Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings

Tomoyuki Kajiwara
Graduate School of System Design
Tokyo Metropolitan University
Tokyo, Japan
kajiwara-tomoyuki@ed.tmu.ac.jp

Mamoru Komachi
Graduate School of System Design
Tokyo Metropolitan University
Tokyo, Japan
komachi@tmu.ac.jp

Abstract

Methods for text simplification using the framework of statistical machine translation have been extensively studied in recent years. However, building the monolingual parallel corpus necessary for training the model requires costly human annotation. Monolingual parallel corpora for text simplification have therefore been built only for a limited number of languages, such as English and Portuguese. To obviate the need for human annotation, we propose an unsupervised method that automatically builds the monolingual parallel corpus for text simplification using sentence similarity based on word embeddings. For any sentence pair comprising a complex sentence and its simple counterpart, we employ a many-to-one method of aligning each word in the complex sentence with the most similar word in the simple sentence and compute sentence similarity by averaging these word similarities. The experimental results demonstrate the excellent performance of the proposed method in a monolingual parallel corpus construction task for English text simplification. The results also demonstrated the superior accuracy in text simplification that use the framework of statistical machine translation trained using the corpus built by the proposed method to that using the existing corpora.

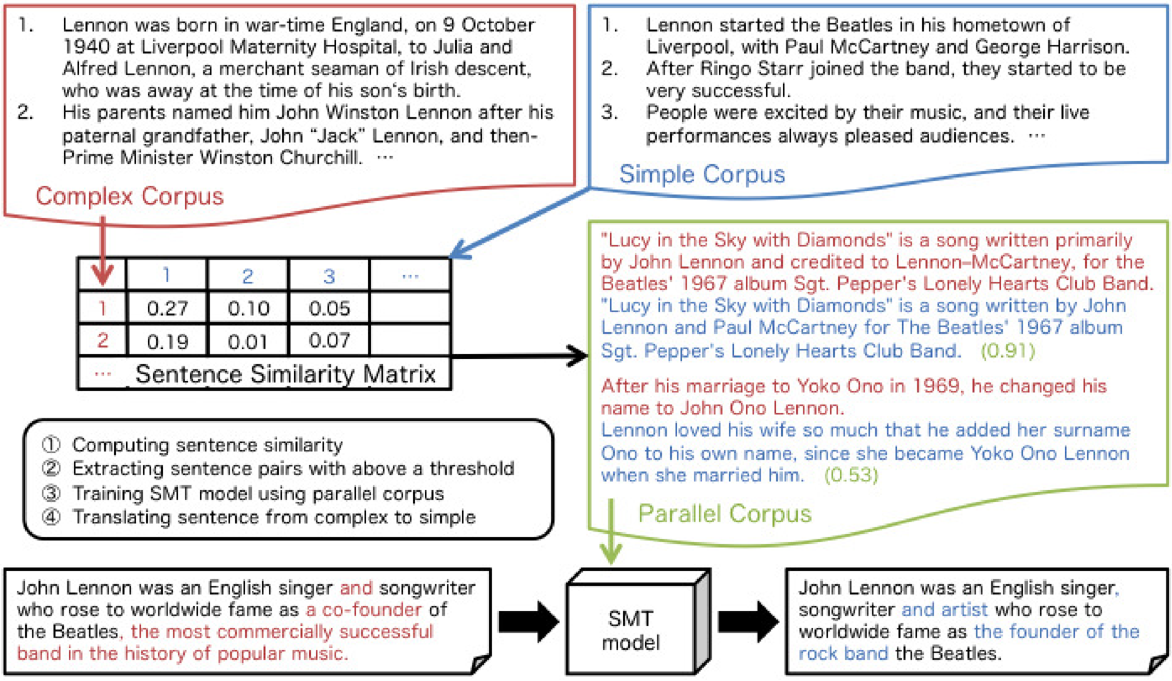
1 Introduction

Text simplification is the process of rewriting a complex text into a simpler form while preserving its meaning. The purpose of text simplification is to assist the comprehension of readers, especially language learners and children. Recent studies have treated text simplification as a monolingual machine translation problem in which a simple synonymous sentence is generated using the framework of statistical machine translation (Specia, 2010; Zhu et al., 2010; Coster and Kauchak, 2011a; Coster and Kauchak, 2011b; Wubben et al., 2012; Štajner et al., 2015a; Štajner et al., 2015b; Goto et al., 2015). However, unlike statistical machine translation, which uses bilingual parallel corpora, text simplification requires a monolingual parallel corpus for training. While bilingual parallel data are available in large quantities, monolingual parallel data are hard to obtain because simplification of a complex text is not a by-product of other tasks. Monolingual parallel corpora for text simplification are available in only seven languages—English (Zhu et al., 2010; Coster and Kauchak, 2011b; Hwang et al., 2015; Xu et al., 2015), Portuguese (Caseli et al., 2009), Spanish (Bott and Saggion, 2011), Danish (Klerke and Søgaard, 2012), German (Klaper et al., 2013), Italian (Brunato et al., 2015), and Japanese (Goto et al., 2015). In addition, only the English corpora are open to the public. We therefore propose an unsupervised method [1] that automatically builds monolingual parallel corpora for text simplification without using any external resources for computing sentence similarity.

