-level text simplification are text summarization(Dong et al., 2018; Cao et al., 2020), paraphrasing (Zhao et al., 2018; Guo et al., 2018), and split & rephrase (Narayan et al., 2017; Surya et al., 2019). Obviously, the paraphrasing and split & rephrase tasks are both sentence-level tasks. The most closely related task is text summarization, which is also a document-level task. We use an example to illustrate the difference between text summarization and our task, as shown in Table 1. We can see that text summarization does not involve rewriting text with simplified versions, though both the two tasks may filter or delete some unimportant text from the original document.

1.3 Our Contributions

In this paper, we are committed to promoting research on document-level text simplification. In summary, the main contributions of our work include:

(1) We define the new task of document-level text simplification and build the D-Wikipedia

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7997–8013
November 7–11, 2021.

Original article	Firefighters or firemen are people whose job is to put out fires and rescue people. Besides fires, firefighters rescue people and animals from car wrecks, collapsed buildings, stuck elevators and many other emergencies. Firefighting is a job which requires bravery, strength, quick thinking and a wide range of skills. Firefighters are based at a building called a " fire station " (also known as a " firehouse " or " fire hall "). When their help is needed, they drive a vehicle called a " fire engine " or " fire truck " to the scene responding code 1 code 2 or code 3. These vehicles can pump water and foam to put out fires. Fire engines also carry ladders, cutting tools and lots of different types of rescue equipment. Most carry first aid kits to help people who are injured or hurt.
Document-level simplification	**The job of** a firefighter is to put out fires and **save lives from many emergencies**. They are based at a building called a " fire station ". They drive a vehicle called a " fire engine " or " fire truck " to the scene. The vehicle carries many types of rescue equipment **to help people in danger**.
Text summarization	Firefighters or firemen are people whose job is to put out fires and rescue people and animals from many emergencies. Firefighters are based at a building called a " fire station ". When their help is needed, they drive a vehicle called a " fire engine " or " fire truck " which may carry different types of rescue equipment to help people who are injured or hurt to the scene .

Table 1: Examples for document-level simplification and text summarization. It can be seen from the **Bold** part that the simplified article not only deletes complicated and unimportant sentences from the original article but rewrites the clause, merges two sentences then simplifies them, replaces difficult words, etc. These are operations that text summarization does not require.

dataset for research[1].

(2) We propose a new automatic evaluation metric called D-SARI that is more suitable for the new task.

(3) We select several representative models and perform both automatic evaluation and human evaluation. The results could serve as the baselines.

2 Related Works

Sentence simplification aims to rewrite an original sentence into a more straightforward sentence (Saggion, 2017; Sulem et al., 2018b). The input and output of the model are just sentences instead of articles. Based on the English Wikipedia and the Simple English Wikipedia, many researchers have built high-quality datasets such as Wikilarge (Zhang and Lapata, 2017), Wikismall (Zhu et al., 2010), and so on (Coster and Kauchak, 2011; Kauchak, 2013). Based on Newsela, Xu et al. (2015) established the Newsela dataset. The above datasets are widely used in the field of sentence simplification.

Most of the early simplification models were based on statistical machine translation (Wubben et al., 2012; Narayan and Gardent, 2014). Nisioi et al. (2017) improved the machine translation model to obtain a new simplification model. Zhang and Lapata (2017) developed a reinforcement learning model and achieved excellent results. Vu et al. (2018) introduced a new memory-augmented neural network to enhance the results. Two new approaches were proposed by Kriz et al. (2019) to solve the problem of long and complicated simplified outputs. Scarton and Specia (2018) and Nishihara et al. (2019) investigated how to simplify sentences to different difficulty levels. Dong et al. (2019) studied the three explicit edit operations in sentence simplification and proposed a new model. Štajner et al. (2017), Paetzold et al. (2017), and Jiang et al. (2020) proposed sentence alignment methods to improve sentence simplification. Sun et al. (2020) used the preceding and following sentences to help simplify a specifically given sentence.

There are very few works related to document-level text simplification. Alva-Manchego et al. (2019) focused on cross-sentence transformations in text simplification and analyzed them, concluding that document-level simplification cannot be achieved by merely selecting parts of the content then simplifying individual sentences. Subsequently, Zhong et al. (2020) used discourse-level factors to predict whether a sentence should be deleted and achieved good results.

Although some previous works focus more or less on document-level information, the task of document-level text simplification is still not

[1]The D-Wikipedia dataset is released at https://github.com/RLSNLP/Document-level-text-simplification.

Operation	Sentence joining	Sentence splitting	Sentence deletion	Sentence reordering	Sentence addition	Anaphora resolution
Percentage(%)	96	84	91	92	92	92

Table 2: Estimated percentage of articles with each of the six document-level simplification operations in the D-Wikipedia dataset. Each of these document-level operations appears in most of articles in the dataset.

clearly defined, and there are no available high-quality datasets and criteria to evaluate the generated articles.

3 Problem Formulation

The document-level text simplification task can be defined as follows. Given an original complex article C, the article consists of n sentences, denoted as $C = \{S_1, S_2, ...S_n\}$. Document-level simplification aims to simplify C into m sentences, which form the simplified article F, denoted as $F = \{T_1, T_2, ...T_m\}$, and m may not be equal to n. F retains the primary meaning of C and is more straightforward than C, making it easier for people to understand.

The operations for sentence-level simplification include word reservation and deletion, synonym replacement, etc. (Xu et al., 2016) Based on the work of Alva-Manchego et al. (2019), we define six types of document-level simplification operations, namely, sentence joining, sentence splitting, sentence deletion, sentence reordering, sentence addition, and anaphora resolution. See Appendix A for the specific definition and example of each operation.

In our definition, document-level simplification should allow the loss of information but should not allow the loss of important information. Zhong et al. (2020) pointed out that sentence deletion is a prevalent phenomenon in document simplification. We believe that information that has little relevance to the primary meaning should be removed to improve readability.

