Discourse Level Factors for Sentence Deletion in Text Simplification

Yang Zhong,[*1] **Chao Jiang,**[1] **Wei Xu,**[1] **Junyi Jessy Li**[2]
[1] Department of Computer Science and Engineering, The Ohio State University
[2] Department of Linguistics, The University of Texas at Austin
{zhong.536, jiang.1530, xu.1265}@osu.edu jessy@austin.utexas.edu

Abstract

This paper presents a data-driven study focusing on analyzing and predicting sentence deletion — a prevalent but understudied phenomenon in document simplification — on a large English text simplification corpus. We inspect various document and discourse factors associated with sentence deletion, using a new manually annotated sentence alignment corpus we collected. We reveal that professional editors utilize different strategies to meet readability standards of elementary and middle schools. To predict whether a sentence will be deleted during simplification to a certain level, we harness automatically aligned data to train a classification model. Evaluated on our manually annotated data, our best models reached F1 scores of 65.2 and 59.7 for this task at the levels of elementary and middle school, respectively. We find that discourse level factors contribute to the challenging task of predicting sentence deletion for simplification.

1 Introduction

Text simplification aims to rewrite an existing document to be accessible to a broader audience (e.g., non-native speakers, children, and individuals with language impairments) while remaining truthful in content. The simplification process involves a variety of operations, including lexical and syntactic transformations, summarization, removal of difficult content, and explicification (Siddharthan 2014).

While recent years saw a bloom in text simplification research (Xu et al. 2016; Narayan and Gardent 2016; Nisioi et al. 2017; Zhang and Lapata 2017; Vu et al. 2018; Sulem, Abend, and Rappoport 2018; Maddela and Xu 2018; Kriz et al. 2019) thanks to the development of large parallel corpora of original-to-simplified sentence pairs (Zhu, Bernhard, and Gurevych 2010; Xu, Callison-Burch, and Napoles 2015), most of the recent work is conducted at the sentence-level, i.e., transducing each complex sentence to its simplified version.

As a result, this line of work does not capture document-level phenomena, among which sentence deletion is the most prevalent, as simplified texts tend to be shorter (Petersen and Ostendorf 2007; Drndarevic and Saggion 2012; Woodsend and Lapata 2011).

This work aims to facilitate better understanding of sentence deletion in document-level text simplification. While prior work analyzed sentence position and content in a corpus study (Petersen and Ostendorf 2007), we hypothesize and show that sentence deletion is driven partially by contextual, discourse-level information, in addition to the content within a sentence.

We utilize the Newsela[2] corpus (Xu, Callison-Burch, and Napoles 2015) which contains multiple versions of a document rewritten by professional editors. This dataset allows us to compare sentence deletion strategies to meet different readability requirements. Unlike prior work that often uses automatically aligned data for analysis (c.f. Section 5), we manually aligned 50 articles of more than 5,000 sentences across three reading levels to provide a reliable ground truth for the analysis and model evaluation in this paper.[3] We find that sentence deletion happens very often at rates of 17.2%–44.8% across reading levels, indicating that sentence deletion prediction is an important task in text simplification. Several characteristics of the original document, including its length and topic, significantly influence sentence deletion. By analyzing the rhetorical structure (Mann and Thompson 1988) of the original articles, we show that the sentence deletion process is also informed by how a sentence is situated in terms of its connections to neighboring sentences, and its discourse salience within a document. In addition, we reveal that the use of discourse connectives within a sentence also influence whether it will be deleted.

To predict whether a sentence in the original article will be deleted during simplification, we utilize noisy supervision obtained from 886 automatically aligned articles with a total of 42,264 sentences. Our neural network model learns from both the content of the sentence itself, as well as the discourse level factors we analyzed. Evaluated on our manually annotated data, our best model that utilizes Gaussian-

* Work started as an undergraduate student at UT Austin.

[2] Newsela is an educational tech company that provides reading material for children in school curricula.

[3] To request our data, please first obtain access to the Newsela corpus at: https://newsela.com/data/, then contact the authors.

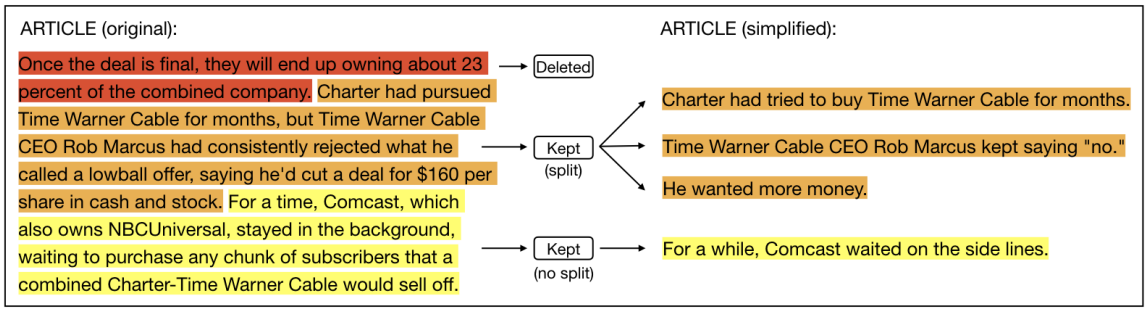

Figure 1: An example paragraph of the original news article (left) and its simplified version by professional editors for elementary school students (right). The arrows signify the sentence alignment between the original and simplified documents. The second sentence in the original paragraph is aligned to three sentences after simplification.

