Large-Scale Hierarchical Alignment for Data-driven Text Rewriting

Nikola I. Nikolov and **Richard H.R. Hahnloser**
Institute of Neuroinformatics, University of Zürich and ETH Zürich, Switzerland
{niniko, rich}@ini.ethz.ch

Abstract

We propose a simple unsupervised method for extracting pseudo-parallel monolingual sentence pairs from comparable corpora representative of two different text styles, such as news articles and scientific papers. Our approach does not require a seed parallel corpus, but instead relies solely on hierarchical search over pre-trained embeddings of documents and sentences. We demonstrate the effectiveness of our method through automatic and extrinsic evaluation on text simplification from the normal to the Simple Wikipedia. We show that pseudo-parallel sentences extracted with our method not only supplement existing parallel data, but can even lead to competitive performance on their own.[1]

1 Introduction

Parallel corpora are indispensable resources for advancing monolingual and multilingual text rewriting tasks. Due to the scarce availability of parallel corpora, and the cost of manual creation, a number of methods have been proposed that can perform large-scale sentence alignment: automatic extraction of *pseudo-parallel* sentence pairs from raw, comparable[2] corpora. While pseudo-parallel data is beneficial for machine translation (Munteanu and Marcu, 2005), there has been little work on large-scale sentence alignment for monolingual text-to-text rewriting tasks, such as simplification (Nisioi et al., 2017) or style transfer (Liu et al., 2016). Furthermore, the majority of existing methods (e.g. Marie and Fujita (2017); Grégoire and Langlais (2018)) assume access to some parallel training data. This impedes their application to cases where there is *no parallel data available whatsoever*, which is the case for the majority of text rewriting tasks, such as style transfer.

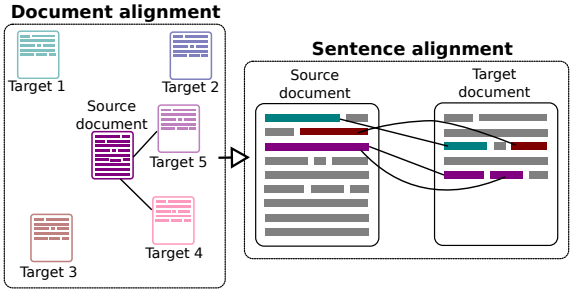

Figure 1: Illustration of large-scale hierarchical alignment (**LHA**). For each document in a *source* dataset, **document alignment** retrieves matching documents from a *target* dataset. In turn, **sentence alignment** retrieves matching sentence pairs from within each document pair.

In this paper, we propose a simple unsupervised method, Large-scale Hierarchical Alignment (**LHA**) (Figure 1; Section 3), for extracting pseudo-parallel sentence pairs from two raw monolingual corpora which contain documents in two different author styles, such as scientific papers and press releases. LHA hierarchically searches for document and sentence nearest neighbors within the two corpora, extracting sentence pairs that have high semantic similarity, yet preserve the stylistic characteristics representative of their original datasets. LHA is robust to noise, fast and memory efficient, enabling its application to datasets on the order of hundreds of millions of sentences. Its generality makes it relevant to a wide range of monolingual text rewriting tasks.

We demonstrate the effectiveness of LHA on automatic benchmarks for alignment (Section 4), as well as extrinsically, by training neural machine translation (NMT) systems on the task of text simplification from the normal Wikipedia to the Simple Wikipedia (Section 5). We show that

[1] Code available at https://github.com/ninikolov/lha.

[2] Corpora that contain documents on similar topics.

Proceedings of Recent Advances in Natural Language Processing, pages 844–853,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_098

pseudo-parallel datasets obtained by LHA are not only useful for augmenting existing parallel data, boosting the performance on automatic measures, but can even be competitive on their own.

2 Background

2.1 Data-Driven Text Rewriting

The goal of text rewriting is to transform an input text to satisfy specific constraints, such as simplicity (Nisioi et al., 2017) or a more general author style, such as political (e.g. democratic to republican) or gender (e.g. male to female) (Prabhumoye et al., 2018; Shen et al., 2017). Rewriting systems can be valuable when preparing a text for multiple audiences, such as simplification for language learners (Siddharthan, 2002) or people with reading disabilities (Inui et al., 2003). They can also be used to improve the accessibility of technical documents, e.g. to simplify terms in clinical records for laymen (Abrahamsson et al., 2014).

