Neural semi-Markov CRF for Monolingual Word Alignment

Wuwei Lan[★1], **Chao Jiang**[★2], **Wei Xu**[2]
[1] Department of Computer Science and Engineering, Ohio State University
[2] School of Interactive Computing, Georgia Institute of Technology
lan.105@osu.edu {chao.jiang, wei.xu}@cc.gatech.edu

Abstract

Monolingual word alignment is important for studying fine-grained editing operations (i.e., deletion, addition, and substitution) in text-to-text generation tasks, such as paraphrase generation, text simplification, neutralizing biased language, etc. In this paper, we present a novel neural semi-Markov CRF alignment model, which unifies word and phrase alignments through variable-length spans. We also create a new benchmark with human annotations that cover four different text genres to evaluate monolingual word alignment models in more realistic settings. Experimental results show that our proposed model outperforms all previous approaches for monolingual word alignment as well as a competitive QA-based baseline, which was previously only applied to bilingual data. Our model demonstrates good generalizability to three out-of-domain datasets and shows great utility in two downstream applications: automatic text simplification and sentence pair classification tasks.[1]

1 Introduction

Monolingual word alignment aims to align words or phrases with similar meaning in two sentences that are written in the same language. It is useful for improving the interpretability in natural language understanding tasks, including semantic textual similarity (Li and Srikumar, 2016) and question answering (Yao, 2014). Monolingual word alignment can also support the analysis of human editing operations (Figure 1) and improve model performance for text-to-text generation tasks, such as text simplification (Maddela et al., 2021) and neutralizing biased language (Pryzant et al., 2020). It has also been shown to be helpful for data augmentation and label projection (Culkin et al., 2021) when combined with paraphrase generation.

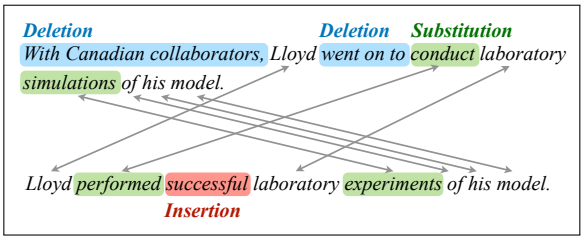

Figure 1: An example that illustrates monolingual word alignment (shown as arrows) can support analysis of human editing process and training of text generation models (§6.1), such as for simplifying complex sentences for children to read.

One major challenge for automatic alignment is the need to handle not only alignments between words and linguistic phrases (e.g., *a dozen* ↔ *more than 10*), but also non-linguistic phrases that are semantically related given the context (e.g., *tensions* ↔ *relations being strained* in Figure 3). In this paper, we present a novel neural semi-Markov CRF alignment model, which unifies both word and phrase alignments though variable-length spans, calculates span-based semantic similarities, and takes alignment label transitions into consideration. We also create a new manually annotated benchmark, **Multi**-Genre **M**onolingual **W**ord **A**lignment (MultiMWA), which consists of four datasets across different text genres and is large enough to support the training of neural-based models (Table 1). It addresses the shortcomings of existing datasets for monolingual word alignment: MTReference (Yao, 2014) was annotated by crowd-sourcing workers and contains many obvious errors (more details in §4); iSTS (Agirre et al., 2016) and SPADE/ESPADA (Arase and Tsujii, 2018, 2020) were annotated based on chunking and parsing results, which may restrict the granularity and flexibility of the alignments.

[1] Our code and data will be available at: https://github.com/chaojiang06/neural-Jacana

★ Authors contributed equally.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6815–6828
August 1–6, 2021.

Our experimental results show that the proposed semi-Markov CRF model achieves state-of-the-art performance with higher precision, in comparison to the previous monolingual word alignment models (Yao et al., 2013a,b; Sultan et al., 2014), as well as another very competitive span-based neural model (Nagata et al., 2020) that had previously only applied to bilingual data. Our model exceeds 90% F1 in the in-domain evaluation and also has very good generalizability on three out-of-domain datasets. We present a detailed ablation and error analysis to better understand the performance gains. Finally, we demonstrate the utility of monolingual word alignment in two downstream applications, namely automatic text simplification and sentence pair classification.

2 Related Work

Word alignment has a long history and was first proposed for statistical machine translation. The most representative ones are the IBM models(Brown et al., 1993), which are a sequence of unsupervised models with increased complexity and implemented the GIZA++ toolkit (Och and Ney, 2003). Many more works followed, such as FastAlign (Dyer et al., 2013). Dyer et al. (2011) also used a globally normalized log-linear model for discriminative word alignment. Bansal et al. (2011) proposed a hidden semi-Markov model to handle both continuous and noncontinuous phrase alignment. These statistical methods promoted the development of monolingual word alignment (MacCartney et al., 2008; Thadani and McKeown, 2011; Thadani et al., 2012). Yao et al. (2013a) proposed a CRF aligner following (Blunsom and Cohn, 2006), then extended it to a semi-CRF model for phrase-level alignments (Yao et al., 2013b). Sultan et al. (2014) designed a simple system with heuristic rules based on word similarity and contextual evidence.

Neural methods have been explored in the past decade primarily for bilingual word alignment. Some early attempts (Yang et al., 2013; Tamura et al., 2014) did not match the performance of GIZA++, but recent Transformer-based models started to outperform. Garg et al. (2019) proposed a multi-task framework for machine translation and word alignment, while Zenkel et al. (2020) designed an alignment layer on top of Transformer for machine translation. Both can be trained without word alignment annotations but rely on millions of bilingual sentence pairs. As for supervised methods, Stengel-Eskin et al. (2019) extracted representations from the Transformer-based MT system, then used convolutional neural network to incorporate neighboring words for alignment. Nagata et al. (2020) proposed a span prediction method and formulated bilingual word alignment as a SQuAD-style question answering task, then solved it by fine-tuning multilingual BERT. We adapt their method to monolingual word alignment as a new state-of-the-art baseline (§5.1). Some monolingual neural models have different settings from this work. Ouyang and McKeown (2019) introduced pointer networks for long, sentence- or clause-level alignments. Arase and Tsujii (2017, 2020) utilized constituency parsers for compositional and non-compositional phrase alignments. Culkin et al. (2021) considered span alignment for FrameNet (Baker et al., 1998) annotations and treated each span pair as independent prediction.

3 Neural Semi-CRF Alignment Model

In this section, we first describe the problem formulation for monolingual word alignment, then present the architecture of our neural semi-CRF word alignment model (Figure 2).