4 The D-Wikipedia Dataset

4.1 Dataset Construction

According to the definition of document-level simplification, we built a new large-scale dataset named D-Wikipedia based on the English Wikipedia and Simple English Wikipedia. We first downloaded dumps from the official website of Wikipedia and created over 170,000 article pairs[2]. Considering that it is not easy to establish a one-to-one correspondence between the contents, i.e., the subheadings, we kept only the main content, which is the abstract below the headings. Meanwhile, we considered that if the article is too long, it will occupy a large amount of memory during training. Therefore, we removed those article pairs whose original article or simplified article is longer than 1,000 words. Finally, we built a dataset containing 143,546 article pairs. The D-Wikipedia dataset not only can be used for document-level simplification research but also can be further aligned to construct a sentence-level simplification dataset.

In this work, we randomly divided the dataset into 132K article pairs as the training set, 3K article pairs as the validation set, and 8K article pairs as the test set. There is no overlap between the training set, validation set, and test set.

4.2 Additional Newsela Test Set

There is also a commonly-used and high-quality corpus named Newsela that might be used for document-level simplification. Each original article in the Newsela corpus corresponds to four articles of different simplification levels. We also removed the article pairs whose original article or simplified article is longer than 1000 words. Given that the number of articles in each simplification level is less than a thousand, we only use them to build four additional test sets of different simplification levels. In addition, using the Newsela corpus requires a license[3], while the D-Wikipedia dataset will be completely open-source.

4.3 Statistics and Comparison

We randomly sampled 100 article pairs from the established D-Wikipedia dataset to estimate the percentage of the articles which contain each of the six document-level simplification operations (mentioned in Section 3). We used Amazon Mechanical

[2]The English Wikipedia dumps are from https://dumps.wikimedia.org/enwiki and the Simple English Wikipedia are from https://dumps.wikimedia.org/simplewiki. The version we used is 2020-08-20. We also used the WikiExtractor (https://github.com/attardi/wikiextractor) to extract and clean text from the dumps.

[3]https://newsela.com/data

Turk to invite three workers to identify the operations in the articles, and the percentage of articles with each operation is shown in Table 2. It can be seen that each simplification operation appears in most of the simplified articles in the dataset. In other words, most articles involve with different simplification operations.

We also calculated the percentage of each simplification operation according to the total occurrences of the operations in the simplified articles, which is shown in Figure 1. It can be seen that the sentence deletion operation occurs most frequently in the dataset.

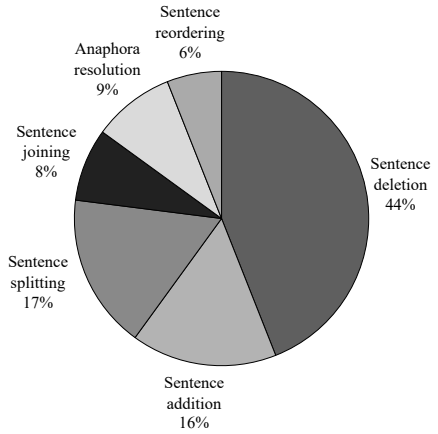

Figure 1: The percentage of each simplification operation. Sentence deletion occurs most frequently, accounting for nearly half of the total simplification operations, while sentence reordering occurs least frequently, accounting for only 6% of the total simplification operations.

To analyze the word-level differences between the original articles and the simplified articles, following Xu et al. (2015), we adopted the odds ratio method proposed by Monroe et al. (2008). The odds ratio of token t between corpus i and corpus j is defined as:

$$r_t^{(i-j)} = \frac{y_t^i / y_t^j}{n^i / n^j} \quad (1)$$

In Equation 1, y_t^i represents the count of token t in corpus i and y_t^j represents the count of token t in corpus j. n^i and n^j represent the size of corpus i and corpus j, respectively. We have found some complex words that occur frequently as examples to show that they are sufficiently simplified, as shown in Table 3.

It can be seen from $R_{original}$ and R_{simple} that the relative frequency of complex words appearing in simplified texts is much lower than in the original texts. We used a chi-square test to show if the odds ratio is significantly different from one[4]. An odds ratio significantly lower than one means that the complex words are well simplified. The reduction of the word "including" may means that clauses are deleted or split into multiple sentences.

	$R_{original}$	R_{simple}	odds ratio↓	p-value
population	47	123	0.49	0
including	68	173	0.49	0
located	79	263	0.38	0
metropolitan	281	904	0.32	0

Table 3: $R_{original}$ and R_{simple} indicate the ranking of the number of occurrences of the word in the original article and the simplified article, respectively. A smaller odds ratio means a greater reduction of the complex word. The closer the p-value of the test is to zero, the more significant the difference between the odds ratio and one.

Sentence splitting is a common operation in document-level simplification. When splitting the conjoined clauses, to preserve the rhetorical relation, Siddharthan (2003) introduced the cue words. We calculated the odds ratio of the conjunction and the cue word, and the results are shown in Table 4. We did not calculate all words because the number of occurrences of some words, such as "hence", was too low to be statistically meaningful.

conjunction	although	though	since	as
odds ratio↓	0.41	0.68	0.74	0.64
p-value	0	0	0	0
conjunction	and	or	but	
odds ratio↓	0.78	0.93	1.02	
p-value	0	0	0.04	

cue word	still	then	also	however
odds ratio↑	1.23	1.18	1.12	0.76
p-value	0	0	0	0

Table 4: The odds ratio of most conjunctions is significantly less than one, and the odds ratio of most cue words is significantly greater than one, indicating that the simplified article may contain more split sentences and the long sentences in the original article have been simplified.

The D-Wikipedia dataset was also analyzed and compared with the Newsela corpus, and the results are shown in Table 5. In terms of the average number of words per article, the compression ratio

[4]We use the script in https://github.com/scipy/scipy/blob/v1.7.1/scipy/stats/contingency.py.