In this study, a monolingual parallel corpus for text simplification is built from a comparable corpus comprising complex and simple texts. This was done in two steps. First, we compute the similarity for all combinations of complex and simple sentences using the alignment between word embeddings. Second, we extract sentence pairs whose similarity exceeded a certain threshold. Figure 1 gives an overview of the method. Monolingual parallel corpus can be used for text simplification in the framework of SMT.

[1] https://github.com/tmu-nlp/sscorpus

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,
pages 1147–1158, Osaka, Japan, December 11-17 2016.



1. Lennon was born in war-time England, on 9 October 1940 at Liverpool Maternity Hospital, to Julia and Alfred Lennon, a merchant seaman of Irish descent, who was away at the time of his son's birth.
2. His parents named him John Winston Lennon after his paternal grandfather, John "Jack" Lennon, and then-Prime Minister Winston Churchill. ⋯

1. Lennon started the Beatles in his hometown of Liverpool, with Paul McCartney and George Harrison.
2. After Ringo Starr joined the band, they started to be very successful.
3. People were excited by their music, and their live performances always pleased audiences. ⋯

	1	2	3	⋯
1	0.27	0.10	0.05	
2	0.19	0.01	0.07	
⋯				

"Lucy in the Sky with Diamonds" is a song written primarily by John Lennon and credited to Lennon–McCartney, for the Beatles' 1967 album Sgt. Pepper's Lonely Hearts Club Band.
"Lucy in the Sky with Diamonds" is a song written by John Lennon and Paul McCartney for The Beatles' 1967 album Sgt. Pepper's Lonely Hearts Club Band. (0.91)

After his marriage to Yoko Ono in 1969, he changed his name to John Ono Lennon.
Lennon loved his wife so much that he added her surname Ono to his own name, since she became Yoko Ono Lennon when she married him. (0.53)

① Computing sentence similarity
② Extracting sentence pairs with above a threshold
③ Training SMT model using parallel corpus
④ Translating sentence from complex to simple

John Lennon was an English singer and songwriter who rose to worldwide fame as a co-founder of the Beatles, the most commercially successful band in the history of popular music.

John Lennon was an English singer, songwriter and artist who rose to worldwide fame as the founder of the rock band the Beatles.

Figure 1: Process flow of building a monolingual parallel corpus and simplifying a sentence using the SMT framework.

We evaluated our proposed method using a benchmarking dataset [2] to construct a corpus for English text simplification. The benchmark dataset contains pairings of complex and simple sentences with a binary label of parallel (the sentence pair is synonymous) or nonparallel (the sentence pair is not synonymous). Intrinsic evaluation using this dataset showed that the proposed method had an improved F1 score. In addition, we built a statistical machine translation model trained on the resulting corpus and compared it with one trained on the existing corpora. Extrinsic evaluation using statistical machine translation for text simplification demonstrated the improved BLEU score of the proposed method.

Our contributions are summarized as follows:

- The proposed method improved the binary classification task between monolingual parallel data and nonparallel data by 3.1 points (0.607 → 0.638), compared with the F1 score from a previous study, and demonstrated high accuracy in building a monolingual parallel corpus for text simplification.

- The SMT-based text simplification model trained using the corpus built by the proposed method had a BLEU score 3.2 points higher (44.3 → 47.5) than an SMT-based text simplification model trained using the state-of-the-art monolingual parallel corpus.

- The proposed method can build a monolingual parallel corpus for text simplification at low cost because it does not require any external resources such as labeled data or dictionaries when computing sentence similarity.

2 Related Work

The statistical machine translation framework has become widely used in text simplification. In English, text simplification using a monolingual parallel corpus extracted from the English Wikipedia and Simple English Wikipedia has been actively studied. Coster and Kauchak (2011b) simplified sentences using the standard phrase-based SMT toolkit Moses (Koehn et al., 2007) and evaluated it using the standard automatic MT evaluation metric BLEU (Papineni et al., 2002). In addition to generic SMT translation models, specialized translation models such as targeting phrasal deletion have been proposed (Zhu et al., 2010; Coster and Kauchak, 2011a; Wubben et al., 2012). These studies reported that models specialized