3.1 Problem Formulation

We formulate word alignment as a sequence tagging problem following previous works (Blunsom and Cohn, 2006; Yao et al., 2013b). Given a source sentence $\boldsymbol{s}$ and a target sentence $\boldsymbol{t}$ of the same language, the span alignment $\boldsymbol{a}$ consists of a sequence of tuples (i, j), which indicates that span s_i in the source sentence is aligned with span t_j in the target sentence. More specifically, $a_i = j$ means source span s_i is aligned with target span t_j. We consider all spans up to a maximum length of D words. Given a source span s_i of d ($d \leq D$) words $[s^w_{b_i}, s^w_{b_i+1}, ..., s^w_{b_i+d-1}]$, where b_i is the beginning word index, its corresponding label a_i means every word within the span s_i is aligned to the target span t_{a_i}. That is, the word-level alignments $a^w_{b_i}, a^w_{b_i+1}, ..., a^w_{b_i+d-1}$ have the same value j. We use $\boldsymbol{a^w}$ to denote the label sequence of alignments between words and $s^w_{b_i}$ to denote the b_ith word in the source sentence. There might be cases where span s_i is not aligned to any words in the target sentence, then $a_i =$ `[NULL]`. When $D \geq 2$, the Markov property would no longer hold for word-

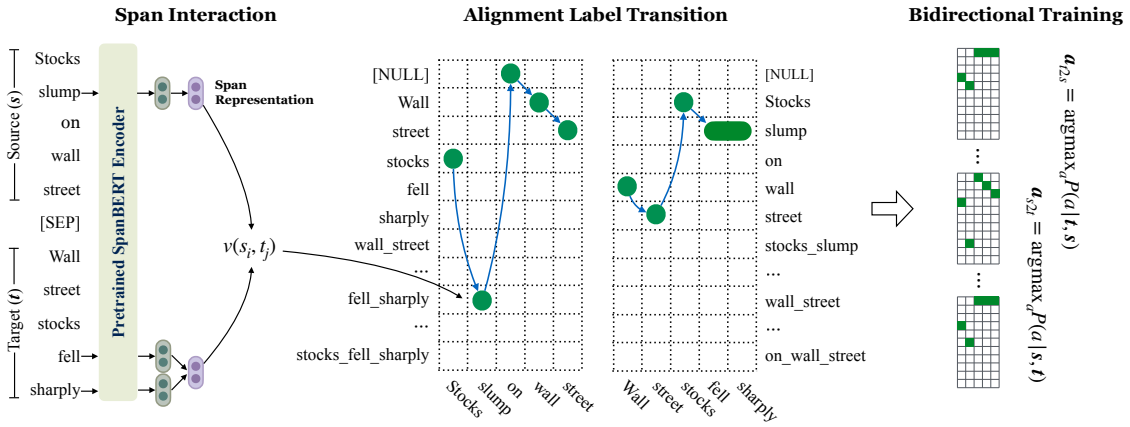

Figure 2: Illustration of our neural semi-CRF word alignment model.

level alignment labels, but for span-level labels. That is, a_i depends on $a^w_{b_i-1}$, the position in the target sentence where the source span (with ending word index $b_i - 1$) that precedes the current span s_i is aligned to. We therefore design a discriminative model using semi-Markov conditional random fields (Sarawagi and Cohen, 2005) to segment the source sentence and find the best span alignment, which we present below. One unique aspect of our semi-Markov CRF model is that it utilizes a varied set of labels for each sentence pair.

3.2 Our Model

The conditional probability of alignment $\boldsymbol{a}$ given a sentence pair $\boldsymbol{s}$ and $\boldsymbol{t}$ is defined as follows:

$$p(\boldsymbol{a}|\boldsymbol{s},\boldsymbol{t}) = \frac{e^{\psi(\boldsymbol{a},\boldsymbol{s},\boldsymbol{t})}}{\sum_{\boldsymbol{a}'\in\mathcal{A}} e^{\psi(\boldsymbol{a}',\boldsymbol{s},\boldsymbol{t})}} \quad (1)$$

where the set $\mathcal{A}$ denotes all possible alignments between the two sentences. The potential function ψ can be decomposed into:

$$\psi(\boldsymbol{a},\boldsymbol{s},\boldsymbol{t}) = \sum_i \upsilon(s_i,t_{a_i}) + \tau(a^w_{b_i-1}, a_i) + cost(\boldsymbol{a},\boldsymbol{a}^*) \quad (2)$$

where i denotes the indices of a subset of source spans that are involved in the alignment $\boldsymbol{a}$; $\boldsymbol{a}^*$ represents the gold alignment sequence at span-level. The potential function ψ consists of three elements, of which the first two compose negative log-likelihood loss: the span interaction function υ, which accounts the similarity between a source span and a target span; the Markov transition function τ, which models the transition of alignment labels between adjacent source spans; the *cost* is implemented with Hamming loss to encourage the predicted alignment sequence to be consistent with gold labels. Function υ and τ are implemented as two neural components which we describe below.

Span Representation Layer. First, source and target sentences are concatenated together and encoded by the pre-trained SpanBERT (Joshi et al., 2020) model. The hidden representations in the last layer of the encoder are extracted for each WordPiece token, then averaged to form the word representations. Following previous work (Joshi et al., 2020), the span is represented by a self-attention vector computed over the representations of each word within the span, concatenated with the Transformer output states of two endpoints.

Span Interaction Layer. The semantic similarity score between source span s_i and target span t_j is calculated by a 2-layer feed-forward neural network FF_{sim} with Parametric Relu (PReLU) (He et al., 2015),[2] after applying layer normalization to each span representation:

$$\upsilon(s_i,t_j) = \text{FF}_{sim}([\boldsymbol{h}^s_i; \boldsymbol{h}^t_j; |\boldsymbol{h}^s_i - \boldsymbol{h}^t_j|; \boldsymbol{h}^s_i \circ \boldsymbol{h}^t_j]) \quad (3)$$

where $[;]$ is concatenation and $\circ$ is element-wise multiplication. We use h^s_i and h^t_j to denote the representation of source span s_i and target span t_j, respectively.

Markov Transition Layer. Monolingual word alignment moves along the diagonal direction in most cases. To incorporate this intuition, we propose a scoring function to model the transition between the adjacent alignment labels $a^w_{b_i-1}$ and a_i. The main feature we use is the distance between the beginning index of current target span and the

[2] We also compared ReLU and GeLU, and found PReLU works slightly better.

end index of the target span that the prior source span is aligned to. The distance is binned into 1 of 13 buckets with the following boundaries [-11, -6, -4, -3, -2, -1, 0, 1, 2, 3, 5, 10], and each bucket is encoded by a 128-dim randomly initialized embedding. It is then transformed into a real-value score by a 1-layer feed forward neural network.