	D-Wikipedia		Newsela corpus			
	Original	Simple	Simp-1	Simp-2	Simp-3	Simp-4
Total articles	143,546		712	775	813	797
Total sentences	707,470	581,513	27,254	34,814	39,489	37,329
Total words	20,349,706	11,286,155	575,077	613,174	570,164	442,173
Avg words per article	141.76	78.62	807.69	791.19	701.31	554.80
-Compression ratio		0.55	1.05	1.01	0.89	0.70
Avg words per sent	28.76	19.41	21.10	17.61	14.44	11.85
-Compression ratio		0.67	0.86	0.72	0.59	0.48

Table 5: Basic statistics of the D-Wikipedia dataset vs. the Newsela Simplification corpus. For the Newsela corpus, the results are different from those reported by Xu et al. (2015) because we deleted too long articles.

of the D-Wikipedia dataset is lower than that of the Newsela corpus of any simplification level. In terms of the average number of words per sentence, the compression ratio of the D-Wikipedia dataset is between the Simp-2 level and the Simp-3 level of the Newsela corpus.

4.4 Human Evaluation

In this section, we employ human judges to evaluate the quality of the D-Wikipedia dataset. Before evaluation, we need to analyze whether the human evaluation indicators used for sentence-level simplification are suitable to evaluate document-level simplification.

In human evaluation, sentence simplification is usually evaluated from the three perspectives of simplicity, meaning, and grammar. Simplicity is the most important indicator. Simplicity indicates if the simplified sentence is simpler than the original sentence. However, we believe that this measure is not a good indicator for scoring document-level simplification. For example, if only the first sentence of the original article is simplified and the other sentences are deleted, the simplified article will be very short and simple and will get a high simplicity score. But such an article does not retain the main information of the original article, which is not what we want.

Therefore, we propose a new indicator named O-simplicity (Overall simplicity with quality guarantee). O-simplicity indicates if the simplified article is simpler than the original article, under the condition of quality guarantee, i.e., it also should read smoothly and can retain the main meaning of the original article. As an indicator to evaluate how good the simplification is, O-simplicity is a more meaningful and comprehensive measure than the original simplicity indicator or simply averaging the simplicity, meaning, and grammar scores.

Following Sulem et al. (2018c), we also use the fine-grained simplicity-phrase and simplicity-structure, which measures the simplification of words and the simplification of sentence structure, respectively. In this way, O-simplicity is an overall indicator that needs comprehensive consideration, and the other four indicators are focusing on specific aspects. More examples and scoring guidelines are given and analyzed in Appendix C.

We invited three workers to evaluate the quality of the D-Wikipedia dataset with the above measures. We randomly selected 100 article pairs from the dataset, and the five-point Likert scale is used for rating. For the results, the average O-simplicity score is 3.94, indicating the simplification of the articles is generally good. The Simplicity-phrase and Simplicity-structure scores reach 4.28 and 4.23, respectively, implying that the simplified article has made considerable lexical and sentence structure simplifications compared to the original article. The grammar score achieves 4.65, probably because the simplified articles are written by humans and are easy to read. The meaning score is 3.69, indicating that the simplified article can preserve the meaning of the original article.

5 The D-SARI Metric

Currently, the most commonly used automatic evaluation metric for sentence-level simplification is the SARI metric (Xu et al., 2016). The correlation of SARI with human judgments of the simplicity indicator proved to be high in sentence-level simplification (Sulem et al., 2018a). However, this metric has shortcomings when used directly to evaluate document-level simplification. For better understanding, in this section, we conduct a qualitative analysis and give the following example:

Original article: *marengo is a town in and the county seat of iowa county , iowa , united states .*

	SARI	F_{keep}	P_{del}	F_{add}	D-SARI	D_{keep}	D_{del}	D_{add}
Simplified article 1	54.24	23.74	88.18	50.80	42.80	23.74	88.18	16.49
Simplified article 2	64.90	33.68	98.08	62.95	41.00	11.86	48.19	62.95
Simplified article 3	66.80	67.63	96.44	36.33	42.91	25.68	66.72	36.33
Simplified article 4	49.93	51.39	91.25	7.14	48.69	50.06	88.88	7.14

Table 6: The SARI and D-SARI values for the four simplified articles. The fourth simplified article does the best job of simplification, not only retaining the main meaning of the original article but also deleting the unimportant information and the difficult words. However, its SARI value is the lowest among the simplified articles.

it has served as the county seat since august 1845 , even though it was not incorporated until july 1859 . the population was 2,528 in the 2010 census , a decline from 2,535 in 2000 .

Simplified article 1: *in the US . 2,528 in 2010 .*

Simplified article 2: *marengo is a city in iowa , the US . it has served as the county seat since august 1845 , even though it was not incorporated . the population was 2,528 in the 2010 census , a decline from 2,535 in 2010 .*

Simplified article 3: *marengo is a town in iowa . marengo is a town in the US . in the US . the population was 2,528 . the population in the 2010 census .*

Simplified article 4: *marengo is a town in iowa , united states . in 2010 , the population was 2,528 .*

Reference article: *marengo is a city in iowa in the US . the population was 2,528 in 2010 .*

The SARI values of the four simplified articles are shown in the left half of Table 6[5]. When using the SARI metric, we take the whole article as input, output and reference.

The simplified article 1 generates "US", a word that does not appear in the original article, which raises the overall SARI value. Intuitively, however, it makes no sense to generate several simplified words if the simplified article is too short to convey the main meaning of the original article. Therefore, we believe that a penalty factor LP_1 should be added to the F_{add} score. If the generated simplified article is shorter than the reference article, the F_{add} score will be penalized.

The simplified article 2 does a good job of simplifying the first sentence of the original article but retains much useless information and difficult words. Paradoxically, its P_{del} score is the highest among the four simplified essays. A more common scenario is that the original article is much longer than the simplified article, then according to the formula of P_{del}, removing fewer words will have a limited effect on P_{del}. Therefore, we believe that a penalty factor LP_2 should be added to the P_{del} score, penalizing the P_{del} score if the generated simplified article is longer than the referenced article.

The simplified article 3 finds the important information in the original article and does a good job of simplifying it. Nevertheless, it performs duplicate generation, which leads to a severe decrease in readability and should not get such a high F_{keep} value. According to the formula of F_{keep}, if the duplicate n-grams also appear in the reference, then the F_{keep} value will not decrease. Therefore, we believe that the LP_2 should also be added to the F_{keep} score. Besides, we add a sentence-level penalty factor SLP to penalize the F_{keep} score if the number of generated sentences is far from the number of sentences in the reference article.

