Gauging the Gap Between Human and Machine Text Simplification Through Analytical Evaluation of Simplification Strategies and Errors

Daichi Yamaguchi **Rei Miyata** **Sayuka Shimada** **Satoshi Sato**
Nagoya University, Nagoya, Japan
{yamaguchi.daichi.e4,shimada.sayuka.y9}@s.mail.nagoya-u.ac.jp
{miyata,ssato}@nuee.nagoya-u.ac.jp

Abstract

This study presents an analytical evaluation of neural text simplification (TS) systems. Because recent TS models are trained in an end-to-end fashion, it is difficult to grasp their abilities to perform particular simplification operations. For the advancement of TS research and development, we should understand in detail what current TS systems can and cannot perform in comparison with human performance. To that end, we first developed an analytical evaluation framework consisting of fine-grained taxonomies of simplification strategies (at both the surface and content levels) and errors. Using this framework, we annotated TS instances produced by professional human editors and multiple neural TS systems and compared the results. Our analyses concretely and quantitatively revealed a wide gap between humans and systems, specifically indicating that systems tend to perform deletions and local substitutions while excessively omitting important information, and that the systems can hardly perform information addition operations. Based on our analyses, we also provide detailed directions to address these limitations.

1 Introduction

Text simplification (TS) is the task of reducing the content and structural complexity of text while retaining the core part of the original meaning (Alva-Manchego et al., 2020). TS can not only facilitate the text reading by children or language learners, but also improve the performance of downstream NLP applications, including machine translation and summarization (Siddharthan et al., 2004; Štajner and Popovic, 2016).

Early studies on TS have separately dealt with lexical simplification (Glavaš and Štajner, 2015) and syntactic simplification (Scarton et al., 2017), and developed simplification techniques specialized for particular linguistic phenomena. In contrast, recent studies have tackled TS as a task of monolingual translation from a complex to a simplified language using deep neural networks. While neural TS has demonstrated steady improvement, few studies have attempted to assess what kind of editing operations are performed by the systems in concrete terms. To further advance TS research and development, we should understand the potential and limitations of current TS technologies and precisely grasp the gap between human and machine TS. To do so, we need analytical frameworks that can be applied to both human and machine TS. In contrast to (machine) translation research and practice, where several frameworks have been developed for analyzing translation strategies (Chesterman, 2016) and errors (Lommel et al., 2014), no well-established framework tailored for TS tasks is available.

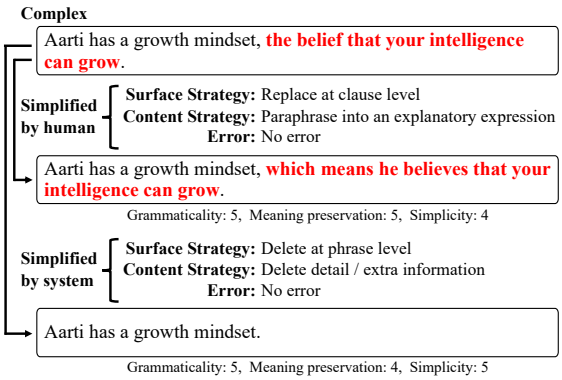

Figure 1: Example of an analytical evaluation in terms of editing strategies and errors.

Therefore, in this study, we first propose an analytical evaluation framework consisting of taxonomies of editing strategies (both at the surface and content levels) and errors. We then report an experiment in which we apply our framework to instances of human and machine TS in various settings, and concretely describe the gap between them. Figure 1 shows an example of the evaluation using our framework, illustrating the detailed differences in editing operations between humans and

Findings of the Association for Computational Linguistics: EACL 2023, pages 359–375

TS systems. Our results revealed that current neural TS systems can frequently replace local spans, while excessively deleting important parts. Moreover, TS systems cannot perform operations related to content addition, such as the addition of detail information. These findings enable us to understand the fundamental challenges of current technologies and pursue a promising avenue to fill the gap between humans and machines.

2 Related Work

To evaluate TS systems, automatic evaluation metrics such as SARI (Xu et al., 2016), BLEU (Papineni et al., 2002), and Flesch–Kincaid Grade Level (FKGL) (Kincaid et al., 1975) are widely used. SARI and BLEU use n-gram overlap between target sentences and human-created references, whereas FKGL uses the number of syllables and words in the output sentences. These metrics can be easily calculated if references are available and are indispensable for the rapid cycle of system development and evaluation. However, their limitations and pitfalls have been acknowledged. Sulem et al. (2018a), for example, reported that BLEU is negatively correlated with human evaluation scores, such as simplicity and grammaticality. Tanprasert and Kauchak (2021) also showed that the FKGL score can easily be manipulated by minor modifications, such as adding periods randomly.

Subjective human evaluation has also been implemented (Štajner and Nisioi, 2018; Sulem et al., 2018b; Al-Thanyyan and Azmi, 2021). In many cases, certain aspects of TS quality, namely grammaticality/fluency, meaning preservation/adequacy, and simplicity, are rated on a three- to five-point Likert scale based on the evaluation criteria.

Importantly, all the abovementioned evaluation methods only provide summative numerical scores. These scores are useful for comparing the general performances of different systems, but do not necessarily provide a guidepost for achieving higher system performance. To gain a detailed understanding of what TS systems can/cannot do vis-à-vis editing operations by humans, analytical evaluation methods are required.

The analytical evaluation of TS can be broadly divided into strategy and error analyses. The former concerns the type of editing operation (strategy) performed to produce the simplified text. Previous studies have acknowledged general strategies, such as paraphrasing, deletion, and splitting (Shardlow, 2014), and document-level strategies, such as sentence reordering and sentence-joining operations (Alva-Manchego et al., 2019b). However, these roughly typify superficial textual changes rather than detailed content-level changes that capture editing operations peculiar to TS. The latter concerns the type of error in the resulting simplified text. In contrast to automatic and human evaluations, fewer attempts have been made to conduct an error analysis (Maddela et al., 2021).[1]

Proper implementation of analytical evaluation requires well-formulated frameworks to classify textual phenomena observed in the outputs. The general editing strategies mentioned above and some guidelines for human writers (Mitkov and Štajner, 2014) are not sufficiently concrete for fine-grained analysis. Although several typologies of simplification operations (e.g., Amancio and Specia, 2014; Brunato et al., 2014; Koptient et al., 2019) and editing guidelines for human writers (Mitkov and Štajner, 2014) have been proposed, the following limitations can be generally acknowledged: (1) content-level operations are not fully covered; (2) their applicability to outputs of automatic TS systems has not been verified. In the field of translation studies, a wide variety of translation strategies have been proposed to describe the differences between source and target texts (e.g., Vinay and Darbelnet, 1958; Molina and Hurtado Albir, 2002). Chesterman (2016), for example, developed a comprehensive taxonomy of translation strategies that consists of syntactic, semantic, and pragmatic categories, and each includes ten strategies. Taxonomies of translation errors have also been developed and are widely used in practice, such as Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). Although these existing frameworks may be useful as points of departure, detailed ones dedicated to TS tasks are still lacking.