Training and Inference. During training, we minimizes the negative log-likelihood of the gold alignment $\boldsymbol{a}^*$, and the model is trained from both directions (source to target, target to source):

$$\sum_{(\boldsymbol{s},\boldsymbol{t},\boldsymbol{a}^*)} -\log p(\boldsymbol{a}^*_{s2t}|\boldsymbol{s},\boldsymbol{t}) - \log p(\boldsymbol{a}^*_{t2s}|\boldsymbol{t},\boldsymbol{s}) \quad (4)$$

where $\boldsymbol{a}^*_{s2t}$ and $\boldsymbol{a}^*_{t2s}$ represent the gold alignment labels from both directions.

During inference, we use the Viterbi algorithm to find the optimal alignment. There are different strategies to merge the outputs from two directions, including intersection, union, grow-diag (Koehn, 2009), bidi-avg (Nagata et al., 2020), etc. It can be seen as a hyper-parameter and decided based on the dev set. In this work, we use intersection in our semi-CRF model for all experiments.

3.3 Implementation Details

We implement our model in PyTorch (Paszke et al., 2017). We use the Adam optimizer and set both the learning rate and weight decay as 1e-5. We set the maximum span size to 3 for our neural semi-CRF model, which can converge within 5 epochs. The neural semi-CRF model has ∼2 hour training time per epoch for MultiMWA-MTRef, measured on a single GeForce GTX 1080 Ti GPU.

4 A Multi-Genre Benchmark for Monolingual Word Alignment

In this section, we present the manually annotated **Multi**-genre **M**onolingual **W**ord **A**lignment (MultiMWA) benchmark that consists of four datasets of different text genres. As summarized in Table 1, our new benchmark is the largest to date and of higher quality compared to existing datasets. In contrast to iSTS (Agirre et al., 2016) and SPADE/ESPADA (Arase and Tsujii, 2018, 2020), our annotation does not rely on external chunking or parsing that may introduce errors or restrict the granularity and flexibility. Our benchmark contains both token alignments and a significant portion of phrase alignments as they are semantically equivalent as a whole. Our benchmark also contains a large portion of semantically similar but not strictly equivalent sentence pairs, which are common in text-to-text generation tasks and thus important for evaluating the monolingual word alignment models under this realistic setting.

For all four datasets, we closely follow the standard 6-page annotation guideline[3] from (Callison-Burch et al., 2006) and further extend it to improve the phrase-level annotation consistency (more details in Appendix B.1). We describe each of the four datasets below.

MultiMWA-MTRef. We create this dataset by annotating 3,998 sentence pairs from the MTReference (Yao, 2014), which are human references used in a machine translation task. The original labels in MTReference were annotated by crowd-sourcing workers on Amazon Mechanical Turk following the guideline from (Callison-Burch et al., 2006). In an early pilot study, we discovered that these crowd-sourced annotations are noisy and contain many obvious errors. It only gets 73.6/96.3/83.4 for Precision/Recall/F_1 on a random sample of 100 sentence pairs, when compared to the labels we manually corrected.

To address the lack of reliable annotation, we hire two in-house annotators to correct the original labels using GoldAlign[4] (Gokcen et al., 2016), an annotation tool for monolingual word alignment. Both annotators have linguistic background and extensive NLP annotation experience. We provide a three-hour training session to the the annotators, during which they are asked to align 50 sentence pairs and discuss until consensus. Following previous work, we calculate the inter-annotator agreement as 84.2 of F_1 score for token-level non-identical alignments by comparing one annotator's annotation against the other's. The alignments between identical words are usually easy for human annotators. After merging the the labels from both annotators, we create a new split of 2398/800/800 for train/dev/test set. To ensure the quality, an adjudicator further exams the dev and test sets and constructs the final labels.

MultiMWA-Newsela. Newsela corpus (Xu et al., 2015b) consists of 1,932 English news articles and their simplified versions written by

[3] http://www.cs.jhu.edu/~ccb/publications/paraphrase_guidelines.pdf

[4] https://github.com/ajdagokcen/goldalign-repo

Datasets	#Train	#Dev	#Test	Length	%aligned	%word/phrase	%id/non-id	Genre	External	License
Existing Monolingual Word Alignment Datesets										
MSR RTE (Brockett, 2007)	800	–	800	29 / 11	37.9	90.0 / 10.0	76.6 / 23.4	Misc.	–	Free
Edinburgh++ (Thadani et al., 2012)	714	–	306	22 / 22	85.7	77.7 / 22.3	67.2 / 32.8	Misc.	–	Free
iSTS (Agirre et al., 2016)	1,506	–	750	9 / 9	74.0	6.5 / 93.5	23.3 / 76.7	News Image captions	Chunking	Free
SPADE / ESPADA† (2018; 2020)	1,916	50	151	23 / 23	81.9	44.0 / 56.0	72.3 / 27.7	News	Parsing	LDC
Our Multi-Genre Monolingual Word Alignment (MultiMWA) Benchmark										
MultiMWA-MTRef	2,398	800	800	22 / 17	88.6	62.0 / 38.0	52.6 / 47.3	News	–	Free
MultiMWA-Wiki	2,514	533	1,052	30 / 29	91.8	95.6 / 4.4	94.1 / 5.9	Wikipedia	–	Free
MultiMWA-Newsela	–	–	500	27 / 23	76.5	74.6 / 25.4	67.1 / 32.9	News	–	Free*
MultiMWA-arXiv	–	–	200	29 / 28	87.8	96.6 / 3.4	93.4 / 6.6	Scientific writing	–	Free
Total	**4,912**	**1,333**	**2,552**	**26 / 23**	**89.4**	**79.3 / 20.7**	**73.8 / 26.2**	**all above**	**–**	**Free**

Table 1: Statistics of our new MultiWMA benchmark and existing datasets. **Length** of the longer/shorter sentence in each pair is measured by the number of tokens. **%aligned** is the percentage of aligned words among all words. **%word/phrase** denotes the percentage of word alignment and phrasal alignment. **%id/non-id** specifies the percentage of identical (e.g., *Lloyd* ↔ *Lloyd*) and non-identical (e.g., *conduct* ↔ *performed*) alignments. **Externel** indicates whether the annotation relies on additional linguistic information. †ESPADA (train) has not been released at the time of writing; statistics are based on the SPADE (dev/test) dataset. *Newsela data is free for academic research but license needs to be requested at: https://newsela.com/data.

professional editors. It has been widely used in text simplification research (Xu et al., 2016; Zhang and Lapata, 2017; Zhong et al., 2020). We randomly select 500 complex-simple sentence pairs from the test set of Newsela-Auto (Jiang et al., 2020),[5] which is the newest sentence-aligned version of Newsela. 214 of these 500 pairs contain sentence splitting. An in-house annotator[6] labels the word alignment by correcting the outputs from GIZA++ (Och and Ney, 2003).