[2] http://ssli.ee.washington.edu/tial/projects/simplification/

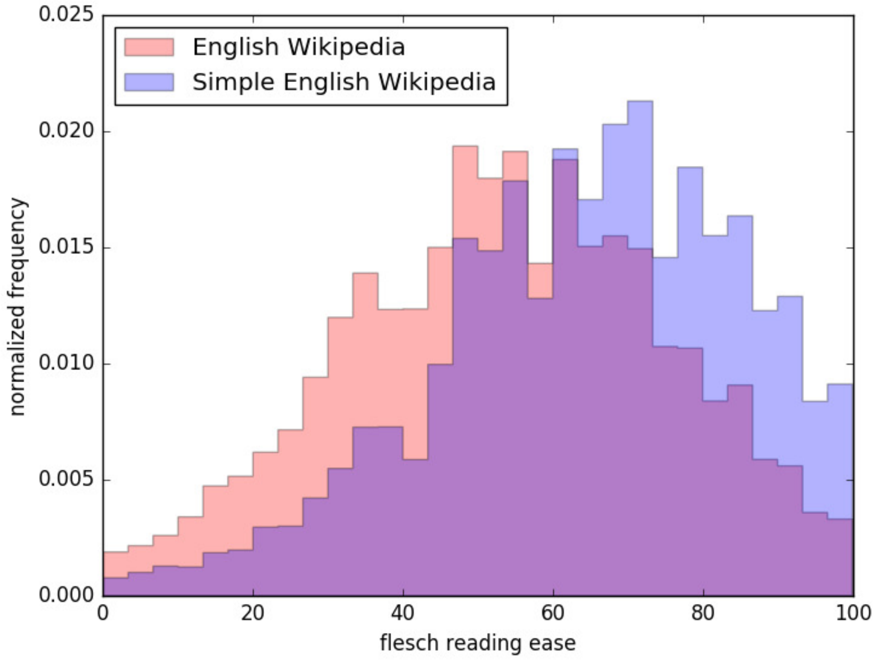


Figure 2: Readability score distribution of English Wikipedia and Simple English Wikipedia. A higher score in Flesch Reading Ease indicates simpler sentences.

in text simplification improved readability and the BLEU score. In languages other than English, text simplification using SMT has been studied for Portuguese (Specia, 2010), Spanish (Štajner et al., 2015b), and Japanese (Goto et al., 2015). We follow these works in applying SMT to text simplification, whilst improving the quality and quantity of the monolingual parallel corpus using an unsupervised method.

Three monolingual parallel corpora for English text simplification have been built from English Wikipedia and Simple English Wikipedia. First, Zhu et al. (2010) [3] pioneered automatic construction of a text simplification corpus using the cosine similarity between sentences represented as TF-IDF vectors. Second, Coster and Kauchak (2011b) [4] extended Zhu et al. (2010)'s work by considering the order of the sentences. However, these methods did not compute similarities between different words. In text simplification, it would be useful to consider similarities between synonymous expressions when computing the similarity between sentences, since concepts are frequently rewritten from a complex to a simpler form. Third, Hwang et al. (2015) [2] computed the similarity between sentences taking account of word-level similarity using the co-occurrence of a headword in a dictionary and its definition sentence. We also consider word-level similarity to compute similarity between sentences but using word embeddings to build a text simplification corpus at low cost without requiring access to external resources.

These text simplification corpora built from English Wikipedia and Simple English Wikipedia received some criticism. Xu et al. (2015) point out that Zhu et al. (2010)'s corpus has 17% of sentence pairs unaligned (two sentences have different meanings or only have partial content overlap) and 33% of sentence pairs become more complex (the simple sentence has the same meaning as the original sentence but is not simpler). However, Simple English Wikipedia contains simpler expressions in general. Figure 2 shows the distribution of the readability scores of Simple English Wikipedia and English Wikipedia. It clearly illustrates that Simple English Wikipedia contains easier sentences than English Wikipedia and supports that it is a good source for text simplification. Štajner et al. (2015a) investigated the quality and quantity of a monolingual parallel corpus using the framework of statistical machine translation and showed that sentence pairs with a moderate level of similarity are effective for training text simplification models. Therefore, we use the sentence similarity method to accurately measure the moderate level of

[3] https://www.ukp.tu-darmstadt.de/data/sentence-simplification/simple-complex-sentence-pairs/

[4] http://www.cs.pomona.edu/~dkauchak/simplification/

similarity.

To address the challenge of computing the similarity between sentences containing different words with similar meanings, many methods have been proposed. In semantic textual similarity task (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015), sentence similarity is computed on the basis of word similarity following the success of word embeddings such as word2vec (Mikolov et al., 2013a). For example, a supervised approach using word embeddings when obtaining a word alignment achieved the best performance in SemEval-2015 Task 2 (Sultan et al., 2015). Word embeddings have also been used in unsupervised sentence similarity metrics (Mikolov et al., 2013b; Song and Roth, 2015; Kusner et al., 2015). These unsupervised sentence similarity metrics can be applied to the automatic construction of a monolingual parallel corpus for text simplification, without requiring the data to be labeled.

3 Sentence Similarity based on Alignment between Word Embeddings

We propose four types of sentence similarity measures for building a monolingual parallel corpus for text simplification, based on alignments between word embeddings that have achieved outstanding performance on different NLP tasks. The methods discussed in Sections 3.1-3.3 are the sentence similarity measures proposed by Song and Roth (2015) for a short text similarity task. The Word Mover's Distance (Kusner et al., 2015) discussed in Section 3.4 is another sentence similarity measure based on alignment between word embeddings that is known to achieve good performance on a document classification task.