3 Framework of Analytical Evaluation

We developed taxonomies of simplification strategies and errors as the analytical evaluation framework for TS. Simplification strategies consist of two independent components: **surface strategies**, which capture superficial operations for grammatical or textual elements, and **content strategies**, which capture semantic or content changes from the viewpoint of simplification. In this framework,

[1]The under-reporting of error analysis is a general problem in NLG literature (van Miltenburg et al., 2021).

each TS instance (i.e., a minimally decomposed editing operation) is first judged as an error category listed in the error taxonomy. If it is not, the instance is then independently labeled a surface and content strategy (see also Figure 1). Our framework includes guidelines for annotating surface and content strategies in the form of a decision tree.[2]

3.1 Taxonomy Construction

Specifically referring to Alva-Manchego et al. (2019b), Chesterman (2016), and Shardlow (2014), we created taxonomies of the simplification strategies and errors through an analysis of human and machine TS instances. As manual simplification data, we used original and simplified news articles from Newsela,[3] which were produced by professional editors and are expected to include various types of editing operations, including creative ones. We selected four articles from Newsela's Popular category. Each article has four simplified versions with different degrees of simplicity, from Lv0 (the original document) to Lv4 (the simplest document). We manually aligned sentences from all adjacent-level documents (e.g., Lv0–Lv1, Lv1–Lv2) and acquired 551 complex–simplified pairs that exhibited any sort of rewriting.[4] Next, we decomposed the rewriting from complex to simplified sentences into minimum edits (see Figure 2).[5] Consequently, we acquired 1,133 minimum editing instances of human simplification.

First, using these instances, we created prototype taxonomies of the simplification strategies in the bottom-up procedures: (i) for each instance, we devised labels for describing surface and content strategies; and (ii) we aggregated and revised the labels to form systematic taxonomies. Edit (3) in Figure 2 is an example of a minimum edit instance: "hard work will help you reach your goals" → "hard work is important". The same editing operation is annotated differently with the surface strategy ("Replace at sentence level") and content strategy ("Paraphrase into a direct expression").

Second, using simplified instances generated by TS systems, we expanded and modified the prototype taxonomies to improve their applicability. At this stage, we created an error taxonomy by analyzing the system errors.[6] We used three neural TS models, Transformer (Jiang et al., 2020), DRESS (Zhang and Lapata, 2017), and SUC (Sun et al., 2020) that were trained on Newsela data.[7] From the same Newsela's articles, we selected 166, 38, and 38 sentences for Transformer, DRESS, and SUC, respectively, to generate their simplified versions.[8] We decomposed 125 outputs that exhibited any sort of rewriting to acquire 217 minimum edits. We then separated the non-error and error instances. Using the prototype taxonomies of the surface and content strategies, we classified non-error instances and, if necessary, modified the taxonomies to properly cover all instances. We also created an error taxonomy by analyzing the error instances.

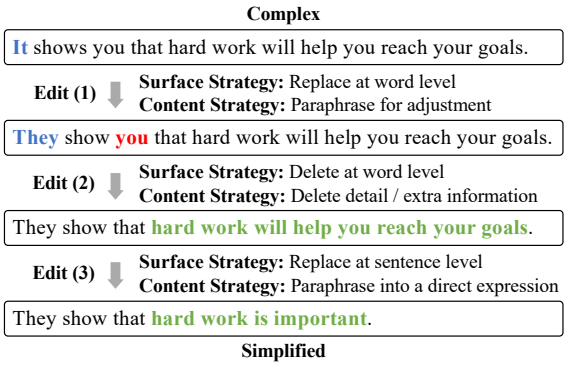

Figure 2: Example of the decomposition of rewriting instances and labeling of strategies.

3.2 Taxonomy of Surface Strategies

We defined 22 surface strategies, S1–S22, under the seven general categories: **Replacement**, **Deletion**, **Addition**, **Integration**, **Splitting**, **Move**, and **No change**.[9] **Replacement**, **Deletion**, and **Addition** have the same set of linguistic focuses, i.e., punctuations, words, phrases, clauses, and sentences.

Note that, if the head of a phrase changes (e.g., "playing the video games" → "the video games), it is classified into the **Replacement** category. If the head of a phrase is retained (e.g., "the video games" → "the games"), it is classified into the **Deletion** rather than **Replacement** category.

[2]The decision trees are shown in Appendix A.2.

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

[4]These pairs included the sentences that were not aligned because we considered such sentences as instances of addition or deletion of a sentence.

[5]Following Miyata and Fujita (2021), we defined a minimum edit as "a small edit that is difficult to be further decomposed into more than one independent edit" and that does not induce "ungrammaticality in the edited sentence" (p. 1541).

[6]In principle, simplification instances by professional human editors seldom include errors.

[7]We explain the detailed implementation in §4.1.

[8]We first used the same set of 38 sentences, but we found out that the outputs of DRESS and SUC consisted of many error instances. To collect a wide range of non-error instances, we added another 128 sentences only for Transformer.

[9]See Table 2 for the detailed surface strategies and Appendix B for the example sentences.

3.3 Taxonomy of Content Strategies

We defined 30 content strategies, C1–C30, under five general categories: **No content change**, **Content deletion**, **Content addition**, **Content change**, and **Document-level adjustment**.[10]

Note that while the content strategies in **No content change**, except for **(C5) Remain unchanged**, change the surface structure or textual element, they do not change the propositional meaning of the sentence. **Document-level adjustment** includes the change of the sentence order in a document and the secondary edits that need to be performed due to changes made to a different sentence; for example, some lexical changes might entail changing the pronouns in the later sentences.

3.4 Error Taxonomy

We defined four error categories: **Inappropriate deletion**, **Inappropriate addition**, **Inappropriate paraphrase**, and **Non-sentence**.[11] The first three categories roughly correspond to **Deletion**, **Addition**, and **Replacement** in the surface strategies. **Non-sentence** covers other error types that make the sentence ungrammatical or unintelligible.

4 Experimental Setup

To clarify the potential and limitations of current TS systems in comparison with human performance, we designed an experiment to annotate TS instances produced by human editors and recent neural TS systems using our taxonomies of simplification strategies and errors described in §3. We also conducted a human evaluation to better understand the general tendencies of how strategies and errors affect the TS quality.

4.1 Neural Text Simplification Systems

We implemented six systems, that is, three neural models below trained separately with Newsela (in-domain setting) or Wikipedia (out-of-domain setting).[12] It should be noted that the training data size and pre-processing methods differed depending on the models, as we aimed to replicate the models described in the original papers as much as possible.

Transformer (Jiang et al., 2020)[13] This BERT-initialized Transformer model is a state-of-the-art model. We used the Newsela and Wikipedia models distributed by the authors.

DRESS (Zhang and Lapata, 2017)[14] This model exploits reinforcement learning, which rewards rewriting. Many studies have used this as a baseline (e.g., Vu et al., 2018; Nassar et al., 2019; Omelianchuk et al., 2021). To train the Newsela model, we used newsela_data_share-20150302 from the Newsela corpus, excluding Lv0–Lv1, Lv1–Lv2, and Lv2–Lv3 pairs, following the original paper. We also excluded sentences that were more than 85 words per sentence or included "/" because the original code could not process them. The remaining 94,635 sentences[15] were used for training after the named entities were tagged with Stanford CoreNLP (Manning et al., 2014).[16] We used processed Wikilarge to train the Wikipedia model.

SUC (Sun et al., 2020)[17] This model uses one target sentence and two preceding and following sentences as input. Only this model exploits the context among these three models. Because Sun et al. (2020) did not provide a Newsela model, we trained it using Newsela-Auto, the same dataset used in the Transformer model above. Excluding Lv0–Lv1, Lv1–Lv2, and Lv2–Lv3 pairs following Zhang and Lapata (2017), we used the 640,867 sentences with context and 173,105 sentences without context. To train the Wikipedia model, we used the first 116,020 sentences with context and all of the 40,893 sentences without context from the distributed dataset. We created the vocabularies for Newsela and Wikipedia models, respectively, from the training data using spaCy (Honnibal and Montani, 2017).[18]

4.2 Annotation of Strategies and Errors

As evaluation data, from three original Newsela articles (Lv0) in the Popular category, we respectively extracted 13, 11, and 22 sequential sentences while retaining the textual cohesion. For these 46

[10] See Table 4 for the detailed content strategies and Appendix B for the example sentences.