MultiMWA-arXiv. The arXiv[7] is an open-access platform that stores more than 1.7 million research papers with their historical versions. It has been used to study paraphrase generation (Dong et al., 2021) and statement strength (Tan and Lee, 2014). We first download the LaTeX source code for 750 randomly sampled papers and their historical versions, then use OpenDetex[8] package to extract plain text from them. We use the trained neural CRF sentence alignment model (Jiang et al., 2020) to align sentences between different versions of the papers and sample 200 non-identical aligned sentence pairs for further annotation. The word alignment is annotated in a similar procedure to that of the MultiMWA-Wiki.

MultiMWA-Wiki. Wikipedia has been widely used in text-to-text tasks, including text simplification (Jiang et al., 2020), sentence splitting (Botha et al., 2018), and neutralizing bias language (Pryzant et al., 2020). We follow the method in (Pryzant et al., 2020) to extract parallel sentences from Wikipedia revision history dump (dated 01/01/2021) and randomly sample 4,099 sentence pairs for further annotation. We first use an earlier version of our neural semi-CRF word aligner (§3) to automatically align words for the sentence pairs, then ask two in-house annotators to correct the aligner's outputs. The inter-annotator agreement is 98.1 at token-level measured by F_1.[9] We split the data into 2514/533/1052 sentence pairs for train/dev/test sets.

5 Experiments

In this section, we present both in-domain and out-of-domain evaluations for different word alignment models on our MultiWMA benchmark. We also provide a detailed error analysis of our neural semi-CRF model and an ablation study to analyze the importance of each component.

5.1 Baselines

We introduce a novel state-of-the-art baseline by adapting the **QA-based method** in (Nagata et al., 2020), which has not previously applied to monolingual word alignment but only bilingual word alignment. This method treats the word alignment problem as a collection of independent predictions

[5] More specifically, we sample from the exact test set used in Table 2 in Maddela et al. (2021).

[6] This annotator has annotated MultiMWA-MTRef.

[7] https://arxiv.org/

[8] https://github.com/pkubowicz/opendetex

[9] The inter-annotator agreement is much higher compared to that of MultiMWA-MTRef, as the parallel sentences extracted from Wikipedia revision history have more overlap.

Models	MultiMWA-MTRef$_{Sure}$ P (P_i / P_n)	R (R_i / R_n)	F_1 (F_{1i} / F_{1n})	EM	MultiMWA-MTRef$_{Sure+Poss}$ P (P_i / P_n)	R (R_i / R_n)	F_1 (F_{1i} / F_{1n})	EM	MultiMWA-Wiki P (P_i / P_n)	R (R_i / R_n)	F_1 (F_{1i} / F_{1n})	EM
JacanaToken (Yao et al., 2013a)	87.9 94.4 / 65.1	72.2 94.7 / 41.3	79.3 94.6 / 50.5	2.6	82.8 93.3 / 61.7	70.5 96.7 / 43.6	76.2 95.0 / 51.1	1.3	98.8 99.3 / 77.1	95.7 99.5 / 71.6	97.2 99.4 / 74.3	59.8
JacanaPhrase (Yao et al., 2013b)	84.4 94.1 / 58.5	72.4 95.3 / 40.7	78.0 94.7 / 48.0	1.9	82.8 93.3 / 61.4	70.0 96.2 / 42.5	75.8 94.8 / 50.3	1.4	92.8 98.5 / 44.4	97.0 99.8 / 49.1	94.9 99.2 / 46.6	27.4
PipelineAligner (Sultan et al., 2014)	**96.0** 98.1 / 78.9	67.7 93.3 / 30.6	79.4 95.6 / 44.1	2.5	**97.1** 98.3 / 82.9	60.8 92.9 / 23.9	74.8 95.5 / 37.1	1.0	**99.5** 99.6 / 66.2	94.9 99.6 / 60.0	97.1 99.6 / 62.9	53.4
QA-based Aligner	88.4 98.2 / 76.3	**92.3** 99.2 / 83.9	90.3 98.7 / 79.9	14.0	91.3 98.5 / 84.1	**92.9** 99.2 / 86.9	92.1 98.9 / 85.5	21.3	97.4 99.5 / 82.3	**97.9** 99.8 / 81.9	**97.6** 99.7 / 82.1	67.4
Neural CRF Aligner	87.6 97.3 / 74.2	91.6 99.5 / 82.2	89.5 98.4 / 78.0	10.8	91.5 98.5 / 83.4	90.2 99.2 / 82.1	90.8 98.8 / 82.7	16.9	96.5 99.3 / 80.6	97.6 99.6 / 80.6	97.1 99.4 / 80.6	63.5
Neural semi-CRF Aligner	90.6 98.9 / 78.9	90.3 98.9 / 79.1	**90.5** 98.9 / 79.0	**14.1**	94.7 99.3 / 89.1	90.2 98.7 / 82.3	**92.4** 99.0 / 85.5	**23.3**	97.7 99.6 / 82.8	97.5 99.7 / 80.8	**97.6** 99.7 / 81.8*	**68.5**

Table 2: In-domain evaluation of different monolingual word alignment models on the MultiMWA benchmark. We report the precision (**P**), recall (**R**), $\mathbf{F}_1$, and exact match (**EM**), which is the percentage of sentence pairs for which model predictions are exactly same as gold labels for the entire sentence. For each metric, we also report the performance on identical alignments (P_i, R_i, F_{1i}) and non-identical alignments (P_n, R_n, F_{1n}) separately. * MultiMWA-Wiki contains only about 5% non-identical alignment.