3.1 Average Alignment

The sentence similarity $\mathrm{STS}_{\mathrm{ave}}(x, y)$ between sentence x and sentence y is computed by averaging the similarities between all pairs of words taken from the two sentences, as follows:

$$\mathrm{STS}_{\mathrm{ave}}(x, y) = \frac{1}{|\boldsymbol{x}||\boldsymbol{y}|} \sum_{i=1}^{|\boldsymbol{x}|} \sum_{j=1}^{|\boldsymbol{y}|} \phi(x_i, y_j) \tag{1}$$

Here, x_i denotes the i-th word in the sentence x ($\boldsymbol{x} = (x_1, x_2, \ldots, x_{|\boldsymbol{x}|})$), y_j denotes the j-th word in the sentence y ($\boldsymbol{y} = (y_1, y_2, \ldots, y_{|\boldsymbol{y}|})$), and $\phi(x_i, y_j)$ denotes the similarity between words x_i and y_j. We employed the cosine similarity as the word similarity $\phi(x_i, y_j)$.

3.2 Maximum Alignment

Average alignment, discussed in Section 3.1, is an intuitive method. However, it is not possible that all word pairs have a high similarity $\phi(x_i, y_j)$, even when considering synonymous sentence pairs. Moreover, it is often the case that many word similarities $\phi(x_i, y_j)$ are noise and are near to zero. Therefore, we utilize only accurate alignments by computing the sentence similarity $\mathrm{STS}_{\mathrm{asym}}(x, y)$ from the most similar word y_j for each word x_i rather than averaging the word similarities between all pairs. Here $\mathrm{STS}_{\mathrm{asym}}(x, y)$ is an inherently asymmetric score. Therefore, we obtain the symmetric sentence similarity $\mathrm{STS}_{\mathrm{max}}(x, y)$ by averaging the two similarities $\mathrm{STS}_{\mathrm{asym}}(x, y)$ and $\mathrm{STS}_{\mathrm{asym}}(y, x)$ as follows:

$$\mathrm{STS}_{\mathrm{asym}}(x, y) = \frac{1}{|\boldsymbol{x}|} \sum_{i=1}^{|\boldsymbol{x}|} \max_{j} \phi(x_i, y_j), \quad \mathrm{STS}_{\mathrm{max}}(x, y) = \frac{1}{2}(\mathrm{STS}_{\mathrm{asym}}(x, y) + \mathrm{STS}_{\mathrm{asym}}(y, x)) \tag{2}$$

3.3 Hungarian Alignment

Average alignment and *maximum alignment* can be considered as sentence similarity measures based on many-to-many word alignments and many-to-one word alignments, respectively. However, since these methods compute the word alignments independently, they do not take into account the sentence-level consistency of alignments. To address this lack of global alignment, we represent two sentences x and y as a bipartite graph in which the vertices consist of words that occur in each sentence and the edges reflect their word-level similarity. The graph is then used to define sentence similarity. This bipartite graph is a weighted complete bipartite graph whose edge is assigned a word similarity $\phi(x_i, y_j)$ as a weight. The one-to-one word alignment that maximizes the sum of the word similarities is obtained by finding the maximum matching of the bipartite graph. This maximum matching problem can be solved using the Hungarian algorithm (Kuhn, 1955). The sentence similarity $\mathrm{STS_{hun}}(x, y)$ is then computed by selecting a word $h(x_i)$ using the Hungarian algorithm for each word x_i:

$$\mathrm{STS_{hun}}(x, y) = \frac{1}{\min(|\boldsymbol{x}|, |\boldsymbol{y}|)} \sum_{i=1}^{|\boldsymbol{x}|} \phi(x_i, h(x_i)) \tag{3}$$

3.4 Word Mover's Distance

Word Mover's Distance (Kusner et al., 2015) also considers the global consistency of word alignments when computing sentence similarity based on a many-to-many word alignment. This is a special case of the Earth Mover's Distance (Rubner et al., 1998) which solves the transportation problem of transporting words from sentence x to sentence y.

$$\mathrm{STS_{wmd}}(x, y) = 1 - \mathrm{WMD}(x, y), \qquad \mathrm{WMD}(x, y) = \min \sum_{u=1}^{n} \sum_{v=1}^{n} \mathcal{A}_{uv} \psi(x_u, y_v) \tag{4}$$

$$\text{subject to}: \sum_{v=1}^{n} \mathcal{A}_{uv} = \frac{1}{|\boldsymbol{x}|} freq(x_u), \qquad \sum_{u=1}^{n} \mathcal{A}_{uv} = \frac{1}{|\boldsymbol{y}|} freq(y_v)$$

Here $\psi(x_u, y_v)$ denotes the dissimilarity (distance) between the two words x_u and y_v. We used the Euclidean distance to denote the word dissimilarity $\psi(x_u, y_v)$. Here, $\mathcal{A}_{uv}$ denotes a weighted matrix of flow from word x_u in the sentence x to word y_v in the sentence y, n denotes the vocabulary size, and $freq(x_u)$ denotes an occurrence frequency of the word x_u in the sentence x.

4 Experiments for Building a Monolingual Parallel Corpus for Text Simplification

We built a monolingual parallel corpus for text simplification by aligning sentences from a comparable corpus using sentence similarity, based on the alignment between word embeddings, and evaluated the effectiveness of the proposed method from the quality of the corpus. First, we evaluated the proposed method in binary classification of a sentence pair as parallel or nonparallel. Next, we built a monolingual parallel corpus for text simplification using the proposed sentence similarity measure, and evaluated it qualitatively. Finally, we trained text simplification models using the SMT framework on our corpus and on existing corpora, to compare their effectiveness.