[11] See Appendix B for the example sentences.

[12] We calculated scores of automatic evaluation metrics and verified that we had appropriately reproduced the implementations reported in the original papers. See Appendix C.1 for details on the automatic evaluation.

[13] https://github.com/chaojiang06/wiki-auto

[14] https://github.com/XingxingZhang/dress

[15] Zhang and Lapata (2017) reported using 94,208 sentences. Although we processed the corpus in the same manner, we could not obtain the same number of sentences.

[16] https://github.com/stanfordnlp/CoreNLP

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

[18] https://github.com/explosion/spaCy

Score	Grammaticality (G)	Meaning preservation (M)	Simplicity (S)
5	Native speaker level fluent	Adequately preserved	Much simpler
4	Non-native speaker level fluent	Mostly preseved	Simpler
3	Understandable	Partially preserved	The same simplicity
2	Partially understandable	Completely different	More difficult
1	Completely unintelligible	Unintelligible	Unintelligible

Table 1: Abridged guidelines for human evaluations. The full version is shown in Appendix A.1.

complex sentences, we extracted 54 corresponding simplified sentences from Newsela's articles (Lv1) as human references[19] and generated 276 simplified sentences (46 sentences × 6 systems) as system outputs. We decomposed 39 and 191 sentences that exhibited any sort of rewriting to acquire 105 and 389 minimum edits, respectively, for human references and system outputs.[20]

Each editing instance was annotated with strategies and error categories based on the classification procedures explained in §3. We counted the sentences that were not rewritten as instances of strategy. The annotation was carried out independently by the first and third authors, who can adequately understand the English text and have a good command of the analytical evaluation framework, i.e., the taxonomies and guidelines. The inter-annotator agreement scores (Cohen's unweighted kappa) for the surface strategies, content strategies, and errors were 0.806, 0.745, and 0.851, respectively, indicating substantial agreement (Landis and Koch, 1977).[21] After the independent annotation, the annotators resolved any disagreement in judgments through discussions to obtain the final labels.

4.3 Human Evaluation

Using the sentence data used in §4.2, we also conducted a subjective human evaluation to assess the grammaticality (G), meaning preservation (M), and simplicity (S) of the simplified sentences generated by the six systems.

The annotators were two professional translators who were familiar with Japanese–English translation, English proofreading, and native language checking. They assigned a score to each sentence using a five-point Likert scale by referring to the evaluation guidelines, an abridged version of which is shown in Table 1.[22] Before commencement of the formal evaluation, they evaluated another 29 sentences as a practice to properly understand the task. They evaluated the same set of sentences that exhibited any sort of rewriting. We consistently gave scores of 5, 5, and 3 for G, M, and S, respectively, to the non-rewritten sentences. The inter-annotator agreement scores (Cohen's quadratic weighted kappa) for G, M, and S were 0.541, 0.257, and 0.628, respectively.[23]

5 Results and Discussions

5.1 Surface Strategies

Table 2 lists the annotation results for the surface strategies with human evaluation scores for the system outputs.[24] Note that for each strategy, the human evaluation score was calculated using *sentences that exhibit the strategy*. As single sentences may include multiple strategies, the scores may be influenced by other strategies. Nevertheless, the general impact of each strategy can be inferred.

All the systems performed **Replacement** less frequently than humans did. The systems chiefly performed **(S2) Replace at word level** and could not perform **(S4) Replace at clause level** or **(S5) Replace at sentence level**, whereas humans performed **Replacement** strategies at various linguistic levels. This indicates the incapability of current models to learn replacement operations for linguistic units larger than phrases.

Deletion was the dominant strategy for the systems; the Transformer systems and in-domain DRESS system performed **Deletion** more frequently than humans. Human evaluation scores suggest the trade-off between meaning preservation and simplicity according to the size of the linguistic unit that is deleted; the deletion of a larger

[19] Because human references include the instances of sentence addition and splitting, the number of simplified sentences is larger than that of complex sentences.

[20] This means that the average rewriting rate for human editors was 2.69 times per sentence and that of systems was 2.03 times.

[21] When calculating the agreement scores for the strategies, we aggregated the annotations for the errors into one class and vice versa. The detailed distributions of annotations are presented in Appendix D.

[22] The detailed guidelines are presented in Appendix A.1.

[23] When calculating the inter-annotator agreement scores, we excluded the non-rewritten sentences. If we include them, the scores for G, M, and S rise to 0.618, 0.433, and 0.725, respectively.

[24] The overall results of the human evaluation are presented in Appendix C.2.

Surface strategy	Human ref.	Number of annotated instances: Transformer IND	Transformer OOD	DRESS IND	DRESS OOD	SUC IND	SUC OOD	Human evaluation G	Human evaluation M	Human evaluation S
Replacement	29	20	12	19	12	12	0			
(S1) Replace at punctuation level	(3)	(3)	(0)	(1)	(0)	(0)	(0)	4.63	3.25	3.88
(S2) Replace at word level	(4)	(10)	(5)	(13)	(11)	(10)	(0)	3.80	3.72	3.14
(S3) Replace at phrase level	(11)	(7)	(6)	(5)	(1)	(2)	(0)	4.26	3.81	3.76
(S4) Replace at clause level	(4)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(S5) Replace at sentence level	(7)	(0)	(1)	(0)	(0)	(0)	(0)	5.00	4.50	4.50
Deletion	30	39	35	32	17	16	0			
(S6) Delete at punctuation level	(4)	(0)	(3)	(1)	(0)	(0)	(0)	3.38	3.50	3.00
(S7) Delete at word level	(6)	(5)	(10)	(5)	(2)	(7)	(0)	3.67	3.31	3.29
(S8) Delete at phrase level	(12)	(16)	(10)	(10)	(2)	(5)	(0)	3.84	3.29	3.78
(S9) Delete at clause level	(3)	(18)	(12)	(16)	(13)	(4)	(0)	3.91	3.12	3.91
(S10) Delete at sentence level	(5)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
Addition	20	1	0	0	0	1	0			
(S11) Add at punctuation level	(0)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(S12) Add at word level	(3)	(1)	(0)	(0)	(0)	(1)	(0)	3.75	4.00	3.50
(S13) Add at phrase level	(8)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(S14) Add at clause level	(1)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(S15) Add at sentence level	(8)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
Integration	0	0	0	0	0	0	0			
(S16) Integrate two sentences	(0)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(S17) Integrate more than two sentences	(0)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
Splitting	3	0	0	0	0	0	0			
(S18) Split by phrase	(0)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(S19) Split by clause	(3)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
Move	8	0	1	0	0	1	0			
(S20) Move constituents	(4)	(0)	(1)	(0)	(0)	(1)	(0)	3.50	2.50	2.25
(S21) Move a sentence	(4)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
No transformation	15	4	16	8	23	16	18			
(S22) Use an identical sentence	(15)	(4)	(16)	(8)	(23)	(16)	(18)	5.00	5.00	3.00
Total	105	64	64	59	52	46	18			
Precision		0.313	0.297	0.288	0.327	0.283	0.278			
Recall		0.190	0.181	0.162	0.162	0.124	0.048			

Table 2: Number of annotated instances for the surface strategies. Three TS systems are trained with in-domain Newsela data (IND) and out-of-domain Wikipedia data (OOD). Human evaluation scores are the averaged scores for system outputs that involve each strategy (G: grammaticality; M: meaning preservation; S: simplicity).

unit can increase the simplicity score, but decrease the meaning preservation score. It is also notable that the number of **(S9) Delete at clause level** performed by the systems was much larger than that performed by humans. This is attributable to the structure of the training data, which are aligned at the single-sentence level. Consider the case in which the complex sentence "I bought an apple and ate it" is split into two sentences, "I bought an apple" and "I ate it". In the current research practices for preparing training data, the two simplified sentences are separately aligned with the complex sentence. This would induce the systems to excessively learn large deletions, such as Examples 1 in Table 3. It is also important to note that none of the systems performed **(S10) Delete at sentence level**. This may be because the training data did not include instances of sentence deletion, as the alignment of such cases is difficult.