Models	MultiMWA-Newsela P (P_i / P_n)	R (R_i / R_n)	F_1 (F_{1i} / F_{1n})	EM	MultiMWA-arXiv P (P_i / P_n)	R (R_i / R_n)	F_1 (F_{1i} / F_{1n})	EM	MultiMWA-Wiki P (P_i / P_n)	R (R_i / R_n)	F_1 (F_{1i} / F_{1n})	EM
JacanaToken (Yao et al., 2013a)	85.5 91.2 / 60.1	74.9 97.5 / 39.7	79.8 94.3 / 47.9	11.0	94.9 97.3 / 72.6	96.8 99.5 / 73.4	95.8 98.4 / 73.0	49.0	94.7 98.4 / 51.2	96.9 99.9 / 50.1	95.8 99.2 / 50.6	33.3
JacanaPhrase (Yao et al., 2013b)	84.3 91.3 / 53.9	75.0 97.4 / 38.6	79.4 94.3 / 45.0	8.2	90.9 97.1 / 53.2	96.6 99.1 / 64.7	93.7 98.1 / 58.4	31.5	92.9 98.5 / 44.9	96.9 99.8 / 49.6	94.9 99.1 / 47.1	28.0
PipelineAligner (Sultan et al., 2014)	**95.2** 96.9 / 64.4	69.4 95.3 / 25.4	80.3 96.1 / 36.5	10.0	**98.5** 98.8 / 68.3	94.6 99.0 / 62.4	96.5 98.9 / 65.2	49.0	**99.5** 99.6 / 66.2	94.9 99.6 / 60.0	97.1 99.6 / 62.9	53.4
QA-based Aligner	84.8 95.3 / 69.4	**87.9** 99.1 / 71.4	86.2 97.1 / 70.4	16.2	93.9 98.0 / 70.7	94.3 95.0 / 79.9	94.1 96.5 / 75.0	27.0	96.1 99.3 / 76.2	**98.2** 99.8 / 78.3	97.2 99.5 / 77.3	57.8
Neural CRF Aligner	88.2 95.3 / 72.3	85.0 99.0 / 66.3	86.6 97.1 / 69.1	15.6	92.9 96.4 / 62.9	**98.7** 99.8 / 73.3	95.7 98.0 / 67.7	43.5	96.1 99.1 / 70.5	98.0 99.9 / 71.9	97.0 94.5 / 71.2	52.1
Neural semi-CRF Aligner	89.4 96.7 / 76.1	85.0 98.4 / 66.5	**87.2** 97.6 / 71.0	**21.6**	96.2 98.9 / 79.3	98.4 99.6 / 83.0	**97.3** 99.3 / 81.1	**62.5**	97.2 99.6 / 80.4	97.6 99.5 / 79.5	**97.4** 99.5 / 79.9	**64.8**

Table 3: Out-of-domain evaluation of different monolingual word alignment models on the MultiMWA benchmark. All the models in this table are trained on the MultiMWA-MTRef$_{Sure+Poss}$ dataset.

from every token in the source sentence to a span in the target sentence, which is then solved by fine-tuning multilingual BERT (Devlin et al., 2019) similarly as for SQuAD-style question answering task. Taking the sentence pair in Figure 1 as an example, the word to be aligned is marked by ¶ in the source sentence and concatenated with the entire target sentence to form the input as "*With Canadian* ··· ¶*conduct*¶ ··· *his model. Lkoyd performed* ··· *his model.* ". A span prediction model based on fine-tuning multilingual BERT is then expected to extract *performed* from the target sentence. The predictions from both directions (source to target, target to source) are symmetrized to produce the final alignment, using a probability threshold of 0.4 instead of the typical 0.5.

We change to use standard BERT in this model for monolingual alignment and find that the 0.4 threshold chosen by Nagata et al. (2020) is almost optimal in maximizing the F_1 score on our MultiMWA-MTRef dataset. This QA-based method alone outperforms all existing models for monolingual word alignment, including: **JacanaToken** aligner (Yao et al., 2013a), which is a CRF model using hand-crafted features and external resources; **JacanaPhrase** aligner (Yao et al., 2013b), which is a semi-CRF model relying on feature templates and external resources; **PipelineAligner** (Sultan et al., 2014), which is a pipeline system that utilizes word similarity and contextual information with heuristic algorithms. We also create a variation of our model, a **Neural CRF aligner**, in which all modules remain the same but the max span length is set to 1, to evaluate the benefits of span-based alignments.

5.2 Experimental Results

Following the literature (Thadani et al., 2012; Yao et al., 2013a,b), we present results under both $Sure$ and $Sure + Poss$ settings for the MultiMWA-MTRef dataset. $Sure + Poss$ setting includes all the annotated alignments, and $Sure$ only contains a subset of them which are agreed by multiple annotators. We consider $Sure + Poss$ as the default setting for all the other three datasets.

The in-domain evaluation results are shown in

Neural semi-CRF Aligner	F_1	EM	Δ_{F_1}/Δ_{EM}
w/ SpanBERT	92.1	23.3	0.0 / 0.0
w/ BERT	90.8	18.9	-1.3 / -4.4
w/o Transition Layer	91.9	21.3	-0.2 / -2.0
w/ post-processing	92.1	23.3	0.0 / 0.0
w/ intersection	92.0	21.8	-0.1/ -1.5
w/ union	91.1	20.1	-1.0 / -3.2
w/ grow-diag	91.5	20.6	-0.6 / -2.7

Table 4: Ablation study of our neural semi-CRF aligner with each component removed or swapped. The results are based on the dev set of $\text{MTRef}_{Sure+Poss}$.

Table 2. The neural models are working remarkably well in comparison to the non-neural methods, especially as measured by Exact Matches (EM). On both MTRef and Wiki datasets, our neural semi-CRF model achieves the best F_1 and EM. QA-based aligner also has competitive performance with strong recall, however, its precision is lower compared to our model. It is worthy to note that our model has a modular design, and can be more easily adjusted than QA-based method to suit different datasets and downstream tasks.

Table 3 presents the out-of-domain evaluation results. Our neural models achieve the best performance across all three datasets. This demonstrates the generalization ability of our model, which can be useful in the downstream applications.

5.3 Ablation Study

Table 4 shows the ablation study for our neural semi-CRF model. F_1 and EM drops by 1.3 and 4.4 points respectively after replacing SpanBERT with BERT, indicating the importance of optimized pre-trained representations. Markov transition layer contributes mainly to the alignment accuracy (EM). We have experimented with different strategies to merge the outputs from two directions: intersection yields better precision, grow-diag and union bias towards recall. Leveraging the span interaction matrix generated by our model (details in §3.2), we design a simple post-processing rule to extend the phrasal alignment to spans that are longer than 3 tokens. Adjacent target words are gradually included if they have very high semantic similarity with the same source span. This rule further improves recall and achieves the best F_1 on the MultiMWA-MTRef.