Text rewriting can be cast as a data-driven task in which transformations are learned from large collections of parallel sentences. Limited availability of high-quality parallel data is a major bottleneck for this approach. Recent work on Wikipedia and the Simple Wikipedia (Coster and Kauchak, 2011; Kajiwara and Komachi, 2016) and on the Newsela dataset of simplified news articles for children (Xu et al., 2015) explore supervised, data-driven approaches to text simplification. Such approaches typically rely on statistical (Xu et al., 2016) or neural (Štajner and Nisioi, 2018) machine translation.

Recent work on unsupervised approaches to text rewriting without parallel corpora is based on variational (Fu et al., 2017) or cross-aligned (Shen et al., 2017) autoencoders that learn latent representations of content separate from style. In (Prabhumoye et al., 2018), authors model style transfer as a back-translation task by translating input sentences into an intermediate language. They use the translations to train separate English decoders for each target style by combining the decoder loss with the loss of a style classifier, separately trained to distinguish between the target styles.

2.2 Large-Scale Sentence Alignment

The goal of **sentence alignment** is to extract from raw corpora sentence pairs suitable as training examples for text-to-text rewriting tasks such as machine translation or text simplification. When the documents in the corpora are *parallel* (labelled document pairs, such as identical articles in two languages), the task is to identify suitable sentence pairs from each document. This problem has been extensively studied both in the multilingual (Brown et al., 1991; Moore, 2002) and monolingual (Hwang et al., 2015; Kajiwara and Komachi, 2016; Štajner et al., 2018) case. The limited availability of parallel corpora led to the development of **large-scale sentence alignment** methods, which is also the focus of this work. The aim of these methods is to extract pseudo-parallel sentence pairs from raw, non-aligned corpora. For many tasks, millions of examples occur naturally within existing textual resources, amply available on the internet.

The majority of previous work on large-scale sentence alignment is in machine translation, where adding pseudo-parallel pairs to an existing parallel dataset has been shown to boost the translation performance (Munteanu and Marcu, 2005; Uszkoreit et al., 2010). The work that is most closely related to ours is (Marie and Fujita, 2017), where authors use pre-trained word and sentence embeddings to extract rough translation pairs in two languages. Subsequently, they filter out low-quality translations using a classifier trained on parallel translation data. More recently, (Grégoire and Langlais, 2018) extract pseudo-parallel translation pairs using a Recurrent Neural Network (RNN) classifier. Importantly, these methods assume that *some parallel training data is already available*, which impedes their application in settings where there is no parallel data whatsoever, which is the case for many text rewriting tasks such as style transfer.

There is little work on large-scale sentence alignment focusing specifically on monolingual tasks. In (Barzilay and Elhadad, 2003), authors develop a hierarchical alignment approach of first clustering paragraphs on similar topics before performing alignment on the sentence level. They argue that, for monolingual data, pre-clustering of larger textual units is more robust to noise compared to fine-grained sentence matching applied directly on the dataset level.

3 Large-Scale Hierarchical Alignment (LHA)

Given two datasets that contain comparable documents written in two different author styles: a

source dataset $\mathbf{S^d}$ consisting of N_S documents $\mathbf{S^d} = \{s_1^d, ..., s_{N_S}^d\}$ (e.g. all Wikipedia articles) and a **target** dataset $\mathbf{T^d}$ consisting of N_T documents $\mathbf{T^d} = \{t_1^d, ..., t_{N_T}^d\}$ (e.g. all articles from the Simple Wikipedia), our approach to large-scale alignment is hierarchical, consisting of two consecutive steps: **document** alignment followed by **sentence** alignment (see Figure 1).

3.1 Document Alignment

For each source document s_i^d, document alignment retrieves K nearest neighbours $\{t_{i_1}^d, ..., t_{i_K}^d\}$ from the target dataset. In combination, these form K pseudo-parallel document pairs $\{(s_i^d, t_{i_1}^d), ..., (s_i^d, t_{i_K}^d)\}$. Our aim is to select document pairs with high semantic similarity, potentially containing good pseudo-parallel sentence pairs representative of the document styles of each dataset.

To find nearest neighbours, we rely on two components: document embedding and approximate nearest neighbour search. For each dataset, we pre-compute document embeddings $e_d()$ as $I_s = [e_d(s_1^d), ..., e_d(s_{N_S}^d)]$ and $I_t = [e_d(t_1^d), ..., e_d(t_{N_T}^d)]$. We employ nearest neighbour search methods[3] to partition the embedding space, enabling fast and efficient nearest neighbour retrieval of similar documents across I_s and I_t. This enables us to find K nearest *target* document embeddings in I_t for each *source* embedding in I_s. We additionally filter document pairs whose similarity is below a manually selected threshold θ_d. In Section 4, we evaluate a range of different document embedding approaches, as well as alternative similarity metrics.