based feature vectorization achieved F1 scores of 65.2 and 59.7 for this task across two different reading levels (elementary and middle school). We show that several of the factors, especially document characteristics, complements sentence content in this challenging task. On the other hand, while our analysis of rhetorical structure revealed interesting insights, encoding them as features does not further improve the model. To the best of our knowledge, this is the first data-driven study that focuses on analyzing discourse-level factors and predicting sentence deletion on a large English text simplification corpus.

2 Data and Setup

We use the Newsela text simplification corpus (Xu, Callison-Burch, and Napoles 2015) of 936 news articles. Each article set consists of 4 or 5 simplified versions of the original article, ranging from grades 3-12 (corresponding to ages 8-18). We group articles into three reading levels: original (grade 12), middle school (grades 6-8) and elementary school (grades 3-5). We use one version of article from each reading level, and study two document-level transformations: original $\rightarrow$ middle and original $\rightarrow$ elementary.

We conduct analysis and learn to predict if a sentence would be dropped by professional editors when simplifying text to the desired reading levels. To obtain labeled data for analysis and evaluation, we manually align sentences of 50 article sets. The resulting dataset is one of the largest manually annotated datasets for sentence alignment in simplification. Figure 1 shows a 3-sentence paragraph in the original article, aligned to the elementary school version. Sentences in the original article that cannot be mapped to any sentence in a lower reading level are considered deleted. To train models for sentence deletion prediction, we rely on noisy supervision from automatically aligned sentences from the rest of the corpus.

Manual alignment. Manual sentence alignment is conducted on 50 sets of the articles (2,281 sentences in the original version), using a combination of crowdsourcing and in-house analysis. The annotation process is designed to be efficient, with rigorous quality control, and includes the following steps:

(1) Align paragraphs between articles by asking in-house annotators to manually verify and correct the automatic alignments generated by the CATS toolkit (Štajner et al. 2018). Automatic methods are much more reliable for aligning paragraphs than sentences, given the longer contexts. We use this step to reduce the number of sentence pairs that need to be annotated.

(2) Collect human annotations for sentence alignment using Figure Eight,[4] a crowdsourcing platform. For every possible pair of sentences within the aligned paragraphs, we ask 5 workers to classify it into three categories: meaning equivalent, partly overlapped, or mismatched.[5] To ensure quality, we embedded a hidden test question in every five questions we asked, and removed workers whose accuracy dropped below 80% on the test questions. The inter-annotator agreement is 0.807 by Cohen's kappa (Artstein and Poesio 2008).

For this study, we consider a sentence in the original article **deleted** by the editor during simplification, if there is no corresponding sentence labeled as meaning equivalent or partly overlap in the elementary or middle school levels. For sentences that are shortened or split, we consider them as being **kept**. The final annotations for sentence alignment are aggregated by majority vote, then verified by the in-house annotators (not the authors).

Automatic alignment. We align sentences between pairs of articles based on the cosine similarity of their vector representations. We use 700-dimensional sentence embeddings pretrained on 16GB English Wikipedia by Sent2Vec (Pagliardini, Gupta, and Jaggi 2018), an unsupervised method that learns to compose sentence embeddings from word vectors along with bigram character vectors. Automatic aligned data includes a total of 42,264 sentences from the original article.

[4] https://figure-eight.com

[5] We provided the following guideline: (i) two sentences are equivalent if they convey the same meaning, though one sentence can be much shorter or simpler than the other sentence; (ii) two sentences partially overlap if they share information in common but some important information differs/is missing; (iii) two sentences are mismatched, otherwise.

	Middle	Elementary
	Avg. (std)	Avg. (std)
automatic alignment	0.146 (±0.158)	0.272 (±0.172)
manual alignment	0.172 (±0.154)	0.448 (±0.161)

Table 1: Fraction of sentences deleted from the original article to meet middle and elementary school reading standards.

	Middle		Elementary	
	Corr.	p-value	Corr.	p-value
# of sentences	0.849	6.8e-15	0.470	5.6e-4
# of tokens	0.845	1.2e-14	0.487	3.3e-4

Table 2: Pearson correlation between deletion rate and document length, measured by the number of sentences and tokens respectively.