5.4 Error Analysis

We sample 50 sentence pairs from the dev set of MultiMWA-MTRef and analyze the errors under **Sure+Poss** setup.[10] Figure 4 shows how the performance of different alignment models would improve, if we resolve each of the 7 types of errors. We discuss the categorization of errors and their breakdown percentages below:

Phrase Boundary (58.6%). The phrase boundary error (see ③ in Figure 3 for an example) is the most prominent error in all models, attributing 7.6 points of F_1 for JacanaPhrase, 5.7 for QA aligner, and 4.7 for neural semi-CRF aligner. For another example, instead of 3x2 alignment *funds for research* ↔ *research funding*, our model captures two 1x1 alignments, *funds* ↔ *funding* and *research* ↔ *research*. This is largely due to the fact that alignments are not limited to linguistic phrases (e.g., noun phrases, verb phrases, etc.), but rather, include non-linguistic phrases. It could also be challenging to handle longer spans, such as *keep his position* ↔ *protect himself from being removed* (more on this in Appendix B.2). Although we use SpanBERT for better phrase representation, there is still room for improvement.

Function Words (19.1%). Function words can be tricky to align when rewording and reordering happens, such as ②. Adding on the complexity, same function word may appear more than once in one sentence. This type of error is common in all the models we experiment with. It attributes 4.7 points of F_1 for JacanaPhrase, 1.3 for QA aligner, and 1.5 for our neural semi-CRF aligner.

Content Words (14.2%). Similar to function words, content words (e.g., *security bureau* ↔ *defense ministry*) can also be falsely aligned or missed, but the difference between neural and non-neural model is much more significant. This error type attributes 7.7 points of F_1 score for Jacana aligner, but only 1.1 and 0.8 for neural semi-CRF aligner and QA aligner, respectively.

Context Implication (5.6%). Some words or phrases that are not strictly semantically equivalent can also be aligned if they appear in a similar context. For example, given the source sentence

[10] The strict **Sure only** labels exclude many alignments that are critical for certain applications, such as label projection. We thus focus on the **Sure+Poss** labels for error analysis.

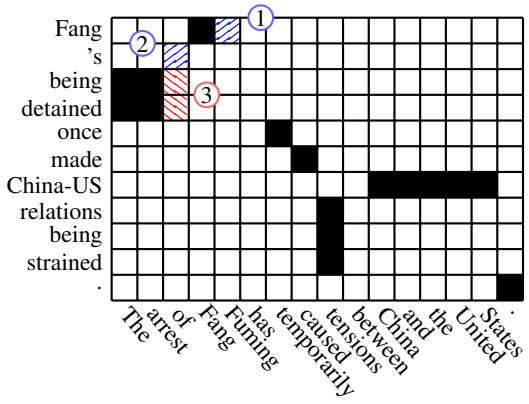

Figure 3: Error examples of the semi-CRF word alignment model on MTRef data. Black-filled boxes denote true positives, boxes filled with blue diagonal lines are false negatives, and red slant lines are false positives.

'Gaza international airport was put into operation the day before' and the target sentence *'The airport began operations one day before'*, the phrase pair *was put into* ↔ *began* can be aligned. This type is related to 2.8 F_1 score improvement for Jacana aligner, but only 0.4 and 0.2 for neural semi-CRF and QA-based aligners, respectively.

Debatable Labels (1.9%). Word alignment annotation can be subjective sometimes. Take phrase alignment *two days of* ↔ *a two-day* for example, it can go either way to include the function word '*a*' in the alignment, or not.

Name Variations (0.6%). While our neural semi-CRF model is designed to handle spelling variations or name abbreviations, it fails sometimes as shown by ① in Figure 3 as an example. Some cases can be very difficult, such as *SAWS* ↔ *the state's supervision and control bureau of safe production*, where *SAWS* stands for *State Administration of Work Safety*.

Skip Alignment (0.0%). Non-contiguous tokens can be aligned to the same target token or phrase (e.g., *owes ... to* ↔ *is a result of*), posing a challenging situation for monolingual word aligners. However, this error is rare, as only 0.6% of all alignments in MTRef dev set are discontinuous.

6 Downstream Applications

In this section, we apply our monolingual word aligner to some downstream applications, including both generation and understanding tasks.

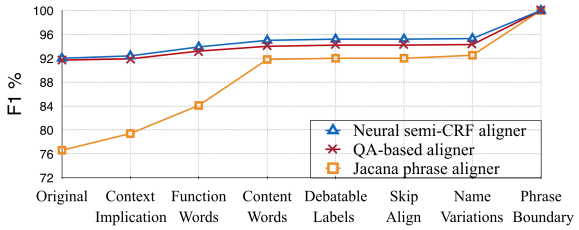

Figure 4: Performance comparison on MTRef dev set for 3 different aligners after resolving each error type.

6.1 Automatic Text Simplification

Text simplification aims to improve the readability of text by rewriting complex sentences with simpler language. We propose to incorporate word alignment information into the state-of-the-art EditNTS model (Dong et al., 2019) to explicitly learn the edit operations, including addition, deletion and paraphrase. The EditNTS model uses a neural programmer-interpreter architecture, which derives the ADD, KEEP and DELETE operation sequence based on the edit-distance measurements during training time. We instead construct this edit sequence based on the neural semi-CRF aligner's outputs (trained on MTRef$_{Sure+Poss}$) with an additional REPLACE tag to train the EditNTS model (more details in Appendix A).

Table 5 presents the text simplification results on two benchmark datasets, Newsela-auto and Wikipedia-auto (Jiang et al., 2020), where we improve the SARI score (Xu et al., 2016) by 0.9 and 0.6, respectively. The SARI score averages the F_1/precision of n-grams inserted (**add**), kept (**keep**) and deleted (**del**) when compared to human references. We also calculate the BLEU score with respect to the input (**s-BL**), the percentage of new words (**%new**) added, and the percentage of system outputs being identical to the input (**%eq**) to show the paraphrasing capability. We manually inspect 50 sentences sampled from Newsela-auto test set and find that both models (EditNTS and EditNTS+Aligner) generate the same output for 10 sentences. For the remaining 40 sentences, the original EditNTS only attempts to paraphrase 4 times (2 are good). Our modified model (EditNTS+Aligner) is more aggressive, generating 25 paraphrases (11 are good). With the help of word aligner, the modified model also produces a higher number of good deletions (20 vs. 13) and a lower number of bad deletions (6 vs. 12), which is consistent with the better **keep** and **del** scores.