The systems seldom performed **Addition**, whereas humans performed this 20 times at various linguistic levels. Although instances of addition were included in the training data, even word- and phrase-level addition strategies were hardly observed. This implies the fundamental difficulties of addition operations for the current models and training data.

The systems used in this study cannot learn **Integration** (**S16** and **S17**)[25], **Splitting** (**S18** and **S19**), and **(S21) Move a sentence** because of the aforementioned data structure problem. Although some end-to-end systems can perform **Splitting** and **Integration** (Scarton and Specia, 2018), these supra-sentential operations remain to be fully achieved in the neural TS research. To address the **Splitting** operation, we can refer to the rich accumulation of linguistically motivated studies on syntactic simplification (Scarton et al., 2017).

The final two rows in Table 2 show the overall

[25]Neither did the humans perform Integration in the evaluation dataset. We observed six instances of Integration in the 1,133 instances used for taxonomy creation.

#	Model	Text
1	Input	But when schools start later, teens get to class on time and find it easier to stay awake, a new study finds.
	Transformer (IND)	but when schools start later , teens get to class on time .
	DRESS (IND)	But when schools start later , teens get to class on time .
	Human reference	A new study finds that when schools start later, teens get to class on time. They also find it easier to stay awake.
2	Input	Not everyone can become a genius or a star athlete, but they can improve the skills they have and develop new ones.
	Transformer (IND)	not everyone can be a genius or a star athlete .
3	Input	Knowing this, schools in several districts have begun to shift their start times.
	Human reference	Knowing this about teens, schools in several districts have begun to shift their start times.
4	Input	Information from millions of cones reaches our brains as electrical signals that communicate all the types of light reflected by what we see, which is then interpreted as different shades of color.
	Human reference	The cones then send information to our brains, which interprets the light we see as different colors.

Table 3: Examples of system outputs and references.

precision and recall of adopted strategies by the systems in comparison with humans' strategies for the same sentences. The precision scores are about 0.2–0.3 and the recall scores are all below 0.2, which means that humans and systems tend to adopt different strategies even for the same sentence. Although further investigations are needed to draw insights from these results, we should be aware of the substantial differences between humans and machines in terms of simplification operations.[26]

5.2 Content Strategies

Table 4 lists the annotation results for the content strategies. While humans performed **(C1) Transform syntactic structure** nine times, the systems rarely did. Most **C1** cases by humans involved sentence splitting, and as previously mentioned, the systems could not learn this operation from the current training data.

The systems generally performed various **Content deletion** strategies. It is worth noting that **(C10) Delete important information**, a large deletion corresponding to **(S9) Delete at clause level** in surface strategies, was performed frequently. Example 2 in Table 3 illustrates the deletion of the latter clause. Although the output can be regarded as a simplified version of the input at the sentence level, this deletion might be inappropriate in terms of logical flow in the entire document. In this sense, these categories might be regarded as **(E1) Inappropriate deletion** in the error taxonomy. Indeed, humans did not adopt this strategy.

The systems did not perform any **Content addition**, which corresponds to the lack of **Addition** of the surface strategies. These strategies require contextual information in many cases like "this" → "this about teens" (See Example 3 in Table 3). However, even the context-aware SUC models cannot perform these addition operations.

As for **Content change**, the Transformer and DRESS performed **(C20) Paraphrase into a similar phrase** more than humans. The neural systems generally have abilities to perform local rewriting like "become" → "be" (see Example 2 in Table 3). Similarly, **(C25) Paraphrase into an essential point**, which substantially concerns deleting or altering local elements like "color production" → "color", was performed well by the in-domain systems. By contrast, the systems cannot perform **(C21) Paraphrase into an explanatory expression** and **(C24) Paraphrase into a concrete expression**. The former requires external or contextual knowledge to add information, such as "the belief that your intelligence can grow" → "which means he believes that your intelligence can grow". The latter requires word sense disambiguation or anaphora resolution to explicitly indicate the hidden meaning, such as "ones" → "friendships". In general, current systems have limitations in performing these sophisticated **Content change** operations, such as Example 4 in Table 3.

The systems hardly performed **Document-level adjustment**. **(C27) Change information flow** corresponds to **(S21) Move a sentence** at surface level, which is architecturally impossible for the systems used in this study. The other strategies, C28–C30, depend on the results of other operations in the document, and are fundamentally difficult for current systems that do not exploit the output-side context.

5.3 Errors

Table 5 lists the annotation results for the simplification errors. As mentioned in §5.2, the number of **(E1) Inappropriate deletion** can increase if we consider **(C10) Delete important information** as an error. The instances of **(E2) Inappropriate**

[26] These differences might be attributed not only to the inability of systems to replicate human performance, but also to the nature of TS tasks. Examining the differences between human editors would be an important future task.

Content strategy	Human ref.	Number of annotated instances: Transformer IND	Transformer OOD	DRESS IND	DRESS OOD	SUC IND	SUC OOD	Human evaluation G	Human evaluation M	Human evaluation S
No content change	31	7	24	11	24	17	18			
(C1) Transform syntactic structure	(9)	(0)	(1)	(0)	(0)	(1)	(0)	3.50	2.50	2.25
(C2) Paraphrase into an abbreviation	(0)	(0)	(0)	(1)	(0)	(0)	(0)	3.00	3.50	4.00
(C3) Paraphrase into a non-abbreviation	(1)	(0)	(2)	(0)	(0)	(0)	(0)	3.50	3.25	2.25
(C4) Paraphrase into standard form	(6)	(3)	(5)	(2)	(1)	(0)	(0)	4.23	3.73	3.55
(C5) Remain unchanged	(15)	(4)	(16)	(8)	(23)	(16)	(18)	5.00	5.00	3.00
Content deletion	24	37	32	31	16	14	0			
(C6) Delete introduction / conclusion	(2)	(1)	(0)	(1)	(0)	(1)	(0)	2.83	2.83	4.00
(C7) Delete a parallel element	(1)	(5)	(1)	(6)	(0)	(1)	(0)	3.50	3.27	3.69
(C8) Delete information for cohesion	(5)	(6)	(6)	(7)	(3)	(3)	(0)	3.94	3.30	3.64
(C9) Delete a modifier	(9)	(6)	(10)	(4)	(2)	(5)	(0)	3.94	3.37	3.46
(C10) Delete important information	(0)	(5)	(6)	(4)	(1)	(1)	(0)	3.91	3.06	3.97
(C11) Delete detail / extra information	(7)	(14)	(9)	(9)	(10)	(3)	(0)	3.93	3.19	3.92
Content addition	17	0	0	0	0	0	0			
(C12) Add introduction / conclusion	(0)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(C13) Add a parallel element	(2)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(C14) Add contextual information	(0)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(C15) Add information for cohesion	(1)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(C16) Add a modifier	(4)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(C17) Add detail / extra information	(10)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
Content change	22	19	8	17	11	15	0			
(C18) Change aspect	(1)	(2)	(0)	(0)	(0)	(1)	(0)	3.67	3.83	3.67
(C19) Change modality	(0)	(1)	(1)	(0)	(0)	(2)	(0)	3.00	3.25	2.50
(C20) Paraphrase into a similar phrase	(2)	(4)	(3)	(7)	(5)	(1)	(0)	3.75	3.60	3.18
(C21) Paraphrase into an explanatory expression	(4)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(C22) Paraphrase into a direct expression	(6)	(3)	(0)	(1)	(2)	(2)	(0)	3.88	3.56	3.44
(C23) Paraphrase into a brief expression	(1)	(1)	(1)	(1)	(0)	(0)	(0)	4.00	3.33	4.33
(C24) Paraphrase into a concrete expression	(1)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(C25) Paraphrase into an essential point	(4)	(6)	(1)	(7)	(2)	(5)	(0)	4.02	3.81	3.55
(C26) Paraphrase into a different view	(3)	(2)	(2)	(1)	(2)	(4)	(0)	4.09	3.95	3.00
Document-level adjustment	11	1	0	0	1	0	0			
(C27) Change information flow	(4)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(C28) Delete for adjustment	(2)	(1)	(0)	(0)	(1)	(0)	(0)	4.25	3.25	4.00
(C29) Add for adjustment	(2)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
(C30) Paraphrase for adjustment	(3)	(0)	(0)	(0)	(0)	(0)	(0)	-	-	-
Total	105	64	64	59	52	46	18			
Precision		0.219	0.250	0.237	0.308	0.239	0.278			
Recall		0.133	0.152	0.133	0.152	0.105	0.048			