We consider a pair of sentences aligned if their similarity exceeds 0.94 when no sentence splitting is involved, or 0.47 when splitting occurs. These thresholds are calibrated on the manually annotated set of article pairs. Empirically tested on the manually labeled data, this alignment strategy is more accurate (72% accuracy) than the alignment method used by (Xu, Callison-Burch, and Napoles 2015) (67% accuracy), which is based on Jaccard similarity.

Corpus Statistics. Table 1 shows the average and standard deviation of the portion of sentences deleted when an article is being simplified from original to middle or elementary levels. Notably, the standard deviation of the deletion ratio is high, which reflects the multi-facet nature of sentence deletion in simplification (c.f. Section 3). Simplifying to the elementary level involves on average 27.6% more deletion than to the middle school level. We also find that automatic alignment results in a much lower deletion rate, indicating that it over-match sentences.

3 Analysis of Discourse Level Factors

We present a series of analyses to study discourse level factors, including document characteristics, rhetorical structure, and discourse relations, that potentially influence sentence deletion during simplification.

3.1 Document Characteristics

Document length We hypothesize that the length of the original article will impact how much content professional editors choose to compress to reach a certain reading level. Table 2 tabulates the Pearson correlation between the number of sentences and words in the original document versus the number of deleted sentences, on the manually aligned articles. The correlations are significant for both middle and elementary levels, yet with the middle level the correlation values are particularly high. Longer documents indeed have higher percentages of sentences being deleted.

	# of articles	Middle	Elementary
Science	282	$-\ 0.0380^{*}$	$-\ 0.0722^{*}$
Health	92	$-\ 0.0253$	$-\ 0.0033$
Arts	79	$-\ 0.0200$	$+\ 0.0014$
War	170	$-\ 0.0192$	$-\ 0.0140$
Kids	179	$+\ 0.0029$	$+\ 0.0147$
Money	160	$+\ 0.0230^{*}$	$+\ 0.0169$
Law	193	$+\ 0.0283^{*}$	$+\ 0.0402$
Sports	95	$+\ 0.0488$	$+\ 0.0300$

Table 3: Average difference between deletion rate of each topic and the average. * indicates statistical significance ($p < 0.05$) comparing the distribution of one topic and the mean of all other topics based on the two sample Kolmogorov-Smirnov test (Massey Jr 1951).

Topics Intuitively, different topics could also lead to varied difficulty in reading comprehension. We compare percentages of sentences deleted across different article categories available in the Newsela dataset. We conduct this particular analysis on all articles, including auto aligned ones, as the distribution of topic labels is sparse on the manually aligned subset. Because of the noise, we compare deletion rates *relative to the mean*. Shown in Table 3, topics vary in their deletion rates. *Science* articles have significantly lower deletion rates for both middle and elementary levels. Articles about *Money* and *Law* have significantly higher deletion rates than others.

3.2 Rhetorical Structure

Rhetorical Structure Theory (RST) describes the relations between text spans in a discourse tree, starting from elementary discourse units (roughly, independent clauses). An argument of a relation can be a nucleus (presents more salient information) or a satellite, illustrated in Figure 2. RST is known to be useful in related applications, including summarization (Marcu 1999; Hirao et al. 2013; Durrett, Berg-Kirkpatrick, and Klein 2016) where information salience plays a central role.

In this section, we focus on how each sentence is situated in the RST tree of the original document, hence we treat each sentence as a discourse unit (that is not necessarily an elementary discourse unit). We use the discourse parser from (Surdeanu, Hicks, and Valenzuela-Escárcega 2015) to process each document.

Depth in discourse tree RST captures the salience of a sentence with respect to its role in the larger context. In particular, the salience of a unit or sentence does not strictly follow the linear order of appearance in the document, but is indicated by its distance to the highest level of topical organization (Cristea, Ide, and Romary 1998). Indeed, we found that the relative position of a sentence is not strongly correlated with the depth of the sentence in the discourse tree: the Pearson correlations are 0.064 and -0.088 for kept and deletion conditions at the elementary level, respectively. To this end, we consider the depth of the current sentence in the RST tree of the document (viewing each sentence as a

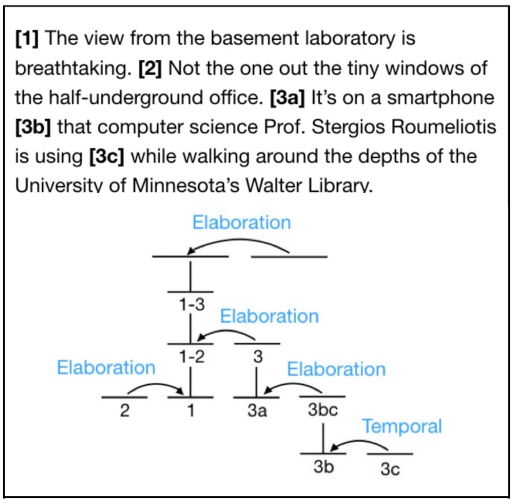

Figure 2: An example RST tree of a segment in an original news article. The arrows represent nucleus (arrow head) and satellite (arrow tail). In the elementary level, [1] is kept and rephrased, [2] is deleted, the third sentence is kept but split into two – [3a] and [3b] as one sentence, and [3c] as another.