Datasets	Models	SARI	add	keep	del	FK	SLen	OLen	CR	%split	s-BL	%new	%eq
Newsela-auto	Complex (input)	11.8	0.0	35.5	0.0	12.3	24.8	24.8	1.0	2.0	100.0	0.0	100.0
	Simple (reference)	86.9	84.7	78.4	97.6	6.5	13.3	13.3	0.63	0.8	25.7	33.5	0.0
	EditNTS	36.6	1.1	32.9	75.7	7.5	14.3	14.3	0.66	2.4	50.2	6.5	1.2
	EditNTS + Aligner	**37.5**	**1.3**	**33.4**	**77.9**	7.2	14.3	14.3	0.66	1.5	49.1	7.6	0.8
Wikipedia-auto	Complex (input)	24.9	0.0	74.6	0.0	13.4	22.6	22.6	1.0	0.8	100.0	0.0	100.0
	Simple (reference)	81.7	66.2	97.5	81.5	12.2	21.7	21.7	0.97	5.4	64.0	14.8	16.2
	EditNTS	36.8	**2.1**	68.4	39.8	12.8	23.6	23.6	1.06	1.7	69.7	12.4	0.6
	EditNTS + Aligner	**37.4**	1.9	**69.5**	**40.9**	12.7	23.6	23.6	1.05	0.6	74.4	10.2	2.8

Table 5: Downstream application on text simplification. By incorporating our monolingual word aligner into the EditNTS (Dong et al., 2019) model, we improve the performance measured by **SARI** score (the main automatic metric for simplification) and its three parts: precision for delete (**del**), F_1 scores for **add** and **keep** operations.

Models	RTE	MRPC	STS-B	STS14	WikiQA	SICK	PIT	URL	TrecQA	QQP	MNLI	SNLI
	2.5k	3.5k	5.7k	8k	8k	10k	11k	42k	53k	363k	392k	549k
	Acc	F_1	r/ρ	r	MAP/MRR	Acc	max_F_1	max_F_1	MAP/MRR	Acc	Acc_m/Acc_mm	Acc
BERT	65.3	88.2	86.7/85.8	83.6	81.8/83.0	86.2	75.0	**78.7**	84.4/**89.6**	90.8	84.8/83.1	**90.5**
BERT + Aligner	**67.3**	**88.9**	**86.8/86.0**	**83.7**	**83.2/84.4**	**87.2**	**75.5**	78.5	**85.1**/87.8	**90.9**	84.8/**83.5**	90.4

Table 6: Downstream applications on natural language inference (RTE, SICK, MNLI, SNLI), paraphrase identification (MRPC, PIT, URL, QQP), question answering (WikiQA, TrecQA), and semantic textual similarity (STS-B, STS14) tasks. The datasets in this table are ordered by the size of their training set, as shown in the second row.

6.2 Sentence Pair Modeling

We can utilize our neural aligner in sentence pair classification tasks (Lan and Xu, 2018), adding conditional alignment probability $p(\boldsymbol{a}|\boldsymbol{s},\boldsymbol{t})$ as an extra feature. We concatenate it with the `[CLS]` representation in fine-tuned BERT and apply the softmax layer for prediction. We experiment with on different datsets for various tasks, including: natural language inference on SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), SICK (Marelli et al., 2014), and RTE (Giampiccolo et al., 2007) from the GLUE benchmark (Wang et al., 2018); semantic textual similarity on STS-B (Cer et al., 2017) and STS14 (Agirre et al., 2014); question answering on WikiQA (Yang et al., 2015) and TrecQA (Wang et al., 2007); paraphrase identification on MRPC (Dolan and Brockett, 2005), URL (Lan et al., 2017), PIT (Xu et al., 2015a), and QQP (Iyer et al., 2017).

We implement the fine-tuned $\text{BERT}_{\text{base}}$ model using Huggingface's library (Wolf et al., 2019). Table 6 shows performance improvement on small (2k-15k) datasets, which include SICK, STS-B, MRPC, RTE, WikiQA, and PIT, but little or no improvement on large (40k-550k) datasets, such as SNLI, MNLI, and QQP. We hypothesize that the Transformer model can potentially learn the latent word alignment through self-attentions, but not as effectively for small data size.

7 Conclusion

In this work, we present the first neural semi-CRF word alignment model which achieves competitive performance on both in-domain and out-of-domain evaluations. We also create a manually annotated **Multi**-Genre **M**onolingual **W**ord Alignment (MultiMWA) benchmark which is the largest and of higher quality compared to existing datasets.

Acknowledgement

We thank Yang Chen, Sarthak Garg, and anonymous reviewers for their helpful comments. We also thank Sarah Flanagan, Yang Zhong, Panya Bhinder, Kenneth Kannampully for helping with data annotation. This research is supported in part by the NSF awards IIS-2055699, ODNI and IARPA via the BETTER program contract 19051600004, ARO and DARPA via the SocialSim program contract W911NF-17-C-0095, and Criteo Faculty Research Award to Wei Xu. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA, ARO, DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

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 (SemEval)*.

Eneko Agirre, Aitor Gonzalez Agirre, Inigo Lopez-Gazpio, Montserrat Maritxalar, German Rigau Claramunt, and Larraitz Uria. 2016. Semeval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*.

Yuki Arase and Junichi Tsujii. 2017. Monolingual phrase alignment on parse forests. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Yuki Arase and Junichi Tsujii. 2018. SPADE: Evaluation dataset for monolingual phrase alignment. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Yuki Arase and Jun'ichi Tsujii. 2020. Compositional phrase alignment and beyond. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING)*.

Mohit Bansal, Chris Quirk, and Robert Moore. 2011. Gappy phrasal alignment by agreement. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics (COLING) and the 44th annual meeting of the Association for Computational Linguistics (ACL)*.

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Chris Brockett. 2007. Aligning the rte 2006 corpus. In *Technical Report MSR-TR-2007-77, Microsoft Research*.

Peter F. Brown, Stephen A. Della-Pietra, Vincent J. Della-Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics (CL)*.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2006. Annotation guidelines for paraphrase alignment. Technical report.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*.

Ryan Culkin, J. Edward Hu, Elias Stengel-Eskin, Guanghui Qin, and Benjamin Van Durme. 2021. Iterative Paraphrastic Augmentation with Discriminative Span Alignment. *Transactions of the Association for Computational Linguistics (TACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP)*.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Chris Dyer, Jonathan H. Clark, Alon Lavie, and Noah A. Smith. 2011. Unsupervised word alignment with arbitrary features. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Ajda Gokcen, Evan Jaffe, Johnsey Erdmann, Michael White, and Douglas Danforth. 2016. A corpus of word-aligned asked and anticipated questions in a virtual patient dialogue system. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First Quora Dataset Release: Question Pairs. In *https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs*.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural crf model for sentence alignment in text simplification. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics (TACL)*.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Wuwei Lan and Wei Xu. 2018. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of International Conference on Computational Linguistics (COLING)*.

Tao Li and Vivek Srikumar. 2016. Exploiting sentence similarities for better alignments. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Bill MacCartney, Michel Galley, and Christopher D Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

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

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics (CL)*.