3.2 Sentence Alignment

Given a pseudo-parallel document pair (s^d, t^d) that contains a **source** document $s^d = \{s_1^s, ..., s_{N_J}^s\}$ consisting of N_J sentences and a **target** document $t^d = \{t_1^s, ..., t_{N_M}^s\}$ consisting of N_M sentences, sentence alignment extracts pseudo-parallel sentence pairs (s_i^s, t_j^s) that are highly similar.

To implement sentence alignment, we first embed each sentence in s^d and t^d and compute an inter-sentence similarity matrix P among all sentence pairs in s^d and t^d. From P we extract K nearest neighbours for each source and each target sentence. We denote the nearest neighbours of s_i^s as $NN(s_i^s) = \{t_{i_1}^s, \ldots, t_{i_K}^s\}$ and the nearest neighbours of t_j^s as $NN(t_j^s) = \{s_{j_1}^s, \ldots, s_{j_K}^s\}$. We remove all sentence pairs with similarity below a manually set threshold θ_s. We then merge *all* overlapping sets of nearest sentences in the documents to produce pseudo-parallel sentence sets (e.g. $(\{s_e^s, s_i^s\}, \{t_j^s, t_k^s, t_l^s\})$ when source sentence i is closes to target sentences j, k, and l and target sentence j is closest to source sentences e and i). This approach, inspired from (Štajner et al., 2018), provides the flexibility to model multi-sentence interactions, such as sentence splitting or compression, as well as individual sentence-to-sentence reformulations. Note that when $K = 1$, we only retrieve individual sentence pairs.

The final output of sentence alignment is a list of pseudo-parallel sentence pairs with high semantic similarity and preserved stylistic characteristics of each dataset. The pseudo-parallel pairs can be used to either augment an existing parallel dataset (as in Section 5), or independently, to solve a new author style transfer task for which there is no parallel data available (see the supplementary material for an example).

3.3 System Variants

The aforementioned framework provides the flexibility of exploring diverse variants, by exchanging document/sentence embeddings or text similarity metrics. We compare all variants in an automatic evaluation in Section 4.

Text embeddings We experiment with four text embedding methods:

1. *Avg*, is the average of the constituent word embeddings of a text[4], a simple approach that has proved to be a strong baseline for many text similarity tasks.

2. In *Sent2Vec*[5] (Pagliardini et al., 2018), the word embeddings are specifically optimized towards additive combinations over the sentence using an unsupervised objective function. This approach performs well on many unsupervised and supervised text similarity tasks, often outperforming more sophisticated supervised recurrent or convolutional architectures, while remaining very fast to compute.

[3] We use the Annoy library https://github.com/spotify/annoy.

[4] We use the Google News 300-dim Word2Vec models.

[5] We use the public unigram Wikipedia model.

3. *InferSent*[6] (Conneau et al., 2017) is a supervised sentence embedding approach based on bidirectional LSTMs, trained on natural language inference data.

4. *BERT*[7] (Devlin et al., 2019) is a state-of-the-art supervised sentence embedding approach based on the Transformer architecture.

Word Similarity We additionally test four word-based approaches for computing text similarity. Those can be used either on their own, or to refine the nearest neighbour search across documents or sentences.

1. We compute the unigram string overlap $o(\mathbf{x}, \mathbf{y}) = \frac{|\{\mathbf{y}\} \cap \{\mathbf{x}\}|}{|\{\mathbf{y}\}|}$ between source tokens $\mathbf{x}$ and target tokens $\mathbf{y}$ (excluding punctuation, numbers and stopwords).

2. We use the *BM25* ranking function (Robertson et al., 2009), an extension of TF-IDF.

3. We use the Word Mover's Distance (*WMD*) (Kusner et al., 2015), which measures the distance the embedded words of one document need to travel to reach the embedded words of another document. WMD has recently achieved good results on text retrieval (Kusner et al., 2015) and sentence alignment (Kajiwara and Komachi, 2016).

4. We use the Relaxed Word Mover's Distance (*RWMD*) (Kusner et al., 2015), which is a fast approximation of the WMD.

4 Automatic Evaluation

We perform an automatic evaluation of LHA using an annotated sentence alignment dataset (Hwang et al., 2015). The dataset contains 46 article pairs from Wikipedia and the Simple Wikipedia. The 67k potential sentence pairs were manually labelled as either *good* simplifications (277 pairs), *good* with a *partial* overlap (281 pairs), *partial* (117 pairs) or *non-valid*. We perform three comparisons using this dataset: evaluating document and sentence alignment separately, as well as jointly.