4.1 Binary Classification between Parallel and Nonparallel Sentences

Hwang et al. (2015) built a benchmark dataset [2] for text simplification extracted from the English Wikipedia and Simple English Wikipedia. They defined four labels: *Good (G) ("The semantics of the sentences completely match, possibly with small omissions.")*, *Good Partial (GP) ("A sentence completely covers the other sentence, but contains an additional clause or phrase that has information which is not contained within the other sentence.")*, *Partial ("The sentences discuss unrelated concepts, but share a short related phrase that does not match considerably.")*, and *Bad ("The sentences discuss unrelated concepts.")*. They annotated 67,853 sentence pairs (277 *G*, 281 *GP*, 117 *Partial*, and 67,178 *Bad*). We classified a sentence pair as parallel or nonparallel using this benchmark dataset to evaluate the sentence similarity measures. We conducted experiments in two settings: a setup (*G vs. O*), in which only sentence pairs labeled *G* were defined as parallel, and the other setup (*G + GP vs. O*), in which sentence pairs labeled either *G* or *GP* were defined as parallel. We evaluated the performance of the binary classification using two measures, the maximum F1 score (MaxF1) and the area under the curve (AUC).

Noise in the word alignment for *average alignment*, *maximum alignment*, and *hungarian alignment* was removed by aligning only those word pairs (x_i, y_j) which had a word similarity $\phi(x_i, y_j) > \theta$. This threshold θ was tuned to maximize MaxF1. We employed 0.89 and 0.95 in the binary classification of *G*

Method		*G vs. O*		*G + GP vs. O*	
		MaxF1	AUC	MaxF1	AUC
Zhu et al.	(Hwang et al., 2015)	0.550	0.509	0.431	0.391
Coster and Kauchak	(Hwang et al., 2015)	0.564	0.495	0.415	0.387
Hwang et al.	(Hwang et al., 2015)	0.712	0.694	0.607	0.529
Additive embeddings		0.691	0.695	0.518	0.487
Average alignment		0.419	0.312	0.391	0.297
Maximum alignment		0.717	0.730	**0.638**	**0.618**
Hungarian alignment		0.524	0.414	0.354	0.275
Word Mover's Distance		**0.724**	**0.738**	0.531	0.499

Table 1: Binary classification accuracy of parallel and nonparallel sentences. *Good (G) vs. Others (O)* is defined for sentence pairs where the label *G* denotes parallel. *G + Good Partial (GP) vs. O* regards the label *GP* as parallel in addition to the label *G*. The labels *G* and *GP* refer to bi- and uni-directional entailment, respectively.

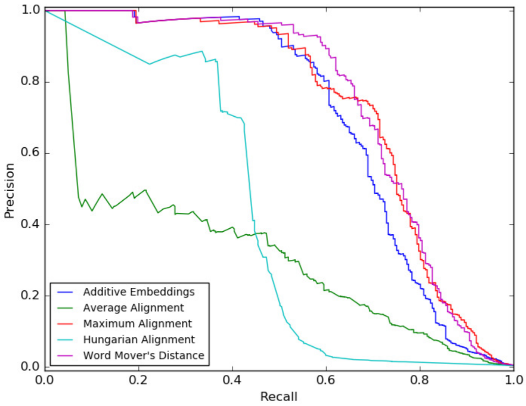


Figure 3: PR curves in binary classification of *G* and *O*.

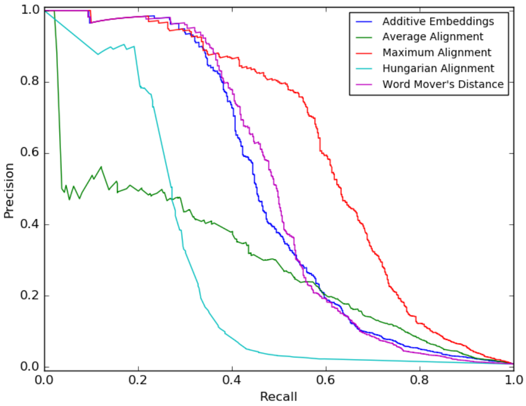


Figure 4: PR curves in binary classification of *G + GP* and *O*.

vs. O and *G + GP vs. O* for *average alignment*, 0.28 and 0.49 in the binary classification of *G vs. O* and *G + GP vs. O* for *maximum alignment*, and 0.98 in the binary classification of *G vs. O* and *G + GP vs. O* for *hungarian alignment*.

Table 1 compares sentence similarity measures in the binary parallel and nonparallel classification task. The top three methods in the upper row are taken from previous studies of monolingual parallel corpus construction for text simplification, and the five methods in the lower rows are the sentence similarity measures based on the word embeddings. *Additive embeddings* provides yet another baseline method, in which sentence embeddings are composed by adding word embeddings without word alignment, and sentence similarity is computed using the cosine similarity between sentence embeddings. We used publicly available [5] pretrained word embeddings to compute sentence similarity. From Table 1, it can be seen that *Word Mover's Distance* performed best in the binary classification task between *G vs. O*, whereas *maximum alignment* performed best in the binary classification task between *G + GP vs. O*.