In summary, based on the SARI metric (Xu et al., 2016), we propose the D-SARI metric for the document-level simplification task. We retain the idea of calculating the scores of add, keep and delete separately in SARI, which proved to be effective in sentence simplification. The D-SARI metric is shown as below:

$$\text{LP}_1 = \begin{cases} 1 & O \geq R \\ e^{\frac{O-R}{O}} & O < R \end{cases}$$
$$\text{LP}_2 = \begin{cases} 1 & O \leq R \\ e^{\frac{R-O}{max(I-R,1)}} & O > R \end{cases} \quad (2)$$

$$\text{SLP} = e^{-\frac{||R_S - O_S||}{max(R_S, O_S)}} \quad (3)$$

$$\begin{aligned} D_{keep} &= F_{keep} * \text{LP}_2 * \text{SLP} \\ D_{add} &= F_{add} * \text{LP}_1 \\ D_{del} &= P_{del} * \text{LP}_2 \end{aligned} \quad (4)$$

$$\text{D-SARI} = (D_{keep} + D_{del} + D_{add}) * 1/3 \quad (5)$$

[5]We use the script in https://github.com/cocoxu/simplification/blob/master/SARI.py.

	D-SARI↑	D_{keep}	D_{del}	D_{add}	SARI↑	F_{keep}	P_{del}	F_{add}	BLEU↑	FKGL↓
Transformer	37.38	31.30	68.80	12.04	44.46	43.14	75.07	15.16	21.70	**17.83**
SUC	12.92	13.05	22.27	3.44	34.13	39.78	59.05	3.56	18.13	59.31
BertSumextabs	**39.88**	**35.71**	**72.06**	11.87	47.39	**50.68**	76.98	14.50	26.96	18.32
BART	37.24	34.34	62.41	**14.98**	**48.34**	50.50	**77.72**	**16.80**	**31.77**	31.90

Table 7: The automatic evaluation results on the D-Wikipedia test set. We use **Bold** to mark the best result and underline the second best result.

	D-SARI↑	SARI↑	BLEU↑	FKGL↓
Transformer	27.03	28.46	0.11	27.06
SUC	5.80	38.37	16.40	343.42
BertSumextabs	27.68	30.23	0.20	26.40
BART	28.56	32.29	0.64	43.39

Table 8: The automatic evaluation results on the Newsela Simp-4 test set. The BART model performs the best in terms of D-SARI, but the results of all models decline to some degree compared to Table 7.

I, O, and R represent the number of words (including punctuation) in the input article, the output article, and the reference article, respectively. O_S and R_S represent the number of sentences in the output article and the reference article, respectively. Due to the limitation of space, please refer to Xu et al. (2016) for the calculation of F_{add}, F_{keep} and P_{del}. We also calculate the D-SARI values for each of the simplified articles in the given example, as shown in the right half of Table 6.

As we analyzed from the given example, it is reasonable to penalize the three components in SARI. In the D-SARI metric, the penalty is based on length. The motivation comes from BLEU (Papineni et al., 2002). A candidate should be neither too long nor too short, and an evaluation metric should enforce this. The difference between the length of a simplified sentence and the original sentence in sentence-level text simplification is not very large, while the opposite is true for document-level text simplification. An original article may be long, while a simplified article may contain only one sentence. It is simple enough, but not a good simplification of the original article. It is a reasonable proposition that the length of the simplified article should be close to the length of the reference article.

We also conduct an empirical analysis of the D-SARI metric. In Section 7.3, we use Spearman's rank correlation coefficient (Zwillinger and Kokoska, 1999) to show that the D-SARI metric has the strongest correlation among several metrics with human ratings.

6 Baseline Models

We selected four representative models as the baselines for the document-level simplification task, which are:

(1) Transformer: It treats the task as a sequence-to-sequence problem. Both the encoder and decoder contain six transformer layers (Vaswani et al., 2017).

(2) SUC: It simplifies each sentence in the article by using use contextual information (Sun et al., 2020).

(3) BertSumextabs: It achieves excellent results on the text summarization task, using the Bert-base model as the encoder (Liu and Lapata, 2019).

(4) BART: It is a recently proposed pre-trained model based on large-scale corpus and achieves state-of-the-art results on many sequence-to-sequence tasks (Lewis et al., 2019).

All the models were tested on our delineated test sets. We used the fairseq toolkit and performed replicate experiments. See Appendix B for detailed parameters.

7 Evaluation Results

7.1 Automatic Evaluation Results

We used the SARI metric, the BLEU metric, the FKGL metric, and the D-SARI metric for automatic evaluation. We have described the SARI and D-SARI metrics in detail in Section 5. BLEU is a method for comparing the similarity between the reference and the output (Papineni et al., 2002)[6]. FKGL is used to measure the readability of the text (Kincaid et al., 1975)[7].

The automatic evaluation results of the D-Wikipedia test set are shown in Table 7. The BertSumextabs model obtains the best results on the D-SARI value. The BART model obtains the best

[6]We used the script in https://github.com/nltk/nltk/blob/develop/nltk/translate/bleu_score.py.

[7]We used the script in https://github.com/shivam5992/textstat/blob/master/textstat/textstat.py.

	Simplicity -phrase	Simplicity -structure	Meaning	Grammar	O-simplicity	Average length
Transformer	**4.77**	**4.74**	2.76	4.50	2.79	58.28
SUC	3.09	2.95	**4.57**	3.78	2.63	194.39
BertSumextabs	4.61	4.50	3.60	**4.70**	3.62	44.21
BART	4.06	4.00	4.23	**4.70**	3.59	86.42
Reference	4.28	4.23	3.69	4.65	**3.94**	81.46

Table 9: The results of human evaluation on the 100 selected article pairs. We use **Bold** to mark the best result and underline the second-best result. The five-point Likert scale is used for rating.