Table 4: Number of annotated instances for the content strategies.

addition were also observed. In particular, SUC produced many such instances. Considering the observation that almost no instance was annotated as **Addition** in Table 2, what the systems added to output sentences were not judged as (successful) strategies but as errors. **(E3) Inappropriate paraphrase** is the most frequent error type for most systems, which includes, for example, "intelligence" (being intellectual) → "spy" and "cells called neurons" → "DNA". These errors are problematic because incorrect information can be conveyed to readers without being noticed as errors.

For Transformer and DRESS, the in-domain systems trained on Newsela generally produced more errors than the out-of-domain systems trained on Wikipedia. Considering the fewer number of **No transformation** cases of in-domain systems (see Table 2), in-domain systems tended to be more aggressive but erroneous than out-of-domain systems.

For all the human evaluation scores, except for the meaning preservation score for **(E2) Inappropriate addition**, the averaged scores are below 3. This indicates that inclusion of any error can lead to an unacceptable output sentence.

6 Conclusions and Outlook

To better advance TS research and practice, in this study, we conducted an analytical evaluation of current neural TS systems and showed their potential and limitations in comparison with human performance. Using our proposed evaluation framework consisting of taxonomies of surface strategies, content strategies, and errors, we annotated both the human references and outputs of six systems (three models trained on in-domain and out-of-domain datasets). The results demonstrated that, while current TS systems can perform deletions and local substitutions, their performance is far behind human parity, owing to the following limitations:

Error category	Number of annotated instances: Transformer IND	Transformer OOD	DRESS IND	DRESS OOD	SUC IND	SUC OOD	Human evaluation G	Human evaluation M	Human evaluation S
(E1) Inappropriate deletion	6	2	7	0	9	1	2.62	2.82	2.80
(E2) Inappropriate addition	6	2	4	4	12	51	2.68	3.86	2.11
(E3) Inappropriate paraphrase	17	5	28	14	15	0	2.83	2.53	2.76
(E4) Non-sentence	0	0	0	1	3	0	1.00	1.25	1.00
Total	29	9	39	19	39	52			

Table 5: Number of annotated instances for the error categories.

- The systems have difficulties in substituting a linguistic unit larger than a phrase, including sentence splitting.
- Excessive deletion of clause-level important information has occurred frequently.
- The systems tried to perform addition operations; however, they always failed to produce correct results.

Our analytical evaluation also suggests detailed paths to overcome these issues. For example, in addition to improving the capacity of end-to-end neural models, utilizing technologies tailored to particular operations such as sentence splitting and explanation generation can be helpful. To mitigate the excessive deletion, it would be effective to refine the alignment methods. Exploiting document-level contexts on both input and output sides and/or document-external knowledge is a necessary task for successful content addition.

Limitations

Applicability. The primary limitation in our study is that we chiefly used the Newsela dataset to build the annotation framework, i.e., the taxonomies and decision trees, and conduct the analytical evaluation. While we assume that the Newsela dataset includes diverse simplification operations as mentioned in §3.1, the applicability of our framework to other domains or datasets, such as Simple Wikipedia, needs to be investigated.[27]

The diversity of adopted TS systems is also limited. As the aim of this pilot study is to demonstrate the usefulness of analytical evaluation, we mainly selected orthodox baseline models. To further improve the applicability, it is important to examine other types of TS models, such as controllable models (e.g., Maddela et al., 2021; Nishihara et al., 2019; Scarton and Specia, 2018) and edit-based models (e.g., Dong et al., 2019; Stahlberg and Kumar, 2020). Further investigation of various document-level models other than SUC used in this study will also be needed (Sun et al., 2021).

Although our taxonomies are mostly language independent, the forms of decision trees for strategy annotation may need to be changed depending on the language because the decision order was defined based on the degree of difficulty in identifying the strategies, which might be language dependent.[28]

Feasibility. The annotation of simplification strategies and errors was conducted by the authors, who were involved in the development of the annotation framework. Although the authors independently conducted the annotation task and substantial inter-annotator agreement was achieved, the feasibility of annotation by those outside this study has not been examined. To improve the feasibility, more detailed instructions and a sufficient training session may be needed. Although sharing the annotated data would be beneficial for the feasibility, it is difficult due to copyright issues.

Acknowledgments

We are grateful to Newsela for sharing the data. This work was supported by JSPS KAKENHI Grant Number 19H05660 and by the Research Grant Program of KDDI Foundation, Japan.

References

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys*, 54(2):1–36.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019a. EASSE: Easier automatic sentence simplification evaluation. 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): System Demonstrations*, pages 49–54, Hong Kong, China.

[27]The characteristics of Simple Wikipedia as a TS data resource have been extensively discussed (Xu et al., 2015).

[28]The details of decision trees are presented in Appendix A.2

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

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

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

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2014. Defining an annotation scheme with a view to automatic text simplification. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it)*, pages 87–92, Pisa, Italy.

Andrew Chesterman. 2016. *Memes of translation: The spread of ideas in translation theory*, 2nd edition. Amsterdam: John Benjamins.

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 (ACL)*, pages 3393–3402, Florence, Italy.

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 (ACL-IJCNLP)*, pages 63–68, Beijing, China.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7943–7960, Online.

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, Institute for Simulation and Training, University of Central Florida.

Anaïs Koptient, Rémi Cardon, and Natalia Grabar. 2019. Simplification-induced transformations: Typology and some characteristics. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 309–318, Florence, Italy.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:455–463.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3536–3553, Online.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 55–60, Baltimore, Maryland, USA.

Ruslan Mitkov and Sanja Štajner. 2014. The fewer, the better? A contrastive study about ways to simplify. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA)*, pages 30–40, Dublin, Ireland.

Rei Miyata and Atsushi Fujita. 2021. Understanding pre-editing for black-box neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1539–1550, Online.

Lucía Molina and Amparo Hurtado Albir. 2002. Translation techniques revisited: A dynamic and functionalist approach. *Meta*, 47(4):498–512.