Figures 3 and 4 show the Precision-Recall curves in the binary classification task between parallel and nonparallel sentences. Figure 4 shows that *maximum alignment* performed better than the other sentence similarity measures based on word embeddings, in the binary classification between *G + GP vs. O*.

Text simplification must take account not only of paraphrases from a complex expression to a simple expression but also of the deletion of unimportant parts of a complex sentence. It is therefore important to include both *G* sentence pairs, where the simple sentence is synonymous with the complex sentence, and *GP* sentence pairs, where the complex sentence entails the simple sentence. For this reason, *maximum alignment*, which performed best in classification between *G + GP vs. O*, was the preferred measure for

[5] `https://code.google.com/archive/p/word2vec/`

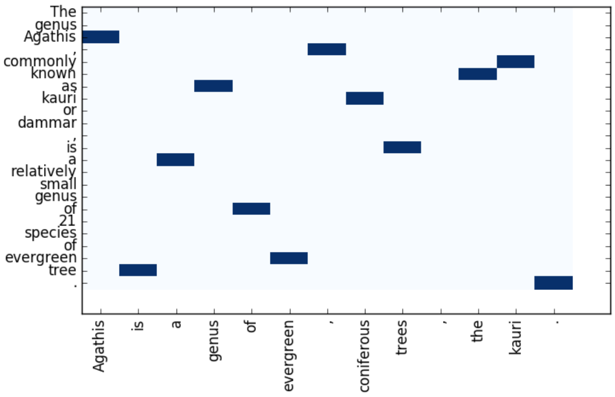


Figure 5: Hungarian word alignment matrix. A vertical axis indicates a sentence from English Wikipedia. A horizontal axis indicates a sentence from Simple English Wikipedia.

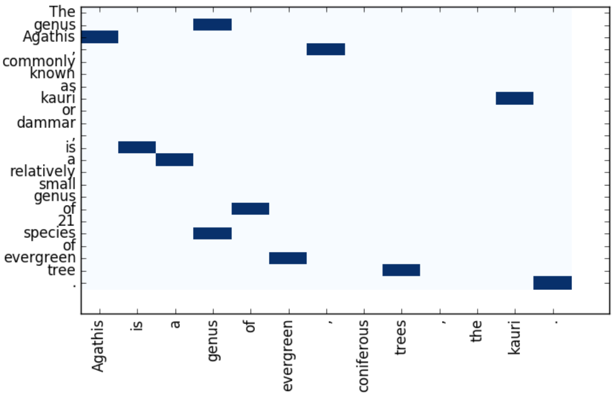


Figure 6: Maximum word alignment matrix. A vertical axis indicates a sentence from English Wikipedia. A horizontal axis indicates a sentence from Simple English Wikipedia.

computing sentence similarity in text simplification.

The experimental results demonstrate the effectiveness of *maximum alignment* in text simplification tasks, but why *maximum alignment* is the best? We present two illustrative figures to explain the reason. First, in *hungarian alignment* (Figure 5), false word alignments such as “as, genus,” “tree, is,” and “commonly, kauri” are found because of the restriction of one-to-one word alignment on the whole. Second, in *maximum alignment* (Figure 6), correct word alignments such as “genus, genus,” “species, genus,” “tree, trees,” and “kauri, kauri” are found because many-to-one word alignment is searched greedily. It may identify ambiguous pairs such as “genus, genus” and “species, genus,” but symmetrization of many-to-one alignment succeeds in reducing this type of noisy alignment. The restriction of *hungarian alignment* is too strict to correctly align content words between the sentences since even function words need to be aligned one-by-one. Also, in text simplification tasks, many-to-one alignment is more appropriate than one-to-one alignment because paraphrase between a phrase and a word occurs frequently.

4.2 Building an English Text Simplification Corpus

We built a monolingual parallel corpus for text simplification from English Wikipedia (normal) [6] and Simple English Wikipedia (simple) [7] using the *maximum alignment* that performed best in the previous experiment. First, we paired articles from the normal and simple editions by an exact match of titles, obtaining 126,725 article pairs. Sentence extraction using WikiExtractor [8] and tokenization using NLTK 3.2.1 [9] gave an average number of words per sentence of 25.1 for the normal articles and 16.9 for the simple articles. The average numbers of sentences per article were 57.7 and 7.65, respectively.

We computed the sentence similarity of all pairings of normal and simple sentences using *maximum alignment*. We based the threshold for word similarity and sentence similarity on the experimental results shown in Table 1. We aligned only those word pairs with a word similarity equal to or greater than 0.49, and aligned only those sentence pairs with a sentence similarity equal to or greater than 0.53. As a result, we obtained 492,993 sentence pairs from 126,725 article pairs.

Table 2 shows examples from the monolingual parallel corpus for text simplification with sentence similarity. We found synonymous expressions (purchased → bought) in sentence pairs with a similarity greater than 0.9 and deletion of unimportant parts of a sentence (such as ...) in sentence pairs with a similarity equal to or greater than 0.7. We also found sentence pairs with only a few words in common with a similarity less than 0.7.