For sentence alignment, the task is to retrieve the 277 good sentence pairs out of the $67k$ possible sentence pairs in total, while minimizing the number of false positives. To evaluate document alignment, we add 1000 randomly sampled articles from Wikipedia and the Simple Wikipedia as noise, resulting in 1046 article pairs in total. The goal of document alignment is to identify the original 46 document pairs out of 1046×1046 possible document combinations.

This set-up additionally enables us to **jointly** evaluate document and sentence alignment, which best resembles the target effort of retrieving good sentence pairs from noisy documents. The two aims of the joint alignment task are to identify the *good* sentence pairs from within either $1M$ document or $125M$ sentence pairs, in the latter case without relying on any document-level information whatsoever.

4.1 Results

Our results are summarized in Tables 1 and 2. For all experiments, we set $K = 1$ and report the maximum F1 score ($\mathbf{F1}_{max}$) obtained from varying the document threshold θ_d and the sentence threshold θ_s. We also report the percentage of true positive ($\mathbf{TP}$) document or sentence pairs that were retrieved when the F1 score was at its maximum, as well as the average speed of each approach (**doc/s** and **sent/s**). The speed becomes of a particular concern when working with large datasets consisting of millions of documents and hundreds of millions of sentences.

On document alignment, (Table 1, left) the *Sent2Vec* approach achieved the best score, outperforming the other embedding methods including the word-based similarity measures. On sentence alignment (Table 1, right), the WMD achieves the best performance, matching the result from (Kajiwara and Komachi, 2016). When evaluating document and sentence alignment jointly (Table 2), we compare our hierarchical approach (*LHA*) to global alignment applied directly on the sentence level (*Global*). *Global* computes the similarities between all $125M$ sentence pairs in the entire evaluation dataset. LHA significantly outperforms Global, successfully retrieving three times more valid sentence pairs, while remaining fast to compute. This result demonstrates that document alignment is beneficial, successfully filtering some of the noise, while also reducing the overall number of sentence similarities to be computed.

The *Sent2Vec* approach to LHA achieves good performance on document and sentence align-

[6] We use the GloVe-based model provided by the authors.
[7] We use the base 12-layer model provided by the authors.

Table 1: Automatic evaluation of Document (left) and Sentence alignment (right). **EDim** is the embedding dimensionality. **TP** is the percentage of true positives obtained at $\mathbf{F1}_{max}$. Speed is calculated on a single CPU thread.

			Document alignment				Sentence alignment			
	Approach	**EDim**	$\mathbf{F1}_{max}$	**TP**	θ_d	**doc/s**	$\mathbf{F1}_{max}$	**TP**	θ_s	**sent/s**
Embedding	Average word embeddings (Avg)	300	0.66	43%	0.69	260	0.675	46%	0.82	1458
	Sent2Vec (Pagliardini et al., 2018)	600	**0.78**	**61%**	0.62	343	0.692	48%	0.69	1710
	InferSent† (Conneau et al., 2017)	4096	-	-	-	-	0.69	49%	0.88	110
	BERT† (Devlin et al., 2019)	768	-	-	-	-	0.65	43%	0.89	25
Word sim	Overlap	-	0.53	29%	0.66	120	0.63	40%	0.5	1600
	BM25 (Robertson et al., 2009)	-	0.46	16%	0.257	60	0.52	27%	0.43	20K
	RWMD (Kusner et al., 2015)	300	0.713	51%	0.67	60	0.704	50%	0.379	1050
	WMD (Kusner et al., 2015)	300	0.49	24%	0.3	1.5	**0.726**	**54%**	0.353	180
	(Hwang et al., 2015)	-	-	-	-	-	0.712	-	-	-
	(Kajiwara and Komachi, 2016)	-	-	-	-	-	0.724	-	-	-

†: These models are specifically designed for sentence embedding, hence we do not test them on document alignment.

Table 2: Evaluation on large-scale sentence alignment: identifying the *good* sentence pairs without any document-level information. We pre-compute the embeddings and use the Annoy ANN library. For the WMD-based approaches, we re-compute the top 50 sentence nearest neighbours of Sent2Vec.

Approach	$\mathbf{F1}_{max}$	**TP**	**time**
LHA (Sent2Vec)	0.54	31%	33s
LHA (Sent2Vec + WMD)	**0.57**	**33%**	1m45s
Global (Sent2Vec)	0.339	12%	15s
Global (WMD)	0.291	12%	30m45s

ment, while also being the fastest to compute. We therefore use it as the default approach for the following experiments on text simplification.