Spearman's ρ	Simplicity -phrase	Simplicity -structure	Meaning	Grammar	O-simplicity
BLEU	-0.14	-0.12	**0.23**	0.09	0.30
SARI	0.28	0.34	-0.22	0.29	0.36
-FKGL	**0.66**	**0.67**	-0.63	**0.47**	0.18
D-SARI	0.42(+0.14)	0.47(+0.13)	-0.30(-0.08)	0.38(+0.09)	**0.42**(+0.06)

Table 10: Correlation of the automatic metrics against the human ratings. We use **Bold** to mark the best result. Because the simpler the article, the lower the FKGL value, we report -FKGL for better comparison. The differences between SARI and D-SARI are also shown in brackets.

results on the SARI value and the BLEU value. The transformer model obtains the best results on the FKGL value. We also give an example to compare the outputs of different models, and we put it in Appendix D.

Spearman's ρ	The length of simplified article
Simplicity-phrase	-0.65
Simplicity-structure	-0.65
Meaning	0.56
Grammar	-0.50
O-simplicity	-0.26

Table 11: Correlation of the simplified article's length against the human ratings. The simplicity-phrase and simplicity-structure score have a strong negative correlation with the article's length, while the O-simplicity score has a weak correlation.

For the Newsela corpus, as mentioned in Section 4.2, we choose a representative test set called Simp-4 to show the automatic results. The models were both trained and validated on the D-Wikipedia dataset, and the results are shown in Table 8.

7.2 Human Evaluation Results

We performed human evaluation according to the method described in Section 4.4. To maintain consistency, we selected the same 100 article pairs in the D-Wikipedia test set that were randomly selected for evaluating the dataset in Section 4.4. We added some fake examples to the questionnaire and checked whether the workers gave a reasonable score to ensure the quality of human evaluation.

The human evaluation results are shown in Table 9. We also report the correlation of the article's length against the human ratings, as shown in Table 11. The results prove that human judges tend to give high simplicity-phrase scores and high simplicity-structure scores to short articles. The O-simplicity indicator places more emphasis on the overall simplification effect, including the retention of the main meaning and the fluency of the sentences. Therefore, as we analyzed in Section 4.4, the O-simplicity indicator can evaluate how good the simplification is, which is better than the simplicity-phrase and the simplicity-structure indicators. Generally, the BART and BertSumextabs models perform better than the other two models, especially on the O-simplicity measure. Directly applying the sentence simplification model SUC does not get good results, which means document-level simplification is very different from sentence-level simplification.

7.3 Correlation of Automatic Metrics with Human Ratings

We calculated Spearman's rank correlation coefficient between each automatic metric and human ratings on the results for the 100 article pairs, and the correlation scores are shown in Table 10. The

D-SARI metric has the highest correlation with the O-simplicity indicator, surpassing both BLEU and SARI. In terms of simplicity-phrase and simplicity-structure, the correlation of D-SARI with human ratings also exceeds that of SARI, and although FKGL has the highest correlation, it does not correlate with the O-simplicity indicator. We also noticed that BLEU has little correlation with the meaning and grammar indicators, probably because the simplification contains lots of splitting operations, which is consistent with the conclusion obtained by Sulem et al. (2018a).

7.4 The Challenge of Document-level Simplification

There are many problems with applying existing models directly to the document-level simplification task. From the automatic evaluation, the D_{keep} values of the baseline models are not high, and the FKGL values also need to be further reduced. From human evaluation, the O-simplicity scores of the articles simplified by the models are still far from that of the reference.

As can be seen from the given example in Appendix D, the best-performing BertSumextabs model among the four models still retains some complex vocabulary and sentence structure compared with the reference, and the model's ability to screen out important information needs further improvement. We also noticed that the results of the SUC model are much lower than all other models, which indicates that document-level simplification cannot be addressed by stitching together the results of sentence simplification as simplified articles.

Above all, we believe that new models designed for document-level simplification could be proposed in the future, which will greatly advance this field.

8 Conclusion

In this paper, we are committed to promoting research on document-level text simplification. We established a large-scale high-quality dataset named D-Wikipedia and proposed a new automatic evaluation metric called D-SARI. We also selected several representative models as baselines for this task. The results demonstrate that the dataset is of high quality and the metric is reliable.

Acknowledgements

We are grateful to the reviewers for their valuable comments.

This work was supported by National Natural Science Foundation of China (61772036), Beijing Academy of Artificial Intelligence (BAAI) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). Xiaojun Wan is the corresponding author.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231.

William Coster and David Kauchak. 2011. Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation*, pages 1–9.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. Editnts: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748.

Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Biological, translational, and clinical language processing*, pages 49–56.

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.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural crf model for sentence alignment in text simplification. *arXiv preprint arXiv:2005.02324*.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73.

David Kauchak. 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.

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.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. 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 3137–3147.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3721–3731.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2020. Controllable text simplification with explicit paraphrasing. *arXiv preprint arXiv:2010.11004*.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445.

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.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. Massalign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4.

Gustavo H Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3761–3767.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.

Advaith Siddharthan. 2003. Preserving discourse structure when simplifying text. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*.

Advaith Siddharthan and Napoleon Katsos. 2010. Reformulating discourse connectives for non-expert readers. In *Human Language Technologies: The 2010 Annual Conference of the North American*

Chapter of the Association for Computational Linguistics, pages 1002–1010.

Sanja Štajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, and Heiner Stuckenschmidt. 2017. Sentence alignment methods for improving text simplification systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 97–102.

Sanja Štajner and Horacio Saggion. 2018. Data-driven text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 19–23.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. Bleu is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018c. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173.

Renliang Sun, Zhe Lin, and Xiaojun Wan. 2020. On the helpfulness of document context to sentence simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1411–1423.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420.

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.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9709–9716.

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.

Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. Crc Press.

A Six Types of Operations in Document-Level Text Simplification

1 Sentence joining and sentence reordering

Src: the fields medal is a prize awarded to two , three , or four mathematicians under 40 years of age at the international congress of the international mathematical union (imu) , a meeting that takes place every four years. the fields medal is regarded as one of the highest honors a mathematician can receive , and has been described as the mathematician 's nobel prize , although there are several key differences , including frequency of award , number of awards , and age limits . according to the annual academic

excellence survey by arwu , the fields medal is consistently regarded as the top award in the field of mathematics worldwide , and in another reputation survey conducted by ireg in 2013–14 , the fields medal came closely after the abel prize as the second most prestigious international award in mathematics. the prize comes with a monetary award which , since 2006 , has been 15,000 . the name of the award is in honour of canadian mathematician john charles fields . fields was instrumental in establishing the award , designing the medal itself , and funding the monetary component. the medal was first awarded in 1936 to finnish mathematician lars ahlfors and american mathematician jesse douglas , and it has been awarded every four years since 1950 . its purpose is to give recognition and support to younger mathematical researchers who have made major contributions .