[6] `https://dumps.wikimedia.org/enwiki/20160501/`

[7] `https://dumps.wikimedia.org/simplewiki/20160501/`

[8] `https://github.com/attardi/wikiextractor/`

[9] `http://www.nltk.org/`

similarity	normal	simple
0.9	Woody Bay Station was **purchased** by the Lynton and Barnstaple Railway Company in 1995 and, after much effort, a short section of railway reopened to passengers in 2004.	Woody Bay Station was **bought** by the Lynton and Barnstaple Railway Company in 1995 and, after much effort, a short section of railway reopened to passengers in 2004.
0.8	This work continued with the 1947 paper "Types of polyploids: their classification and significance", which **detailed a system for the classification of polyploids and** described Stebbins' ideas about the role of paleopolyploidy in angiosperm evolution**, where he argued that chromosome number may be a useful tool for the construction of phylogenies**.	This work continued with the 1947 paper "Types of polyploids: their classification and significance", which described Stebbins' ideas about the role of paleopolyploidy in angiosperm evolution.
0.7	Mir **has been** a significant influence on late 20th-century art, in particular the American abstract expressionist artists **such as Motherwell, Calder, Gorky, Pollock, Matta and Rothko, while his lyrical abstractions and color field paintings were precursors of that style by artists such as Frankenthaler, Olitski and Louis and others**.	Mir **was** a significant influence on late 20th-century art, in particular the American abstract expressionist artists.
0.6	**The couple** has **four children:**	**She** has **two daughters and two sons.**
0.5	Ithaca is **in the rural Finger Lakes region about northwest of New York City; the nearest larger cities, Binghamton and Syracuse, are an hour's drive away by car, Rochester and Scranton are two hours, Buffalo and Albany are three.**	Ithaca is **a city in upstate New York, America.**

Table 2: Examples from our text simplification corpus ranked by similarity.

4.3 English Text Simplification

We trained SMT-based text simplification models using our corpus and existing text simplification corpora (Zhu et al., 2010; Coster and Kauchak, 2011b; Hwang et al., 2015). The results were compared to evaluate the effectiveness of our text simplification corpus. We treated text simplification as a translation problem from the normal sentence to the simple one and modeled it using a phrase-based SMT trained as a log linear model. In each corpus, we randomly sampled 500 sentence pairs for tuning with MERT (Och, 2003) and used the remainder for training. Moses was used as the phrase-based SMT tool. We employed GIZA++ (Och and Ney, 2003) to obtain the word alignment, and KenLM (Heafield, 2011) to build the 5-gram language model from the entire Simple English Wikipedia [7]. As test data, we used 277 sentence pairs labeled *G* and 281 sentence pairs labeled *G* + *GP* from the Hwang et al. (2015) dataset and evaluated the accuracy using BLEU.

Table 3 shows the number of sentences, range of vocabulary, average number of words per sentence, and BLEU scores of the text simplification models trained on each corpus. The text simplification model trained on our corpus achieved the best BLEU score. To compare the learning curves of our corpus with that from Hwang et al. (2015), we recorded the BLEU scores while changing the corpus size. We discovered that the difference in performance was not only due to the corpus size, as the BLEU scores of the model trained on our corpus remained higher than those of the model trained on the Hwang et al. (2015) corpus at all corpus sizes. The model trained on the Coster and Kauchak (2011b) corpus performed slightly better than that trained on our corpus in 100,000 sentence pairs of a *G* (bi-directional entailment) test set; however, in a *G* + *GP* (uni-directional entailment) test set that requires more various

Text simplification corpus	#sents.	#vocabulary		Avg. #words per sent.		BLEU	
		normal	simple	normal	simple	*G*	*G* + *GP*
Baseline (None)						42.1	22.3
Zhu et al.	100,000	173,463	143,030	21.2	17.4	41.8	22.1
Zhu et al. (All)	107,516	181,459	149,643	21.2	17.4	42.0	22.1
Coster and Kauchak	100,000	112,744	102,418	23.7	21.1	43.8	23.4
Coster and Kauchak (All)	136,862	132,567	120,620	23.6	21.1	44.3	23.8
Hwang et al.	100,000	117,474	103,427	25.3	21.2	42.9	22.7
Hwang et al. (G)	154,305	152,419	133,825	25.2	21.2	42.9	22.7
Hwang et al.	200,000	175,416	145,773	25.6	20.5	43.1	22.7
Hwang et al. (G + GP)	284,238	212,138	164,979	26.0	19.8	43.9	23.1
Hwang et al.	300,000	217,699	167,945	26.1	19.7	42.9	22.7
Hwang et al. (All)	391,116	248,510	184,521	26.5	19.4	43.1	22.8
Ours	100,000	122,390	112,670	23.9	21.8	43.2	23.6
Ours	200,000	180,776	151,815	24.7	20.1	45.7	24.8
Ours	300,000	219,628	174,576	25.2	19.0	46.4	25.3
Ours (All)	492,493	274,775	198,043	25.3	17.9	**47.5**	**26.3**

Table 3: SMT-based English text simplification performance. *Baseline* does not do any simplification.