5 Empirical Evaluation

To test the suitability of pseudo-parallel data extracted with LHA, we perform empirical experiments on text simplification from the normal Wikipedia to the Simple Wikipedia. We chose simplification because some parallel data are already available for this task, allowing us to experiment with mixing parallel and pseudo-parallel datasets. In the supplementary material[8] we experiment with an additional task for which there is no parallel data: style transfer from scientific journal articles to press releases.

We compare the performance of neural machine translation (NMT) systems trained under three different scenarios: 1) using **existing** parallel data for training; 2) using a **mixture** of parallel and pseudo-parallel data extracted with LHA; and 3) using pseudo-parallel data **on its own**.

5.1 Experimental Setup

NMT model For all experiments, we use a single-layer LSTM encoder-decoder model (Cho et al., 2015) with an attention mechanism (Bahdanau et al., 2015). We train our models on the subword level (Sennrich et al., 2015), capping the vocabulary size to 50k. We re-learn the subword rules separately for each dataset, and train until convergence using the Adam optimizer (Kingma and Ba, 2015). We use beam search with a beam of 5 to generate all final outputs.

Evaluation metrics We report a diverse range of automatic metrics and statistics. *SARI* (Xu et al., 2016) is a recently proposed metric for text simplification which correlates well with simplicity in the output. SARI takes into account the total number of changes (additions, deletions) of the input when scoring model outputs. *BLEU* (Papineni et al., 2002) is a precision-based metric for machine translation commonly used for evaluation of text simplification (Xu et al., 2016; Štajner and Nisioi, 2018) and of style transfer (Shen et al., 2017). Recent work has indicated that BLEU is not suitable for assessment of simplicity (Sulem et al., 2018), it correlates better with meaning preservation and grammaticality, in particular when using multiple references. We also report the average Levenshtein distance (LD) from the model outputs to the input (LD_{src}) or the target reference (LD_{tgt}). On simplification tasks, LD correlates well with meaning preservation and grammaticality (Sulem et al., 2018), complementing BLEU.

Extracting pseudo-parallel data We use LHA with *Sent2Vec* (see Section 3) to extract pseudo-parallel sentence pairs for text simplification. To ensure some degree of lexical similarity, we exclude pairs whose string overlap (defined in Section 3.3) is below 0.4, and pairs in which the target sentence is more than 1.5 times longer than the source sentence. We use $K = 5$ in all of our align-

[8] Available in our arXiv paper at https://arxiv.org/abs/1810.08237

Table 3: Datasets used to extract pseudo-parallel monolingual sentence pairs in our experiments.

Dataset	Type	Documents	Tokens	Sentences	Tok. per sent.	Sent. per doc.
Wikipedia	Articles	5.5M	2.2B	92M	25 ± 16	17 ± 32
Simple Wikipedia	Articles	134K	62M	2.9M	27 ± 68	22 ± 34
Gigaword	News	8.6M	2.5B	91M	28 ± 12	11 ± 7

Table 4: Example pseudo-parallel pairs extracted by our Large-scale hierarchical alignment (LHA) method.

Dataset	Source	Target
wiki-simp-65	However, Jimmy Wales, Wikipedia's co-founder, denied that this was a crisis or that Wikipedia was running out of admins, saying, "The number of admins has been stable for about two years, there's really nothing going on."	But the co-founder Wikipedia, Jimmy Wales, did not believe that this was a crisis. He also did not believe Wikipedia was running out of admins.
wiki-news-74	Prior to World War II, Japan's industrialized economy was dominated by four major zaibatsu: Mitsubishi, Sumitomo, Yasuda and Mitsui.	Until Japan 's defeat in World War II , the economy was dominated by four conglomerates , known as " zaibatsu " in Japanese . These were the Mitsui , Mitsubishi , Sumitomo and Yasuda groups .

Table 5: Statistics of the pseudo-parallel datasets extracted with LHA. μ^{src}_{tok} and μ^{tgt}_{tok} are the mean src/tgt token counts, while $\%^{src}_{s>2}$ and $\%^{tgt}_{s>2}$ report the percentage of items that contain more than one sentence.

Dataset	Pairs	μ^{src}_{tok}	μ^{tgt}_{tok}	$\%^{src}_{s>2}$	$\%^{tgt}_{s>2}$
wiki-simp-72	25K	26.72	22.83	16%	11%
wiki-simp-65	80K	23.37	15.41	17%	7%
wiki-news-74	133K	25.66	17.25	19%	2%
wiki-news-70	216K	26.62	16.29	19%	2%

ment experiments, which enables extraction of up to 5 sentence nearest neighbours.