Islam Nassar, Michelle Ananda-Rajah, and Gholamreza Haffari. 2019. Neural versus non-neural text simplification: A case study. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association (ALTA)*, pages 172–177, Sydney, Australia.

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 (ACL SRW)*, pages 260–266, Florence, Italy.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 11–25, Online.

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 (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. 2017. MUSST: A multilingual syntactic simplification tool. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP): System Demonstrations*, pages 25–28, Taipei, Taiwan.

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 (ACL)*, pages 712–718, Melbourne, Australia.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing 2014*, 4(1):58–70.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2004. Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 896–902, Geneva, Switzerland.

Felix Stahlberg and Shankar Kumar. 2020. Seq2Edits: Sequence transduction using span-level edit operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159, Online.

Sanja Štajner and Sergiu Nisioi. 2018. A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 3026–3033, Miyazaki, Japan.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 230–242, Riga, Latvia.

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 (EMNLP)*, pages 738–744, Brussels, Belgium.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 162–173, Melbourne, Australia.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7997–8013, Online and Punta Cana, Dominican Republic.

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 (COLING)*, pages 1411–1423, Barcelona, Spain (Online).

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)*, pages 1–14, Online.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation (INLG)*, pages 140–153, Aberdeen, Scotland, UK.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais*. Didier, Paris, trans. and ed. by J. C. Sager & M.-J. Hamel (1995) as *Comparative Stylistics of French and English: A Methodology for Translation.* John Benjamins, Amsterdam.

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 (NAACL-HLT)*, pages 79–85, New Orleans, Louisiana, USA.

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 (TACL)*, 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 (TACL)*, 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 (EMNLP)*, pages 584–594, Copenhagen, Denmark.

Score	Grammaticality/Fluency (G)	Meaning preservation/Adequacy (M)	Simplicity (S)
5	The target sentence is fluent (native speaker level) and grammatically correct.	The target sentence adequately conveys the core meaning of the original sentence.	The target sentence is much simpler than the original sentence.
4	The target sentence is almost fluent (non-native speaker level) and grammatically correct	The target sentence mostly conveys the core meaning of the original sentence.	The target sentence is simpler than the original sentence.
3	The target sentence is less fluent with some ungrammatical parts, but understandable	The core meaning of the original text is not conveyed, but the information of the the original text is partially preserved.	The target sentence is as simple/difficult as the original sentence.
2	The target sentence is ungrammatical, but partially understandable.	The meaning of the target sentence is completely different from that of the original sentence.	The target sentence is more difficult than the original sentence.
1	The target sentence is completely unintelligible.	It is impossible to assess the meaning of the target sentence because of its unintelligibility.	It is impossible to assess the simplicity of the target sentence because of its unintelligibility.

Table 6: Guidelines for human evaluation.

A Guidelines

A.1 Guidelines for Human Evaluation

Table 6 lists the guidelines for human evaluations. We instructed annotators to consider document-level coherence when evaluating each sentence. Additionally, we instructed them to give an S score of 1 to the sentence that was given an M score of 1 or 2.

A.2 Annotation Guidelines for Simplification Strategies

Figures 3 and 4 show the guidelines, i.e., decision trees, for the annotation of simplification strategies. The procedures to build a decision tree were as follows: (1) through the classification of sample instances by trial and error, the first author created the prototype decision tree in a way that easier decisions can be made in earlier stages; (2) the third author validated the prototype by classifying sample instances using it; (3) based on the feedback from the third author, the first author refined the prototype.

L1 represents the category of strategy, and L2 represents the strategy. Note that S# and C# in the figures do not indicate the strategy numbers. In the annotation task described in §4.2, we used the Japanese versions.

B Examples of Strategies and Errors

Tables 7 and 8 list examples of the surface and content strategies. Table 9 lists examples of errors. These sentences were extracted from the instances of human simplification,[29] which are based on Newsela articles (see §3.1 for detail).

C Additional Evaluation Results

C.1 Automatic Evaluation Scores

Table 10 shows the overall results of the automatic evaluation in terms of SARI, BLEU, and FKGL, all of which were measured by using EASSE (Alva-Manchego et al., 2019a)[30] at the corpus level.

For preparing the evaluation data, we manually aligned complex–simplest sentences for five Newsela articles. To properly implement SUC, we excluded sentences that do not have two preceding or following sentences and that consist of less than four words. We finally used 1,010 sentences for the automatic evaluation.

C.2 Overall Human Evaluation Scores

Table 11 shows the overall results of the human evaluation. The evaluation guidelines are presented in Appendix A.1.

[29] An exception is **(C10) Delete important information** in Table 8, the example of which was extracted from the outputs of Transformer.

[30] https://github.com/feralvam/easse

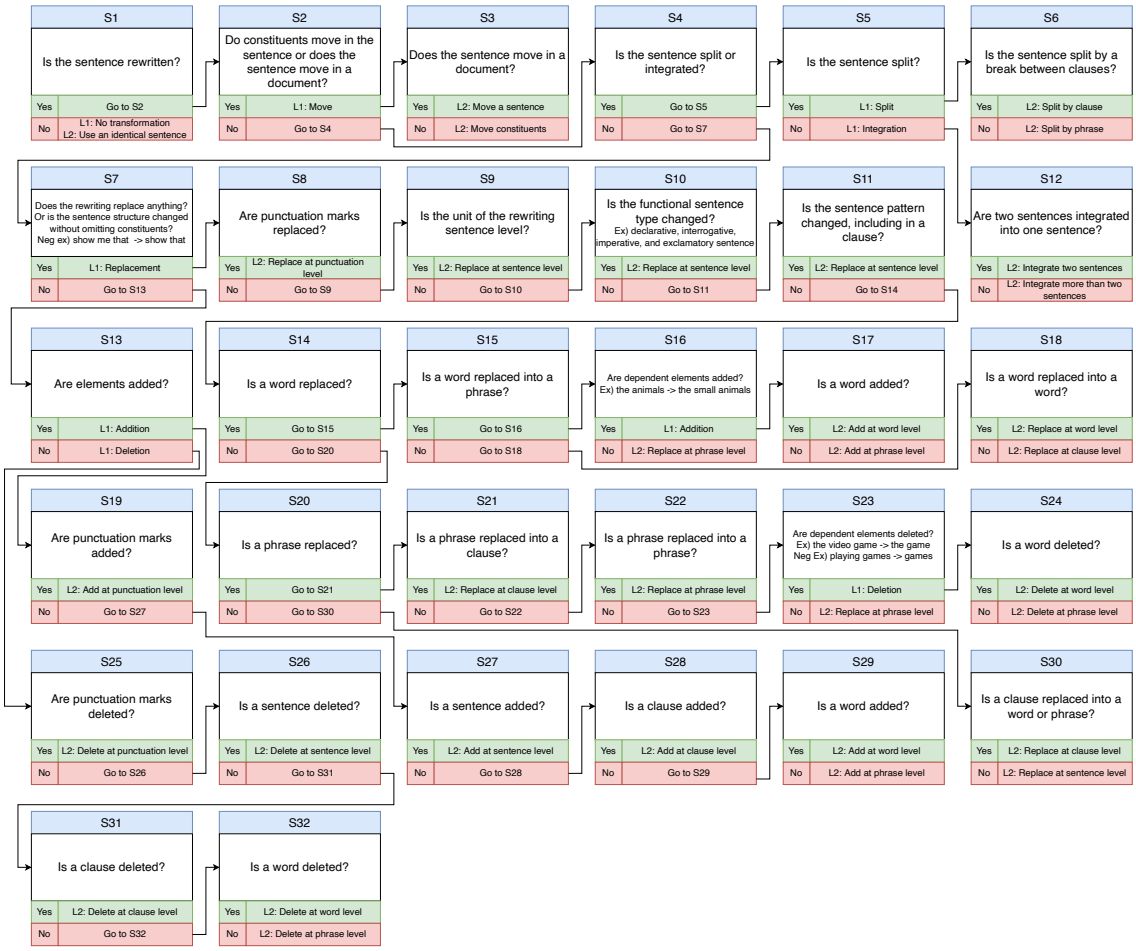

Figure 3: Annotation guidelines for surface strategies.