Input	Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart was more likely to use clarinets in A major than in any other key besides E-flat major.
Reference	Mozart used clarinets in A major often.
Zhu et al.	Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart **which he** more likely to use clarinets in A major than in any other key besides E-flat major.
Coster and Kauchak	**Mozart was** Clarinet Concerto and Clarinet Quintet are both in A major, and **Mozart used clarinets in A major often**.
Hwang et al.	Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart was more likely to use clarinets in A major than in any other key besides E-flat major.
Ours	Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and **Mozart used clarinets in A major often**.

Table 4: Examples of text simplification trained on different text simplification corpora.

substitution and phrasal definition, the model trained on our corpus performed slightly better than their corpus.

Our corpus gave a larger difference in the average number of words between normal and simple sentences than the other corpora, with values closer to the average numbers of words per sentence in the entire Wikipedia (25.1 and 16.9, respectively). This suggests that *maximum alignment* was able to compute sentence similarity more accurately than the other measures regardless of the sentence length.

Table 4 shows examples of text simplification trained on different text simplification corpora. The model trained on our corpus generated a simple sentence that appropriately entailed the reference. The model trained on the Coster and Kauchak (2011b) corpus simplified the input sentence appropriately but also performed incorrect substitutions and generated an ungrammatical sentence. The model trained on the *G* + *GP* part of the Hwang et al. (2015) corpus did not rewrite the input sentence. The model trained on the Zhu et al. (2010) corpus used almost the same amount of Coster and Kauchak (2011b)'s corpus but performed incorrect substitutions and generated an ungrammatical sentence. Coster and Kauchak (2011b) extended Zhu et al. (2010)'s work by considering the order of sentences. In a Wikipedia article, sentences are arranged in a characteristic order, e.g., a definition sentence appears in the first sentence. Therefore, they may obtain similar sentence pairs effectively. In contrast, our simple proposed method achieved equivalent or higher performance than their method without considering any ordering of sentences.

5 Conclusions

We proposed an unsupervised method for building a monolingual parallel corpus for text simplification. Four types of sentence similarity metric were proposed, based on alignment between word embeddings. Experimental results demonstrated the effectiveness of the sentence similarity measure using many-to-one word alignment to align each word in the complex sentence with the most similar word in the simple sentence. Our proposed method achieved state-of-the-art performance in both an intrinsic evaluation based on classifying sentence pairs from the English Wikipedia and Simple English Wikipedia into a parallel and nonparallel data, and in an extrinsic evaluation in which a complex sentence was translated into a simple sentence.

We successfully built an English monolingual parallel corpus for text simplification from comparable corpus with different levels of difficulty. However, such large-scale comparable corpus is unavailable in many languages. In future work, we will build a monolingual parallel corpus from a raw corpus by combining our sentence similarity measure with a readability assessment. Specifically, we will divide the raw corpus into complex and simple corpora based on readability, and use the paired corpora to align complex sentences with simple ones. This approach should be applicable to any language, and open new opportunities in the field of text simplification.

Acknowledgements

This research was (partly) supported by Grant-in-Aid for Research on Priority Areas, Tokyo Metropolitan University, Research on social bigdata.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393, Montréal, Canada.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 81–91, Dublin, Ireland.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, USA.

Stefan Bott and Horacio Saggion. 2011. An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 20–26, Portland, Oregon, USA.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and Annotation of the First Italian Corpus for Text Simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA.

Helena M. Caseli, Tiago F. Pereira, Lucia Specia, Thiago A.S. Pardo, Caroline Gasperin, and Sandra M. Aluísio. 2009. Building a Brazilian Portuguese Parallel Corpus of Original and Simplified Texts. In *Advances in Computational Linguistics, Research in Computer Science*, pages 59–70, Mexico City, Mexico.

William Coster and David Kauchak. 2011a. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Portland, Oregon, USA.

William Coster and David Kauchak. 2011b. Simple English Wikipedia: A New Text Simplification Task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA.

Isao Goto, Hideki Tanaka, and Tadashi Kumano. 2015. Japanese News Simplification: Task Design, Data Set Construction, and Analysis of Simplified Text. In *Proceedings of MT Summit XV*, pages 17–31, Miami, Florida, USA.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado, USA.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria.

Sigrid Klerke and Anders Søgaard. 2012. DSim, a Danish Parallel Corpus for Text Simplification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 4015–4018, Istanbul, Turkey.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Harold W. Kuhn. 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2:83–97.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 957–966, Lille, France.

Tomas Mikolov, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at the International Conference on Learning Representations*, Scottsdale, Arizona, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, Nevada, USA.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 1998. A Metric for Distributions with Applications to Image Databases. In *Proceedings of the Sixth International Conference on Computer Vision*, pages 59–66, Washington, DC, USA.

Yangqiu Song and Dan Roth. 2015. Unsupervised Sparse Vector Densification for Short Text Similarity. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1275–1280, Denver, Colorado, USA.

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. *Lecture Notes in Computer Science*, 6001:30–39.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 148–153, Denver, Colorado, USA.

Sanja Štajner, Hannah Bechara, and Horacio Saggion. 2015a. A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 823–828, Beijing, China.

Sanja Štajner, Iacer Calixto, and Horacio Saggion. 2015b. Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 618–626, Hissar, Bulgaria.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence Simplification by Monolingual Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 1015–1024, Jeju Island, Korea.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1353–1361, Beijing, China.