Tgt: the fields medal is a prize given to mathematicians who are not over 40 years of age . it is given at each international congress of the international mathematical union . this is a meeting that takes place every four years. the canadian mathematician john charles fields was the first to propose this medal and it was first awarded in 1936 . it has been regularly awarded since 1950 . its purpose is to support younger mathematicians who made major contributions. the fields medal is viewed , at least in the media , as the top honor a mathematician can receive .

Analysis: Sentence joining means combining two or more sentences into one sentence. In src, the two sentences marked in red are merged into the sentence marked in red in tgt. Some information in the original two sentences is removed. Sentence reordering implies a change in the structure of the article. In src, the sentences marked in red appear before the sentence marked as underlined, but in tgt, the simplified sentence marked as underlined appears before the sentence marked in red.

2 Sentence splitting

Src: it is a decentralized digital currency without a central bank or single administrator that can be sent from user to user on the peer-to-peer bitcoin network without the need for intermediaries .

Tgt: bitcoin is a digital and global money system currency . it allows people to send or receive money across the internet , even to someone they do n't know or do n't trust . money can be exchanged without being linked to a real identity .

Analysis: In contrast to sentence joining, sentence splitting is the division of a long sentence into two or more sentences. The sentences in tgt are simplified from the parts of the sentence in src marked with the corresponding colors.

3 Sentence addition

Src: 104.6 rtl is a private radio station that is produced in a hot adult contemporary format . it is transmitted from studios in kurfürstendamm in berlin-charlottenburg . according to german media analysis 2011/ii , the station reaches 209,000 listeners in an average transmitting hour (mon-fri , 6am-6pm) with a total of 709,000 listeners per day and thereby is one of the most listened to radio programs in berlin and brandenburg .

Tgt: 104.6 rtl is a german radio station . it first aired on 9 september , 1991 . it broadcasts in berlin and hopes that the 14-39 age group will listen . the studios are at the kurfürstendamm in berlin-charlottenburg . in 2005 the radio channel has been awarded the german radio award for the best morning show .

Analysis: Sentence addition means that there is a sentence in tgt for which the corresponding sentence is not found in src. Sentence addition introduces additional information, often used for explanation and clarification.

4 Sentence deletion

Src: landudal is a commune in the finistère department of brittany in north-western france . the writer angèle jacq , winner of the cezam prix littéraire inter ce in 2000 for her novel " le voyage de jabel " , was born in landudal .

Tgt: landudal is a commune . it is found in the region brittany in the finistère department in the northwest of france .

Analysis: Sentence deletion means that there is a sentence in src that does not find a corresponding sentence in tgt. This is usually because the original sentence is difficult to simplify and the deletion does not affect the main meaning of the text.

5 Anaphora resolution

Src: Winston Churchill was a great politician and statesman. He also won the Nobel Prize for literature in 1953.

Tgt: Winston Churchill won the Nobel Prize in 1953.

Analysis: Anaphora resolution is usually associated with sentence deletion. the first sentence in src is deleted, then the pronoun "he" in the second sentence is replaced with the person's name in tgt.

B Implementation Details

We used the fairseq toolkit[8] to implement the transformer model and the BART model. We used the code on the github to implement the BertSumextabs model[9] and the SUC model[10]. All the models except the SUC model are trained on the training set of the D-Wikipedia dataset we constructed. The SUC model is trained on the Wikipedia dataset and when testing, the original articles in our delineated test set are simplified with this model sentence by sentence, and then the output sentences are stitched together to get the simplified articles. All the models are trained on Nvidia GTX 1080ti. The batchsize we set can make full use of its video memory. The hyperparameters are shown in the following tables.

hyperparameter	value
learning rate	1e-3
dropout	0.1
max tokens	2048
update freq	4
label smoothing	0.1
weight decay	1e-4
num updates	1e5

Table 12: The hyperparameters of the transformer model.

hyperparameter	value
learning rate	1e-4
dropout	0.1
max tokens	2048
update freq	4
label smoothing	0.1
weight decay	1e-4
num updates	1e5

Table 13: The hyperparameters of the BART model.

hyperparameter	value
learning rate	0.1
activation function	GELU
batchsize	16
encoder layers	4
decoder layers	4
multi-heads	4
max epochs	50

Table 14: The hyperparameters of the SUC model.

hyperparameter	value
max learning rate	2e-3
dropout	0.1
batchsize	500
update fraq	8
num updates	50000

Table 15: The hyperparameters of the BertSumextabs model (ext).

hyperparameter	value
max learning rate(encoder)	2e-3
max learning rate(decoder)	0.2
dropout	0.2
batchsize	12
update fraq	20
num updates	50000

Table 16: The hyperparameters of the BertSumextabs model (abs).

C Human Evaluation Guideline

The goal of this review is to evaluate the simplification quality of different articles. In this review, you will be given an original article and its corresponding simplified articles. You should evaluate the quality of the simplification in the following five ways:

(1) **Simplicity-phrase.** Are the words in the simplified article simpler than those in the original article?

(2) **Simplicity-structure.** Are the sentence structures in the simplified article simpler than those in the original article?

(3) **Meaning.** The text simplification operation can remove some sentences from the original article, but the main meaning of the original article should be kept intact.

(4) **Grammar.** The simplified article should be grammatically correct and fluent.

(5) **O-simplicity.** The simplified article should be

[8] https://github.com/pytorch/fairseq

[9] https://github.com/nlpyang/PreSumm

[10] https://github.com/RLSNLP/Document-Context-to-Sentence-Simplification

simpler than the original article, and it also should read smoothly and can retain the main meaning of the original article.