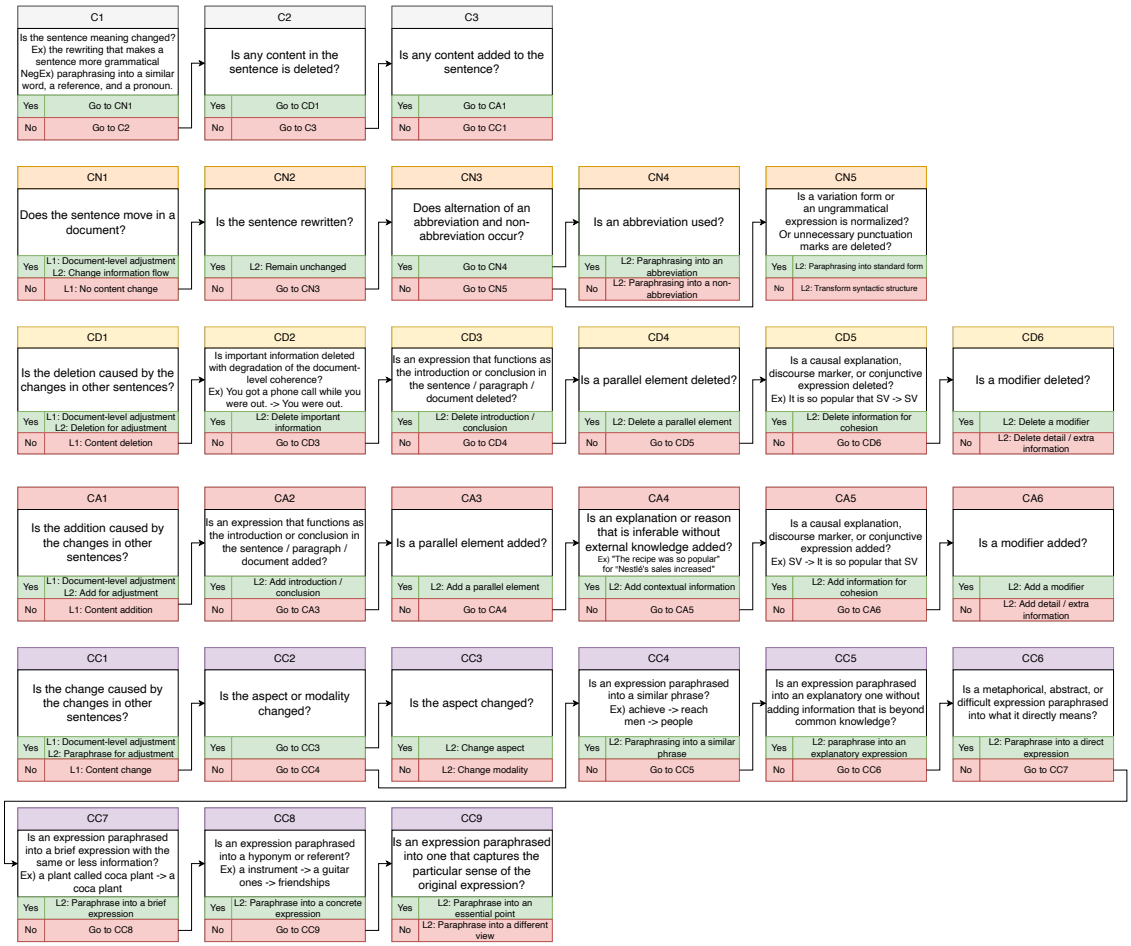

Figure 4: Annotation guidelines for content strategies.

Strategy		Example
Replacement		
(S1) Replace at punctuation level	Comp.	... 10,000 other neurons**!**
	Simp.	... 10,000 other neurons**.**
(S2) Replace at word level	Comp.	... make **better** surgeons.
	Simp.	... make **good** surgeons.
(S3) Replace at phrase level	Comp.	... about **playing video games** is friendship.
	Simp.	... about **video games** is friendship.
(S4) Replace at clause level	Comp.	The persistence **you use in games** ... .
	Simp.	The persistence **in games** ... .
(S5) Replace at sentence level	Comp.	**People who tried the syrup liked the taste**.
	Simp.	**People liked the taste of the syrup**.
Deletion		
(S6) Delete at punctuation level	Comp.	... **"**Toll House Chocolate Crunch Cookies.**"**
	Simp.	... Toll House Chocolate Crunch Cookies.
(S7) Delete at word level	Comp.	... inside and outside the **video** game.
	Simp.	... inside and outside the game.
(S8) Delete at phrase level	Comp.	Beating the final boss **or another really good player** ... .
	Simp.	Beating another really good player ... .
(S9) Delete at clause level	Comp.	He licked the ice **that was stuck around it**.
	Simp.	He licked the ice.
(S10) Delete at sentence level	Comp.	**So does saving a teammate when they're down.**
	Simp.	ϕ
Addition		
(S11) Add at punctuation level		*This strategy does not exist in our collected instances.*
(S12) Add at word level	Comp.	Remember games are ... .
	Simp.	Remember **that** games are ... .
(S13) Add at phrase level	Comp.	You have to be smart.
	Simp.	**In video games**, you have to be smart.
(S14) Add at clause level	Comp.	... — the monkeys, apes, and gorillas — ... .
	Simp.	... — the monkeys, apes, and gorillas **that are most like human** — ... .
(S15) Add at sentence level	Comp.	ϕ
	Simp.	S**cientists have studied video games.**
Integration		
(S16) Integrate two sentences	Comp.	Epperson pulled the stick. **He licked the frozen juice.**
	Simp.	Epperson pulled the stick **and licked the frozen juice**.
(S17) Integrate more than two sentences	Comp.	**We got masks. We got gloves. We got all those hand wipes. They're everywhere.**
	Simp.	**Now masks, gloves, hand wipes, and other material are everywhere**
Splitting		
(S18) Split by phrase	Comp.	**Think about how boring it can be to play an easy game**
	Simp.	**Think about playing an easy game. It can get boring.**
(S19) Split by clause	Comp.	Ruth Wakefield was an expert chef, **and the inn became famous for its desserts**.
	Simp.	Ruth Wakefield was an expert chef. **The inn became famous for its desserts.**
Move		
(S20) Move constituents	Comp.	It can **also** help fix broken ones.
	Simp.	It **also** can help fix broken ones.
(S21) Move a sentence		*The complex and simplified sentences are identical.*
No transformation		
(S22) Use an identical sentence		*The complex and simplified sentences are identical.*

Table 7: Examples of surface strategies (Comp.: Complex sentence; Simp.: Simplified sentence).