Parallel data As a parallel baseline dataset, we use an existing dataset from (Hwang et al., 2015). The dataset consists of 282K sentence pairs obtained after aligning the parallel articles from Wikipedia and the Simple Wikipedia. This dataset allows us to compare our results to previous work on data-driven text simplification. We use two versions of the dataset in our experiments: `full` contains all 282K pairs, while `partial` contains 71K pairs, or 25% of the full dataset.

Evaluation data We evaluate our simplification models on the testing dataset from (Xu et al., 2016), which consists of 358 sentence pairs from the normal Wikipedia and the Simple Wikipedia. In addition to the ground truth simplifications, each input sentence comes with 8 additional references, manually simplified by Amazon Meachanical Turkers. We compute *BLEU* and *SARI* on the 8 manual references.

Pseudo-parallel data We align two dataset pairs, obtaining pseudo-parallel sentence pairs for text simplification (statistics of the datasets we use for alignment are in Table 3). First, we align the normal `Wikipedia` to the `Simple Wikipedia` using document and sentence similarity thresholds $\theta_d = 0.5$ and $\theta_s = \{0.72, 0.65\}$, producing two datasets: `wiki-simp-72` and `wiki-simp-65`. Because LHA uses no document-level information in this dataset, alignment leads to new sentence pairs, some of which may be distinct from the pairs present in the existing parallel dataset. We monitor for and exclude pairs that overlap with the testing dataset. Second, we align `Wikipedia` to the `Gigaword` news article corpus (Napoles et al., 2012), using $\theta_d = 0.5$ and $\theta_s = \{0.74, 0.7\}$, resulting in two additional pseudo-parallel datasets: `wiki-news-74` and `wiki-news-70`. With these datasets, we investigate whether pseudo-parallel data extracted from a *different domain* can be beneficial for text simplification. We use slightly higher sentence alignment thresholds for the news articles because of the domain difference.

We find that the majority of the pairs extracted contain a single sentence, and 15-20% of the source examples and 5-10% of the target examples contain multiple sentences (see Table 5 for additional statistics). Most multi-sentence examples contain two sentences, while 0.5-1% contain 3 to 5 sentences. Two example aligned outputs are in Table 4 (additional examples are available in the supplementary material). They suggest that our method is capable of extracting high-quality pairs that are similar in meaning, even spanning across multiple sentences.

Randomly sampled pairs We also experiment with adding random sentence pairs to the parallel dataset (`rand-100K`, `rand-200K` and `rand-300K` datasets, containing 100K, 200K and 300K random pairs, respectively). The

Table 6: Empirical results on text simplification from Wikipedia to the Simple Wikipedia. The highest SARI/BLEU results from each category are in bold. `input` and `reference` are not generated using Beam Search.

Method or Dataset	Total pairs (% pseudo)	Beam hypothesis 1					Beam hypothesis 2				
		SARI	BLEU	μ_{tok}	LD_{src}	LD_{tgt}	SARI	BLEU	μ_{tok}	LD_{src}	LD_{tgt}
input	-	26	99.37	22.7	0	0.26	-	-	-	-	-
reference	-	38.1	70.21	22.3	0.26	0	-	-	-	-	-
NTS	282K (0%)	30.54	84.69	-	-	-	35.78	77.57	-	-	-
Parallel + Pseudo-parallel or Randomly sampled data (Using full parallel dataset, 282K parallel pairs)											
baseline-282K	282K (0%)	30.72	85.71	18.3	0.18	0.37	36.16	82.64	19	0.19	0.36
+ wiki-simp-72	307K (8%)	30.2	87.12	19.43	0.14	0.34	36.02	81.13	19.03	0.19	0.36
+ wiki-simp-65	362K (22%)	**30.92**	**89.64**	19.8	0.13	0.33	36.48	83.56	19.37	0.18	0.35
+ wiki-news-74	414K (32%)	30.84	89.59	19.67	0.13	0.33	36.57	**83.85**	19.13	0.18	0.35
+ wiki-news-70	498K(43%)	30.82	89.62	19.6	0.13	0.33	36.45	83.11	18.98	0.19	0.36
+ rand-100K	382K (26%)	30.52	88.46	19.7	0.14	0.34	**36.96**	82.86	19	0.2	0.36
+ rand-200K	482K (41%)	29.47	80.65	19.3	0.18	0.36	34.36	74.67	18.93	0.23	0.38
+ rand-300K	582K (52%)	28.68	75.61	19.57	0.23	0.4	32.34	68.9	18.35	0.3	0.43
Parallel + Pseudo-parallel data (Using partial parallel dataset, 71K parallel pairs)											
baseline-71K	71K (0%)	**31.16**	69.53	17.45	0.29	0.44	32.92	67.29	19.14	0.3	0.44
+ wiki-simp-65	150K (52%)	31.0	**81.52**	18.26	0.21	0.38	**35.12**	**77.38**	18.16	0.25	0.39
+ wiki-news-70	286K(75%)	31.01	80.03	17.82	0.23	0.4	34.14	76.44	17.31	0.28	0.43
Pseudo-parallel data only											
wiki-simp-all	104K (100%)	29.93	60.81	18.05	0.36	0.47	30.13	57.46	18.53	0.39	0.49
wiki-news-all	348K (100%)	22.06	28.51	13.68	0.6	0.63	23.08	29.62	14.01	0.6	0.64
pseudo-all	452K (100%)	**30.24**	**71.32**	17.82	0.3	0.43	**31.41**	**65.65**	17.65	0.33	0.45