You will do this using a 1-5 rating scale, where 5 is the best and 1 is the worst. There are no "correct" answers and whatever choice is appropriate for you is a valid response. For example, if you are given the following original article and simplified articles:

Original article: the fields medal is a prize awarded to two, three, or four mathematicians under 40 years of age at the international congress of the international mathematical union (imu), a meeting that takes place every four years. the fields medal is regarded as one of the highest honors a mathematician can receive, and has been described as the mathematician 's nobel prize, although there are several key differences, including frequency of award, number of awards, and age limits. according to the annual academic excellence survey by arwu, the fields medal is consistently regarded as the top award in the field of mathematics worldwide, and in another reputation survey conducted by ireg in 2013–14, the fields medal came closely after the abel prize as the second most prestigious international award in mathematics. the prize comes with a monetary award which, since 2006, has been 15,000. the name of the award is in honour of canadian mathematician john charles fields. fields was instrumental in establishing the award, designing the medal itself, and funding the monetary component. the medal was first awarded in 1936 to finnish mathematician lars ahlfors and american mathematician jesse douglas, and it has been awarded every four years since 1950. its purpose is to give recognition and support to younger mathematical researchers who have made major contributions. in 2014, the iranian mathematician maryam mirzakhani became the first female fields medalist. in all, sixty people have been awarded the fields medal. the most recent group of fields medalists received their awards on 1 august 2018 at the opening ceremony of the imu international congress, held in rio de janeiro, brazil. the medal belonging to one of the four joint winners, caucher birkar, was stolen shortly after the event. the icm presented birkar with a replacement medal a few days later.

Simplified article 1: (Score: Simplicity-phrase 5 Simplicity-structure 5 Meaning 5 Grammar 5 O-simplicity 5)

the fields medal is an award given to mathematicians under 40 years of age. the name of the prize is in honor of the canadian mathematician john charles field. and it is awarded every four years since 1950. the fields medal is regarded as the highest award in the field of mathematics in the world. it is intended to be used to encourage young mathematicians.

Simplified article 2: (Score: Simplicity-phrase 4 Simplicity-structure 5 Meaning 5 Grammar 5 O-simplicity 5)

the fields medal is a prize given to mathematicians who are not over 40 years of age. it is given at each international congress of the international mathematical union. this is a meeting that takes place every four years. the canadian mathematician john charles fields was the first to propose this medal and it was first awarded in 1936. it has been regularly awarded since 1950. its purpose is to support younger mathematicians who made major contributions. the fields medal is viewed, at least in the media, as the top honor a mathematician can receive. it comes with a monetary award. in 2006 the award was $ 15,000 (us $ 13,400 or €10,550). the abel prize has similar prestige, and more money.

Simplified article 3: (Score: Simplicity-phrase 4 Simplicity-structure 3 Meaning 2 Grammar 5 O-simplicity 2)

the fields medal is consistently regarded as the top award in the field of mathematics worldwide. Since 2006, the prize of this award has been 15,000. the most recent group of fields medalists received their awards on 1 august 2018 at the opening ceremony of the imu international congress. the medal belonging to one of the four joint winners, caucher birkar , was stolen shortly after the event .

Simplified article 4: (Score: Simplicity-phrase 1 Simplicity-structure 1 Meaning 5 Grammar 5 O-simplicity 1)

the fields medal is a prize awarded to two, three, or four mathematicians under 40 years of age at the international congress of the international mathematical union (imu) a meeting that takes place every four years. according to the annual academic excellence survey by arwu, the fields medal is consistently regarded as the top award in the field of mathematics worldwide, and in another reputation survey conducted by ireg in 2013–14, the fields medal came closely after the abel prize as the second most prestigious international award in mathematics. the name of the award is in honour of

canadian mathematician john charles fields. he was instrumental in establishing the award, designing the medal itself, and funding the monetary component. the purpose of the fields medal is to give recognition and support to younger mathematical researchers who have made major contributions.

Simplified article 5: (Score: Simplicity-phrase 5 Simplicity-structure 5 Meaning 4 Grammar 1 O-simplicity 2)

the fields medal is to a prize giving to mathematicians who are not over 40 years of age. but it is awarding every four years since 1950. the prize is in honor of the canadian mathematician john charles field. the fields metal described as the mathematician 's nobel prize as the mathematician 's nobel prize. its purpose are to support younger mathematicians who made major contributions.

Analysis: The **Simplified article 1** does a good job on the simplification of words and sentence structures. The simplification includes removing difficult vocabulary, splitting and simplifying long sentences, etc.. So, it scores full marks for the simplification-phrase and the simplification-structure. It is equally able to summarize the main meaning of the original article, so it scores full marks for meaning. It reads smoothly, like it is written by humans, so it scores full marks for grammar. The overall feeling of the article is very good. It reads very simple, fluently and maintains the main meaning, so it scores full marks for the O-simplicity. The **Simplified article 2** has some words that need further simplification, such as "prestige" and "monetary". So, it scores a little bit lower than the simplified article 1 on the simplicity-phrase. However, it reads smoothly and the main meaning is well maintained. One will also feel that the simplification effect is very good when reading this article. These two articles also illustrate that articles that score high marks can be presented in different ways. Obviously, **simplified article 3** does not retain the main meaning of the original article, but rather some non-essential information. Therefore, it scores very low on meaning. Besides, it contains long and complex sentences and the sentence structures are not simple enough compared to the original article. One' s experience of reading such an article is not very good, because it deviates from the main meaning and is not simple enough. The **Simplified article 4** is able to find those relatively important sentences in the original article. But unfortunately, it does little simplification operation and is not easy to read, so it scores very low on the simplification-phrase and the simplification-structure. Children and non-native speakers will not be able to read such an article, so it scores very low on the O-simplicity. The **Simplified article 5** contains many grammatical errors and repetition of some phrases, making it look less like it is written by a human. Therefore, it scores very low on grammar. Although its words and sentence structures are very simple, the existence of grammatical errors makes it difficult to read, so it scores low on the O-simplicity.