	Middle		Elementary	
	Kept	Deleted	Kept	Deleted
Mean depth (std)	7.63 ($\pm$3.05)	9.06 ($\pm$3.78)	7.58 ($\pm$3.09)	8.36 ($\pm$3.44)

Table 4: Depth distribution of sentences in RST trees. Deleted sentences are located significantly ($p < 0.05$) lower than the ones that are kept, using the Wilcoxon ranked test (Wilcoxon 1992).

discourse unit). The distributions of depth for both deleted and kept sentences are shown in Table 4. We observe that sentences that are deleted locate at significantly lower levels in their discourse trees comparing to those that are kept. Since salient sentences tend to locate closer to the root of the discourse tree, this indicates that salience plays some role in the decision whether a sentence should be deleted.

Nuclearity In addition to global salience captured by depth above, we also look into local information salience. An approximate of the importance of a sentence among its neighboring sentences is the nuclearity information between the relations of the sentence and its neighbors. Therefore, we compare how often sentences that are nuclei of their parent relation are deleted vs. satellite ones, across each level of simplification. While we found that satellite sentences tend to be deleted for the elementary level, the differences are small, and a Chi-Squared test yield no significance ($p = 0.2$ for both reading levels).

3.3 Discourse Relations

Discourse relations signal relationships (e.g., contrast, causal) between clauses and sentences. These relations capture pragmatic aspects of text and are prominent players in text simplification (Siddharthan 2014; 2003). Prior work also suggested that sentences in an instantiating role tend to be very detailed (Li and Nenkova 2016), and different relations could indicate different levels of importance for the article as a whole (Louis, Joshi, and Nenkova 2010). In this work, we look at (a) relations that connect a sentence to the rest of the document, and (b) the usage of explicit discourse connectives within a sentence.

	Middle		Elementary	
	Kept	Deleted	Kept	Deleted
Root	0.084	0.057 $\downarrow$	0.115	0.038 $\downarrow$
Elaboration	0.793	0.816	0.752	0.840 $\uparrow$
Contrast	0.040	0.042	0.041	0.033
Background	0.019	0.012	0.022	0.021
Evaluation	0.017	0.010	0.016	0.018
Explanation	0.019	0.011 $\downarrow$	0.020	0.016

Table 5: Fraction of discourse relations that govern the sentence's sub-tree. Arrows indicate significance ($p < 0.05$) using the Wilcoxon ranked test; $\uparrow$: higher presence among deleted sentences than the kept ones; $\downarrow$: lower.

Inter-sentential relations We first consider how a sentence is connected with the rest of the document such that its appearance renders the document coherent, especially when information more "salient" to this sentence is present and less likely to be deleted. To this end, we study the lowest ancestor relation to which it is attached as a satellite, henceforth the "governing" relation. Table 5 shows the fraction of sentences belonging to the top 6 governing relations.[6] Observe that the *elaboration* relation is the most frequent relation in the dataset; sentences serving as an elaboration of another sentence are more likely to be removed during simplification (statistically significant for the elementary level). Important sentences that are not satellites to any relation (*root*) is significantly less likely to be deleted across both levels. Furthermore, sentences that serve as an *explanation* of an existing sentence are less likely to be deleted during simplification (significantly, for the middle school level).

Discourse connectives Discourse connectives are found to have different rates of mental processing in cognitive experiments (Sanders and Noordman 2000). To identify discourse connectives, we parse each document with the NUS parser (Lin, Ng, and Kan 2014) for the Penn Discourse Treebank (PDTB) (Prasad et al. 2008). The fraction of sentences containing a connective is shown in Table 6. Higher reading levels (original and middle) contain significantly ($p = 1e - 48$) more discourse connectives per sentence.

We first compare how often discourse connectives appear in deleted vs. kept sentences in the original version, and the relation senses they signal: contingency, comparison, expan-

[6] We did not include other relations as their frequencies are less than 0.5% in the dataset.

	Original	Middle	Elementary
	Avg. (std)	Avg. (std)	Avg. (std)
% of sents	0.38 (±0.17)	0.34 (±0.16)	0.22 (±0.09)

Table 6: Fraction of sentences that contain explicit discourse connectives.