random pairs are uniformly sampled from the Wikipedia and the Simple Wikipedia, respectively. With the random pairs, we aim to investigate how model performance changes as we add an increasing number of sentence pairs that are non-parallel but are still representative of the two dataset styles.

5.2 Automatic Evaluation

The simplification results in Table 6 are organized in several sections according to the type of dataset used for training. We report the results of the top two beam search hypotheses produced by our models, considering that the second hypothesis often generates simpler outputs (Štajner and Nisioi, 2018).

In Table 6, `input` is copying the normal Wikipedia input sentences, without making any changes. `reference` reports the score of the original Simple Wikipedia references with respect to the other 8 references available for this dataset. `NTS` is the previously best reported result on text simplification using neural sequence models (Štajner and Nisioi, 2018). `baseline-{282K, 71K}` are our parallel LSTM baselines, trained on 282K and 71K parallel pairs, respectively.

The models trained on a *mixture* of *parallel* and *pseudo-parallel* data generate longer outputs on average, and their output is more similar to the `input`, as well as to the original Simple Wikipedia `reference`, in terms of the LD. Adding pseudo-parallel data frequently yields BLEU improvements on both Beam hypotheses: over the *NTS* system, as well as over our baselines trained solely on parallel data. The BLEU gains are larger when using the smaller parallel dataset, consisting of 71K sentence pairs. In terms of SARI, the scores remain either similar or slightly better than the baselines, indicating that simplicity in the output is preserved. The second Beam hypothesis yields higher SARI scores than the first one, in agreement with (Štajner and Nisioi, 2018). Interestingly, adding out-of-domain pseudo-parallel news data (`wiki-news-*` datasets) results in an increase in BLEU despite the potential change in style of the target sequence.

Larger pseudo-parallel datasets can lead to bigger improvements, however noisy data can result in a decrease in performance, motivating careful data selection. In our *parallel and random* setup, we find that an increasing number of random pairs added to the parallel data progressively degrades model performance. However, those models still manage to perform surprisingly well, even when over half of the pairs in the dataset are random. Thus, neural machine translation can successfully learn target transformations despite substantial data corruption, demonstrating robustness to noisy or non-parallel data for certain tasks.

When training solely on *pseudo-parallel* data, we observe lower performance on average in comparison to parallel models. However, the results are encouraging, demonstrating the potential of our approach in tasks for which there is no parallel data available. As expected, the out-of-domain news data (`wiki-news-all`) is

less suitable for simplification than the in-domain data (wiki-simp-all), because of the change in output style of the former. Results are best when mixing all pseudo-parallel pairs into a single dataset (pseudo-all). Having access to a small amount of *in-domain* pseudo-parallel data, in addition to *out-of-domain* pairs, seems to be beneficial to the success of our approach.

5.3 Human Evaluation

Due to the challenges of automatic evaluation of text simplification systems (Sulem et al., 2018), we also perform a human evaluation. We asked 8 fluent English speakers to rate the grammaticality, meaning preservation, and simplicity of model outputs produced for 100 randomly selected sentences from our test set. We exclude any model outputs which leave the input unchanged. Grammaticality and meaning preservation are rated on a Likert scale from 1 (Very bad) to 5 (Very good). Simplicity of the output sentences, in comparison to the input, is rated following (Štajner et al., 2018), between: -2 (much more difficult), -1 (somewhat more difficult), 0 (equally difficult), 1 (somewhat simpler) and 2 (much simpler).