D Case Study

Input: atal bihari vajpayee (; 25 december 1924 – 16 august 2018) was an indian statesman who served three terms as the prime minister of india , first for a term of 13 days in 1996 , then for a period of 13 months from 1998 to 1999 , followed by a full term from 1999 to 2004 . a member of the bharatiya janata party (bjp) , he was the first indian prime minister not of the indian national congress to serve a full term in office . he was also noted as a poet and a writer . he was a member of the indian parliament for over five decades , having been elected ten times to the lok sabha , the lower house , and twice to the rajya sabha , the upper house . he served as the member of parliament for lucknow , retiring from active politics in 2009 due to health concerns . he was among the founding members of the bharatiya jana sangh (bjs) , of which he was president from 1968 to 1972 . the bjs merged with several other parties to form the janata party , which won the 1977 general election . in march 1977 , vajpayee became the minister of external affairs in the cabinet of prime minister morarji desai . he resigned in 1979 , and the janata alliance collapsed soon after . former members of the bjs formed the bjp in 1980 , with vajpayee its first president . during his tenure as prime minister , india carried out the pokhran - ii nuclear tests in 1998 . vajpayee sought to improve diplomatic relations with pakistan , travelling to lahore by bus to meet with prime minister nawaz sharif . after the 1999 kargil war with pakistan , he sought to restore relations through engagement with president pervez musharraf , inviting him to india for a summit at agra . the administration of narendra modi declared in 2014 that vajpayee 's birthday , 25 december , would be marked as good governance

day . in 2015 , he was conferred india 's highest civilian honour , the bharat ratna , by the president of india , pranab mukherjee . he died on 16 august 2018 of age - related illness .

Reference: atal bihari vajpayee (25 december 1924 – 16 august 2018) was an indian politician . he served as the 10th prime minister of india for six years from 1998 through 2004 . he was also serving as prime minister from may 1996 through june 1996 . he founded the bharatiya janata party . vajpayee was born on 25 december 1924 in gwalior , united provinces , british india (now madhya pradesh , india) . he studied at victoria college (now laxmibai college) and at dav college , kanpur . vajpayee was never married and had no children . vajpayee , who had dementia and diabetes , died at a new delhi hospital from multiple organ failure on 16 august 2018 , aged 93 .

Transformer: amar prasad (25 december 1924 – 16 august 2018) was an indian politician . he was a member of the indian national congress party . he served as prime minister of india from 1999 to 2004 . he was also the first prime minister of india . he was born in <unk> , <unk> died on 16 august 2018 at the age of 93 .

SUC: atal bihari vajpayee (; 25 december 1924 – 16 august 2018) was an indian statesman who served three terms as the prime minister of india , first for a term of 13 days in 1996 , then for a period of 13 months from 1998 to 1999 , followed by a full term from 1999 to 2004 . a member of the bharatiya janata party (bjp) , he was the first indian prime minister not of the indian national congress to be elected into the full term in the first indian prime minister . he was the first indian prime minister of the indian national congress and was called the bharatiya janata party (bjp) . he was a member of the indian parliament for over five decades and has been elected ten times to the lok sabha , the lower house , and twice to the rajya sabha , the upper house . he was elected ten times to the lok sabha , the lower house , and twice to the rajya sabha , the upper house . he served as the member of parliament for lucknow , retiring from active politics in 2009 due to health concerns . he was among the founding members of the bharatiya jana sangh (bjs) , of which he was president from 1968 to 1972 . the bjs merged with several other parties to form the janata party , which won the 1977 general election . it to make the janata party , he won the 1977 general election . in march 1977 , vajpayee became the minister of external affairs in the cabinet of prime minister morarji desai . he became the minister of external affairs in the cabinet of prime minister morarji desai . he left in 1979 and left the alliance collapsed soon after . in 1980 's first president.during members of prime minister , india carried out the bjp nuclear tests as prime minister . in 1998 he became prime minister , india , and created the pokhran-ii nuclear tests . vajpayee sought to improve diplomatic relations with pakistan , travelling to lahore by bus to meet with prime minister nawaz sharif . after the 1999 kargil war with pakistan , he sought to restore relations through engagement with president pervez musharraf , inviting him to india for a summit at agra.the administration of narendra modi declared in 2014 that vajpayee 's birthday , 25 december , would be marked as good governance day . in 2015 , he was conferred india 's highest civilian honour , the bharat ratna , by the president of india , pranab mukherjee . he died on 16 august 2018 of age-related illness .

BertSumextabs: atal bihari vajpayee (; 25 december 1924 – 16 august 2018) was an indian politician . he served three terms as the prime minister of india , first for a term of 13 days in 1996 , then for a period of 13 months from 1998 to 1999 , followed by a full term from 1999 to 2004 . a member of the bharatiya janata party (bjp) , he was the first indian prime minister not of the indian national congress .

BART: atal bihari vajpayee (25 december 1924 – 16 august 2018) was an indian politician . he was the first indian prime minister not of the indian national congress to serve a full term in office.he was elected as a member of the indian parliament for over five decades , having been elected ten times to the lok sabha , the lower house , and twice to the rajya sabha , the upper house . he served as the member of parliament for lucknow , retiring from active politics in 2009 due to health concerns.he was among the founding members of the bharatiya jana sangh (bjs) , of which he was president from 1968 to 1972.in march 1977 , vajpayee became the minister of external affairs in the cabinet of prime minister morarji desai . he resigned in 1979 , and

the janata alliance collapsed soon after . former members of the bjs formed the bjp in 1980 , with vajpayee its first president.during his term as prime minister , india carried out the pokhran-ii nuclear tests in 1998 . vajpayee sought to improve diplomatic relations with pakistan , travelling to lahore by bus to meet prime minister nawaz sharif . after the 1999 kargil war with pakistan , he sought to restore relations through engagement with president pervez musharraf , inviting him to india for a summit at agra. vajpayee died on 16 august 2018 in lucknow , aged 93 .

Analysis: We use red to mark sentences with factual errors. We use blue to mark sentences that should have been deleted but are not deleted and not simplified, and we use cyan to mark sentences with grammatical errors. The output articles of the SUC model and the BART model are too long and retain a large number of unsimplified sentences in the input article. The output article of the Transformer model contains many factual errors and is poorly readable. The BertSumextabs model simplifies a less important sentence in the original article, and the simplification is not reasonable. Because it removes the critical information of “to serve a full term in office", the meaning of the sentence may be changed. Besides, the BertSumextabs model do not keep the information about the person’s death from the original article.