	Middle		Elementary	
	Kept	Deleted	Kept	Deleted
% of sents	0.305	0.337↑	0.312	0.352↑
Contingency	0.077	0.079↑	0.081	0.087↑
Comparison	0.064	0.085↑	0.066	0.094↑
Expansion	0.118	0.125	0.117	0.132↑
Temporal	0.111	0.099↓	0.098	0.107↑

Table 7: Fraction of sentences that contain the explicit discourse connectives. Arrows indicate significance ($p < 0.05$) using the Wilcoxon ranked test; ↑: higher presence among deleted sentences than the kept ones; ↓: lower.

sion, or temporal, following the PDTB taxonomy. We did not conduct analysis on fine-grained levels of the taxonomy due to label sparsity. Table 7 shows the fraction of sentences containing at least one connective, as well as fraction of sentences containing connectives of a certain sense. Deleted sentences are in general significantly more likely to have a connective, a potential signal that the sentence is complex (i.e., has more than one clause). This is especially evident for the elementary level in that this holds for all relations. Deleted sentences are significantly less likely to have temporal ones in the middle level. One explanation could be that temporal connectives presuppose the events involved (Lascarides and Oberlander 1993), hence we think they need to be included for the reader to be able to comprehend text as a whole.

We also study the position of these explicit connectives. Specifically, connective positions (start of a sentence vs. not) are strong indicators of whether the relation they signal is intra- or inter-sentential (Lin, Ng, and Kan 2014; Biran and McKeown 2015). Table 8 reveals that when targeting the middle level, sentences with connectives at the beginning ("sent-initial") are much more likely ($p = 9e-6$) to be kept. This could indicate editors being unwilling to delete closely related sentences when the simplification strategy is less aggressive.

4 Predicting Sentence Deletion

We run our experiments on two tasks, first on building a classification model to see if it can predict whether a sentence should be deleted when simplifying to middle and element level. Second, we perform the feature ablation to determine whether in practice document and discourse signals help under noisy supervision.

	Middle		Elementary	
	Kept	Deleted	Kept	Deleted
Sent-initial	0.584	0.195↓	0.405	0.421
Non-initial	0.740	0.147↓	0.561	0.395↓

Table 8: Fraction of sentences with a discourse connective at the start of the sentence or otherwise. ↓ indicates a significantly ($p < 0.05$) lower presence among deleted sentences than the kept ones based on two sample Kolmogorov-Smirnov test.

Experiment Setup Given a sentence in the original article, we (i) predict whether it will be deleted when simplifying to the middle school level, trained on noisy supervision from automatic alignments; (ii) predict the same for the elementary level. We use 15 of the manually aligned articles as the validation set and the other 35 articles as test set.

Method We use logistic regression (LR) and feedforward neural networks (FNN) as classifiers,[7] and experiment with features from multiple, potentially complementary aspects. To capture **sentence-level semantics**, we consider the average of GloVe word embeddings (Pennington, Socher, and Manning 2014). The sparse features (SF) include the relative position of the sentence in the whole article, as well as in the paragraph it resides. Additionally, we include readability scores for the sentence following (Scarton, Paetzold, and Specia 2018)[8]. Leveraging our corpus analysis (Section 3), we incorporate **document-level features**, including the total number of sentences and number of words in the document, as well as the topic of the document. Our **discourse features** include the depth of the current sentence, indicator features for nuclearity and the governing relation of the current sentence in the RST tree, whether there is an explicit connective of one of the four relations we analyzed, and the position of the connective. We also use the **position** of the sentence, as sentences appearing later in an article are more likely to be dropped (Petersen and Ostendorf 2007).

To improve the prediction performance, we adopted a smooth binning approach (Maddela and Xu 2018) and project each of the sparse features, which are either binary or numerical, into a k-dimentional vector representation by applying k Gaussian radial basis functions.

Implementation Details We use Pytorch to implement the neural network model. To get a sentence representation, we take the average word embeddings as input and stack two hidden layers with ReLU activation, and a single-node linear output layer if only embeddings are used for classification. To combine the learned embedding features and sparse fea-

[7] We also tried BiLSTM based feature encoding and it gives similar results for the prediction.

[8] Flesch Reading Ease, Flesch-Kincaid Grade Level, SMOG Index, Gunning Fog Index, Automated Readability Index, Coleman-Liau Index, Linsear Write Formula and Dale-Chall Readability Score.

Model (Elementary)	Precision	Recall	F1
Random	42.1	48.8	45.2
LR Embedding	58.1	58.0	58.1
FNN Embedding	59.2	68.0	63.2
LR All Sparse Features	69.6	45.9	55.3
LR All SF binning	59.9	65.7	62.7
FNN All Sparse Features	72.3	47.7	57.4
FNN All SF binning	70.2	54.0	61.0
LR Embed & Sparse Feature	70.4	43.1	53.4
LR Embed & SF binning	62.2	64.0	63.1
FNN Embed & SF binning	65.9	64.5	65.2

Table 9: Performance of predicting sentence deletions for elementary school level simplification.