The results are reported in Table 7, where we compare our parallel baseline (baseline-272K in Table 6) to our best model trained on a mixture of parallel and pseudo-parallel data (wiki-simp-65) and our best model trained on pseudo-parallel data only (pseudo-all). We also evaluate the original Simple Wikipedia references (reference) for comparison. In terms of simplicity, our pseudo-parallel systems are closer to the result of reference than is baseline-272K, indicating that they better match the target sentence style. baseline-272K and wiki-simp-65 perform similarly to the references in terms of grammaticality, with baseline-272K having a small edge. In terms of meaning preservation, both do worse than the references, with wiki-simp-65 having a small edge. pseudo-all performs worse on both grammaticality and meaning preservation, but is on par with the simplicity result of wiki-simp-65.

In Table 8, we also show example outputs of our best models (additional examples are available in the supplementary material). The models trained on parallel plus additional pseudo-parallel data produced outputs that preserve the meaning of 'Jeddah' as a city better than our parallel baseline, while correctly simplifying *principal* to *main*. The model trained solely on pseudo-parallel data produces a similar output, apart from wrongly replacing *jeddah* with *islam*.

Table 7: Human evaluation of the Grammaticality (**G**), Meaning preservation (**M**) and Simplicity (**S**) of model outputs (on the first Beam hypothesis).

Method	G	M	S
reference	4.53	4.34	0.69
baseline-272K	4.51	3.68	0.9
+ wiki-simp-65	4.39	3.76	0.74
pseudo-all	4.02	2.96	0.77

Table 8: Example model outputs (first Beam hypothesis).

Method	Example
input	jeddah is the **principal** gateway to mecca , islam ' s holiest city , which able-bodied muslims are required to visit at least once in their lifetime .
reference	jeddah is the **main** gateway to mecca , the holiest city of islam , where able-bodied muslims must go to at least once in a lifetime .
baseline-282K	it is the highest gateway to mecca , islam .
+ wiki-sim-65	jeddah is the **main** gateway to mecca , islam 's holiest city .
+ wiki-news-74	it is the **main** gateway to mecca , islam ' s holiest city .
pseudo-all	islam is the **main** gateway to mecca , islam 's holiest city .

6 Conclusion

We developed a hierarchical method for extracting pseudo-parallel sentence pairs from two monolingual comparable corpora composed of different text styles. We evaluated the performance of our method on automatic alignment benchmarks and extrinsically on automatic text simplification. We find improvements arising from adding pseudo-parallel sentence pairs to existing parallel datasets, as well as promising results when using the pseudo-parallel data on its own.

Our results demonstrate that careful engineering of pseudo-parallel datasets can be a successful approach for improving existing monolingual text-to-text rewriting tasks, as well as for tackling novel tasks. The pseudo-parallel data could also be a useful resource for dataset inspection and analysis. Future work could focus on improvements of our system, such as refined approaches to sentence pairing.

Acknowledgments

We acknowledge support from the Swiss National Science Foundation (grant 31003A_156976).

References

Emil Abrahamsson, Timothy Forni, Maria Skeppstedt, and Maria Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 57–65.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *Proc. of ICLR 2015*.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics.

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176. Association for Computational Linguistics.

Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. 2015. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *Proc. of EMNLP 2017*.

William Coster and David Kauchak. 2011. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proc. of NAACL 2019*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation. *arXiv preprint arXiv:1711.06861*.

Francis Grégoire and Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. *arXiv preprint arXiv:1806.05559*.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard wikipedia to simple wikipedia. In *HLT-NAACL*, pages 211–217.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 9–16. Association for Computational Linguistics.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Matt J Kusner, Yu Sun, Nicholas I Kolkin, Kilian Q Weinberger, et al. 2015. From word embeddings to document distances. In *ICML*, volume 15, pages 957–966.

Gus Liu, Pol Rosello, and Ellen Sebastian. 2016. Style transfer with non-parallel corpora.

Benjamin Marie and Atsushi Fujita. 2017. Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 392–398.

Robert C Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 135–144. Springer.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics.

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

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540. Association for Computational Linguistics.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6833–6844.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *Language Engineering Conference, 2002. Proceedings*, pages 64–71. IEEE.

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 Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. In *EMNLP*.

Jakob Uszkoreit, Jay M Ponte, Ashok C Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1101–1109. Association for Computational Linguistics.

Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. CATS: A Tool for Customized Alignment of Text Simplification Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

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