Strategy		Example
No content change		
(C1) Transform syntactic structure	Comp.	Some people think ... are waste of time **or bad for you**.
	Simp.	Some people think ... are a waste of time. **Some people think they are bad for you.**
(C2) Paraphrase into an abbreviated form	Comp.	... seem like **they've** been around forever.
	Simp.	... seem like **they have** been around forever.
(C3) Paraphrase into a non-abbreviated form	Comp.	Helping build ... .
	Simp.	Helping **to** build ... .
(C4) Paraphrase into a standard form	Comp.	**But** ... .
	Simp.	**However**, ...
(C5) Paraphrase into an identical sentence		*The complex and simplified sentences are identical.*
Content deletion		
(C6) Delete introduction / conclusion	Comp.	**Think about your favorite games.**
	Simp.	ϕ
(C7) Delete a parallel element	Comp.	... feel strong **and popular**.
	Simp.	... feel strong.
(C8) Delete information for cohesion	Comp.	**The treats were so popular that** Epperson started ... .
	Simp.	Epperson started ... .
(C9) Delete a modifier	Comp.	He licked the ice **that was stuck around it**.
	Simp.	He licked the ice.
(C10) Delete important information	Comp.	**Winkler teamed up with another scientist named Greg Bryant**, a professor ... .
	Simp.	He is a professor ... . [Transformer IND]
(C11) Delete detail / extra information	Comp.	... created **the semi-sweet morsel**, or chocolate chip.
	Simp.	... created chocolate chip.
Content Addition		
(C12) Add introduction / conclusion	Comp.	ϕ
	Simp.	**Chocolate chip cookies seem like they've been around forever.**
(C13) Add a parallel element	Comp.	... a different culture.
	Simp.	... a different culture **or speak a different language**.
(C14) Add contextual information	Comp.	ϕ
	Simp.	**The company was selling more and more chocolate bars.**
(C15) Add information for cohesion	Comp.	People can recognize it, even if ... .
	Simp.	**Laughter is so important to humans that** people can recognize it, even if ... .
(C16) Add a modifier	Comp.	It can help fix broken ones.
	Simp.	It can **also** help fix broken ones.
(C17) Add detail / extra information	Comp.	... to connect and bond.
	Simp.	... to connect and bond **with others**.
Content change		
(C18) Change aspect	Comp.	You might **do** ... .
	Simp.	You might **start doing** ... .
(C19) Change modality	Comp.	They **can** teach ... .
	Simp.	They teach ... .
(C20) Paraphrase into a similar phrase	Comp.	... Nestlé's sales **soared**.
	Simp.	... Nestlé's sales **increased**.
(C21) Paraphrase into an explanatory expression	Comp.	... to make **a headache medicine**.
	Simp.	... to make **a medicine to fix headaches**.
(C22) Paraphrase into a direct expression	Comp.	... to **shred** a guitar in real life.
	Simp.	... to **play** a guitar in real life.
(C23) Paraphrase into a brief expression	Comp.	... parts of **a plant called the coca plant**.
	Simp.	... parts of **the coca plant**.
(C24) Paraphrase into a concrete expression	Comp.	It also can fix broken **ones**.
	Simp.	It also can fix broken **friendships**.
(C25) Paraphrase into an essential point	Comp.	It is one of many **benefits** ... .
	Simp.	It is one of many **good things** ... .
(C26) Paraphrase into a different view	Comp.	The **cookies** became so popular ... .
	Simp.	The **recipe** became so popular ... .
Documet-level adjustment		
(C27) Change information flow		*The complex and simplified sentences are identical.*
(C28) Delete for adjustment	Comp.	**This makes you see that** solving problems can be fun.
	Simp.	Solving problems can be fun.
(C29) Add for adjustment	Comp.	You also have to be smart.
	Simp.	**In video games**, you also have to be smart.
(C30) Paraphrase for adjustment	Comp.	**It** shows you that ... .
	Simp.	**They** shows you that ... .

Table 8: Examples of content strategies (Comp.: Complex sentence; Simp.: Simplified sentence).

Error		Example
(E1) Inappropriate deletion	Input	When you think, feel, move, or use your senses, **signals travel through this network**.
	Output	When you think , feel , move , or use your senses . [DRESS IND]
(E2) Inappropriate addition	Input	It's how we tell friends that we find their joke funny, ...
	Output	it's how we tell friends that we find their joke funny **funny**, ... [Transformer IND]
(E3) Inappropriate paraphrase	Input	... but rats can make a very high-pitched **trill**.
	Output	... but rats can make a very high-pitched **noise**. [Transformer IND]
(E4) Non-sentence	Input	**The animals that laugh the most include primates like monkeys,** rats, and mammals that live in the ocean like dolphins.
	Output	**humans, on the other hand, like monkeys,** rats and mammals that live in the ocean like dolphins. [Transformer IND]

Table 9: Examples of errors.

	Transformer		DRESS		SUC	
	IND	OOD	IND	OOD	IND	OOD
SARI ↑	**37.57**	30.89	37.08	31.83	31.09	22.24
BLEU ↑	32.20	38.22	37.11	**39.29**	31.92	24.12
FKGL ↓	3.00	4.40	3.27	4.02	4.20	**2.61**

Table 10: Results of automatic evaluation. The upper/down arrow indicates that the higher/lower the score, the better the performance.

	Transformer		DRESS		SUC	
	IND	OOD	IND	OOD	IND	OOD
Grammaticality/Fluency	4.30	**4.60**	3.55	4.45	3.23	3.76
Meaning preservation/Adequacy	3.52	4.21	3.38	4.25	3.35	**4.64**
Simplicity	**3.74**	3.42	3.20	3.15	2.46	2.40

Table 11: Results of human evaluation using a five-point Likert scale.

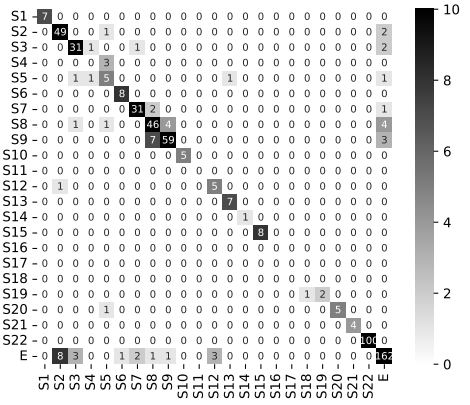

Figure 5: The distribution of annotations for surface strategies.

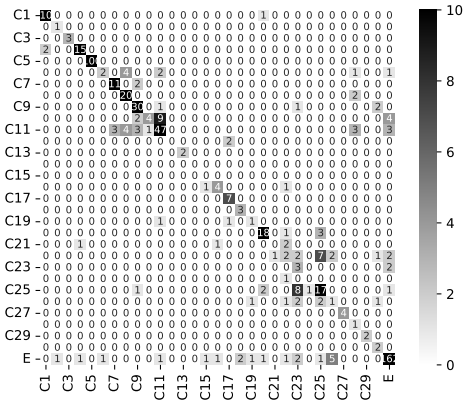

Figure 6: The distribution of annotations for content strategies.

D Distributions of Annotations

Figures 5, 6, and 7 show the distributions of annotations conducted by the two annotators in §4.2. S#, C#, and E# correspond to the surface strategy, content strategy, and error, respectively. When displaying the distributions for the strategies, we aggregated the annotations for the errors into one class and vice versa.

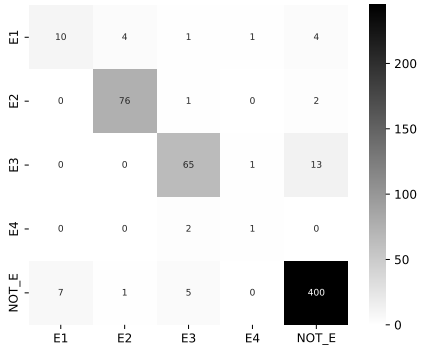

Figure 7: The distribution of annotations for errors.