Model (Middle)	Precision	Recall	F1
Random	21.1	49.0	29.5
LR Embedding	31.4	51.5	39.0
FNN Embedding	34.2	61.7	44.0
LR All Sparse Features	56.0	58.9	57.4
LR All SF binning	42.4	80.1	55.4
FNN All Sparse Features	55.9	58.6	57.2
FNN All SF binning	55.9	63.8	59.6
LR Embed & Sparse Feature	55.9	58.6	57.2
LR Embed & SF binning	41.7	75.9	53.9
FNN Embed & SF binning	56.4	63.6	59.7

Table 10: Performance of predicting sentence deletions for middle school level simplification.

tures, we concatenate the output of second hidden layer of the embedding network with the binned sparse features, and feed them into a multi-layer feedforward network with two hidden layers and one last sigmoid layer for classification. We use 300-dimensional GloVe embeddings and a total of 35 sparse features as inputs, and half of the input size nodes in each hidden layer. The training objective is to minimize the binary cross entropy loss between the logits and the true binary labels. We use Adam (Kingma and Ba 2015) for optimization and also apply a dropout of 0.5 to prevent overfitting. We set the learning rate to 1e-5 and 2e-5 for experiments in Tables 9 and 10 respectively. We set the batch size to 64. We followed (Maddela and Xu 2018) and set the number of bins k to 10 and the adjustable fraction γ to 0.2 for the Gaussian feature vectorization layer. We implemented the logistic regression classifier using Scikit-learn (Pedregosa et al. 2011). Since our data heavily skews towards keeping all sentences, we use downsampling to balance the two classes.

Results The experimental results are shown in Tables 9 and 10. As baseline, we consider randomly removing sentences according to deletion rates in the training set. We then look at using the semantic content of the sentences, captured by GLoVe embeddings, and/or using the sparse features. In general, we find this a challenging task. Predicting sentence deletion at the middle level is more difficult than for elementary, as fewer sentences are deleted (c.f. Table 1). Comparing the uses of features, we find that middle level deletion and elementary level deletion depend on different features. As shown in Table 9 and Table 10, models using only sentence embedding have a higher performance on element level deletion prediction, yet embeddings are not that informative compared to sparse features in middle level. We also tried the two neighbor sentences' embedding as a bonus feature, but find that they have a negative effects over the middle level deletion task and barely no improvement over element level. This reflects the difficulty in middle level deletion prediction, as professional editors might not use specific strategies to pick up deletion candidates based on their sentence semantics alone. On the other hand, using sparse features only for both levels gives comparable results to the best model that utilizes both categories of features. We also find that the Gaussian binning methods proposed by (Maddela and Xu 2018) significantly help the model to make use of the sparse features, which are initially a small number of discrete features.

Model (Elementary)	Precision	Recall	F1
LR Embed & SF binning	62.2	64.0	63.1
– Discourse	62.1	63.7	62.9
– Document	60.7	59.2	60.0↓
– Position	58.5	62.7	60.5↓
Model (Middle)	**Precision**	**Recall**	**F1**
LR Embed & SF binning	41.7	75.9	53.9
– Discourse	41.7	75.9	53.9
– Document	36.0	64.1	46.1↓
– Position	39.1	78.0	52.1↓

Table 11: Feature ablation analysis for predicting sentence deletion by removing one feature category at a time. ↓ indicates significant drop of performance ($p < 0.05$) compared to using all features based on the bootstrapping test (Berg-Kirkpatrick, Burkett, and Klein 2012).

For feature ablation, we perform the experiments over the Logistic Regression model since neural models can be heavily influenced by hyper-parameters and random initialization (Yang et al. 2019). For both levels, document characteristics matter more than position, especially in the middle level task. One reason could be that middle level simplification is based more on content, while editors tend to shorten the whole texts in elementary level simplification by dropping sentences near the end of an article. Overall, the RST and discourse relation features do not help too much, possibly because these features tend to have much lower triggering rates than others, e.g., not every sentence has explicit discourse features as shown in Table 7. Another reason could be the noise introduced during automatic alignment, for example, potential noise in partial match, that could render signals from discourse relations uninformative during training.

5 Related Work

Most existing work on text simplification focuses on word/phrase-level (Yatskar et al. 2010; Biran, Brody, and Elhadad 2011; Specia, Jauhar, and Mihalcea 2012; Glavaš and Štajner 2015; Paetzold and Specia 2017; Maddela and Xu 2018) or sentence-level simplifications (Zhu, Bernhard, and Gurevych 2010; Xu et al. 2016; Štajner and Nisioi 2018; Dong et al. 2019). Only a few projects conducted corpus analyses and automatic prediction on sentence deletion during document-level simplification, including the pioneer work by Petersen and Ostendorf (2007). They analyzed a corpus from Literacyworks (unfortunately, inaccessible by other researchers), and reported the prediction on which sentences will be dropped is "little better than always choosing the majority class (not dropped)" using a decision tree based classifier.

In contrast, we study the Newsela corpus which has been widely used among researchers since its release (Xu, Callison-Burch, and Napoles 2015), as it offers a sizable collection of news articles written by professional editors at five different readability levels. It exhibits more significant sentence dropping and discourse reorganization phenomena. We present an in-depth analysis, focusing on various discourse-level factors that are important to understand for developing document-level automatic simplification systems, very different from prior studies of Newsela (Xu, Callison-Burch, and Napoles 2015; Scarton, Paetzold, and Specia 2018) that focused on vocabulary usage and sentence readability. Other related works include Štajner, Drndarević, and Saggion (2013)'s on Spanish, Gasperin et al. (2009)'s on Brazilian Portuguese, and Gonzalez-Dios, Aranzabe, and de Ilarraza (2018)'s on Basque.

More importantly, nearly all the existing studies on sentence deletion and splitting in simplification are based on automatically aligned sentence pairs, without manually labeled ground truth to gauge the reliability of the findings. This is largely due to the scarcity and cost of manually labeled sentence alignment data. In this paper, we present an efficient crowdsourcing methodology and the first manually annotated, high-quality sentence alignment corpus for simplification. To the best of our knowledge, the most comparable dataset is that created by Hwang et al. (2015) using Wikipedia data, which is inherently noisy as shown by Xu, Callison-Burch, and Napoles (2015), due to the lack of quality control and strict editing guideline in creating the Simple English Wikipedia.

6 Conclusion

This paper presents a parallel text simplification corpus with manually aligned sentences across multiple reading levels from the Newsela dataset. Our corpus analysis show that discourse-level factors are important when editors drop sentences as they simplify. We further show that document characteristic features help in predicting whether a sentence will be deleted during simplification, a challenging task given the low deletion rate when simplifying to the middle school level. To the best of our knowledge, this is the first data-driven study that focuses on analyzing discourse factors and predicting sentence deletion on a large English text simplification corpus. We hope this work will spur more future research on automatic document simplification.

Acknowledgments

We thank NVIDIA and Texas Advanced Computing Center at UT Austin for providing GPU computing resources, the anonymous reviewers and Mounica Maddela for their valuable comments. We also thank Sarah Flanagan, Bohan Zhang, Raleigh Potluri, and Alex Wing for their help with data annotation. This research was partly supported by the NSF under grants IIS-1822754, IIS-1850153, a Figure-Eight AI for Everyone Award and a Criteo Faculty Research Award to Wei Xu. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of the U.S. Government.

References

Artstein, R., and Poesio, M. 2008. Survey article: inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4):555–596.

Berg-Kirkpatrick, T.; Burkett, D.; and Klein, D. 2012. An empirical investigation of statistical significance in NLP. In *EMNLP*, 704–714.

Biran, O., and McKeown, K. 2015. PDTB discourse parsing as a tagging task: The two taggers approach. In *SIGDIAL*, 96–104.

Biran, O.; Brody, S.; and Elhadad, N. 2011. Putting it simply: a context-aware approach to lexical simplification. In *ACL*, 496–501.

Cristea, D.; Ide, N.; and Romary, L. 1998. Veins theory: a model of global discourse cohesion and coherence. In *ACL*, 281–285.

Dong, Y.; Li, Z.; Rezagholizadeh, M.; and Cheung, J. C. K. 2019. EditNTS: an neural programmer-interpreter model for sentence simplification through explicit editing. In *ACL*, 3393–3402.

Drndarevic, B., and Saggion, H. 2012. Reducing text complexity through automatic lexical simplification: an empirical study for Spanish. *Procesamiento del Lenguaje Natural* 49:13–20.

Durrett, G.; Berg-Kirkpatrick, T.; and Klein, D. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *ACL*, 1998–2008.

Gasperin, C.; Specia, L.; Pereira, T.; and Aluísio, S. 2009. Learning when to simplify sentences for natural text simplification. In *ENIA*, 809–818.

Glavaš, G., and Štajner, S. 2015. Simplifying lexical simplification: do we need simplified corpora? In *ACL*, 63–68.

Gonzalez-Dios, I.; Aranzabe, M. J.; and de Ilarraza, A. D. 2018. The corpus of Basque simplified texts. *Language Resources and Evaluation* 52(1):217–247.

Hirao, T.; Yoshida, Y.; Nishino, M.; Yasuda, N.; and Nagata, M. 2013. Single-document summarization as a tree knapsack problem. In *EMNLP*, 1515–1520.

Hwang, W.; Hajishirzi, H.; Ostendorf, M.; and Wu, W. 2015. Aligning sentences from standard Wikipedia to simple Wikipedia. In *NAACL*, 211–217.

Kingma, D. P., and Ba, J. 2015. Adam: a method for stochastic optimization. In *ICLR*.

Kriz, R.; Sedoc, J.; Apidianaki, M.; Zheng, C.; Kumar, G.; Miltsakaki, E.; and Callison-Burch, C. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. In *NAACL*, 3137–3147.

Lascarides, A., and Oberlander, J. 1993. Temporal connectives in a discourse context. In *EACL*, 260–268.

Li, J. J., and Nenkova, A. 2016. The instantiation discourse relation: a corpus analysis of its properties and improved detection. In *NAACL*, 1181–1186.

Lin, Z.; Ng, H. T.; and Kan, M.-Y. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering* 20(2):151–184.

Louis, A.; Joshi, A.; and Nenkova, A. 2010. Discourse indicators for content selection in summarization. In *SIGDIAL*, 147–156.

Maddela, M., and Xu, W. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *EMNLP*, 3749–3760.

Mann, W. C., and Thompson, S. A. 1988. Rhetorical structure theory: toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse* 8(3):243–281.

Marcu, D. 1999. Discourse trees are good indicators of importance in text. In *Advances in Automatic Text Summarization*, 123–136.

Massey Jr, F. J. 1951. The kolmogorov-smirnov test for goodness of fit. *JASA* 46(253):68–78.

Narayan, S., and Gardent, C. 2016. Unsupervised sentence simplification using deep semantics. In *INLG*, 111–120.

Nisioi, S.; Štajner, S.; Ponzetto, S. P.; and Dinu, L. P. 2017. Exploring neural text simplification models. In *ACL*, 85–91.

Paetzold, G., and Specia, L. 2017. Lexical simplification with neural ranking. In *EACL*, 34–40.

Pagliardini, M.; Gupta, P.; and Jaggi, M. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *NAACL*, 528–540.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: machine learning in Python. *JMLR* 12:2825–2830.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: global vectors for word representation. In *EMNLP*, 1532–1543.

Petersen, S. E., and Ostendorf, M. 2007. Text simplification for language learners: a corpus analysis. In *SLaTE*, 69–72.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A. K.; and Webber, B. L. 2008. The Penn Discourse TreeBank 2.0. In *LREC*.

Sanders, T. J., and Noordman, L. G. 2000. The role of coherence relations and their linguistic markers in text processing. *Discourse processes* 37–60.

Scarton, C.; Paetzold, G.; and Specia, L. 2018. Text simplification from professionally produced corpora. In *LREC*.

Siddharthan, A. 2003. Preserving discourse structure when simplifying text. In *ENLG*, 103–110.

Siddharthan, A. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics* 165(2):259–298.

Specia, L.; Jauhar, S. K.; and Mihalcea, R. 2012. SemEval-2012 Task 1: English lexical simplification. In *SemEval*, 347–355.

Štajner, S., and Nisioi, S. 2018. A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification. In *LREC*.

Sulem, E.; Abend, O.; and Rappoport, A. 2018. Simple and effective text simplification using semantic and neural methods. In *ACL*, 162–173.

Surdeanu, M.; Hicks, T.; and Valenzuela-Escárcega, M. A. 2015. Two practical rhetorical structure theory parsers. In *NAACL: Software Demonstrations*, 1–5.

Štajner, S.; Franco-Salvador, M.; Rosso, P.; and Ponzetto, S. P. 2018. CATS: a tool for customized alignment of text simplification corpora. In *LREC*.

Štajner, S.; Drndarević, B.; and Saggion, H. 2013. Corpus-based sentence deletion and split decisions for Spanish text simplification. *Computacion y Systemas* 17(2):251–262.

Vu, T.; Hu, B.; Munkhdalai, T.; and Yu, H. 2018. Sentence simplification with memory-augmented neural networks. In *NAACL*, 79–85.

Wilcoxon, F. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*. Springer. 196–202.

Woodsend, K., and Lapata, M. 2011. Wikisimple: Automatic simplification of wikipedia articles. In *AAAI*, 927–932.

Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; and Callison-Burch, C. 2016. Optimizing statistical machine translation for text simplification. *TACL* 4:401–415.

Xu, W.; Callison-Burch, C.; and Napoles, C. 2015. Problems in current text simplification research: new data can help. *TACL* 3:283–297.

Yang, W.; Lu, K.; Yang, P.; and Lin, J. 2019. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *SIGIR*, 1129–1132.

Yatskar, M.; Pang, B.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2010. For the sake of simplicity: unsupervised extraction of lexical simplifications from Wikipedia. In *NAACL*, 365–368.

Zhang, X., and Lapata, M. 2017. Sentence simplification with deep reinforcement learning. In *EMNLP*, 584–594.

Zhu, Z.; Bernhard, D.; and Gurevych, I. 2010. A monolingual tree-based translation model for sentence simplification. In *COLING*, 1353–1361.