Jessica Ouyang and Kathy McKeown. 2019. Neural network alignment for sentential paraphrases. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Sunita Sarawagi and William W Cohen. 2005. Semimarkov conditional random fields for information extraction. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment. In *Proceedings of Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics (TACL)*.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2014. Recurrent neural networks for word alignment model. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Kapil Thadani, Scott Martin, and Michael White. 2012. A joint phrasal and dependency model for paraphrase alignment. In *Proceedings of International Conference on Computational Linguistics (COLING)*.

Kapil Thadani and Kathleen McKeown. 2011. Optimal and syntactically-informed decoding for monolingual phrase-based alignment. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 2007. What is the Jeopardy model? A quasi-synchronous grammar for qa. In *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wei Xu, Chris Callison-Burch, and William B. Dolan. 2015a. SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015b. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics (TACL)*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics (TACL)*.

Nan Yang, Shujie Liu, Mu Li, Ming Zhou, and Nenghai Yu. 2013. Word alignment modeling with context dependent deep neural network. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Xuchen Yao. 2014. *Feature-driven Question Answering With Natural Language Alignment*. Ph.D. thesis, Johns Hopkins University.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013a. A lightweight and high performance monolingual word aligner. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013b. Semi-markov phrase-based monolingual alignment. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms giza++. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

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

Complex sentence: ['With','Canadian', 'collaborators,', 'Lloyd', 'went', 'on', 'to', 'conduct', 'laboratory', 'simulations', 'of', 'his', 'model.']
Simple sentence: ['Lloyd', 'performed', 'successful', 'laboratory', 'experiments', 'of', 'his', 'model.']
Expert program from EditNTS: [DEL, DEL, DEL, KEEP, ADD('performed'), ADD('successful'), DEL, DEL, DEL, DEL, 'KEEP', ADD('experiments'), DEL, KEEP, KEEP, KEEP]
Expert program from EditNTS with Aligner: [DEL, DEL, DEL, KEEP, ADD('performed'), ADD('successful'), DEL, DEL, DEL, DEL, 'KEEP', 'REPLACE-S', ADD('experiments'), 'REPLACE-E', KEEP, KEEP, KEEP]

Table 7: Expert program comparison between the original EditNTS and our modified version with word alignment for the example in Figure 1.

Models	SARI	add	keep	del	FK	SLen	OLen	CR	%split	s-BL	%new	%eq
Complex (input)	11.8	0.0	35.5	0.0	12.3	24.8	24.8	1.0	2.0	100.0	0.0	100.0
Simple (reference)	86.9	84.7	78.4	97.6	6.5	13.3	13.3	0.63	0.8	25.7	33.5	0.0
EditNTS (original)	36.6	1.1	32.9	75.7	7.5	14.3	14.3	0.66	2.4	50.2	6.5	1.2
EditNTS (original) + Aligner	36.6	1.2	32.6	75.9	7.4	14.1	14.1	0.65	2.2	49.3	6.0	1.5
EditNTS (new)	36.9	1.2	**33.8**	75.8	8.3	16.3	16.3	0.73	1.6	56.5	5.7	0.7
EditNTS (new) + Aligner	**37.5**	**1.3**	33.4	**77.9**	7.2	14.3	14.3	0.66	1.5	49.1	7.6	0.8

Table 8: Comparison experiments on Newsela-auto dataset with different versions of EditNTS model. + Aligner means using the neural semi-CRF aligner output, EditNTS (new) means adding the REPLACE-S/E tags to the original EditNTS model.

A EditNTS with Aligner

The original EditNTS model constructs expert program with the shortest edit path from complex sentence to simple sentence, specifically, it calculates the Levenshtein distances without substitutions and recovers the edit path with three labels: ADD, KEEP and DEL. Since edit distance relies on word identity to match the sentence pair, it cannot produce lexical paraphrases (e.g. *conduct* ↔ *performed* and *simulations* ↔ *experiments* in Figure 1,). The final edit sequence will mix paraphrase words (*performed* and *experiments*) and normal added words (*successful*) together under the same ADD label. In order to differentiate these two types of added words, we introduced special tags (REPLACE-S and REPLACE-E) to refer to lexical paraphrases specifically. During the edit label construction process, after checking the word pair identity for KEEP label, we additionally check whether they are aligned by our neural semi-CRF aligner, if so, we produce REPLACE-S/E tags, otherwise we do normal ADD/DEL tags. See Table 7 for a specific example. Word alignment can arbitrarily align any words in the target sentence, this can break the sequential dependency of the edit labels, we therefore discard some lexical paraphrases to guarantee such propriety (*conduct* ↔ *performed* in Table 7).

In order to show the effectiveness of our modified model, we compared two more versions of EditNTS in Table 8: EditNTS (original) + Aligner, where we directly add word alignment information to the original EditNTS model without any REPLACE tags; EditNTS (new), where we keep the REPLACE tags but don't use any word alignments. The results show that EditNTS model with REPLACE tags can improve the performance, but it is not significant. After adding the word alignment information, we can further improve the SARI score significantly, which can demonstrate the effectiveness of our modified EditNTS with aligner.

B More Details for MultiMWA Benchmark

B.1 Updated Annotation Guideline

After the first round of annotation, we discovery that the definition of phrasal alignment can be ambiguous, which will hinder the development and error analysis for word alignemnt models. There-

fore, we further extend the standard 6-page annotation guideline[11] from (Callison-Burch et al., 2006) to cover three linguistics phenomena to improve the phrase-level annotation consistency.

- "a/an/the + noun" should be aligned together with noun if both nouns are same.
- noun_1 should be only aligned to noun_1 in the phrase "noun_1 and noun_2".
- noun should be only aligned to noun in the "adjective + noun" phrase.

Utilizing the constituency parser implemented in the AllenNLP package (Gardner et al., 2018), we first write a script to implement these rules and apply them to all the training/dev/test sets of MultiMWA-MTRef. Then, we manually go through both dev and test sets to further ensure the annotation consistency.

B.2 Statistics of Alignment Shape

We also analyze the shape of alignment in each dataset, and the statistics can be found in Table 9. Statistical result showes that the dev and test of MultiMWA-MTRef contain a similar portion of phrasal alignment, and less than the training set. There even exists 1×10 alignment annotations in MultiMWA-MTRef, which are actually correct based on our manual inspection. Both MultiMWA-Newsela and MultiMWA-arXiv contain significantly larger portion of 1×1 alignment, especially the latter one contains only 3.2% of phrasal alignment.

		#pairs	1×1	%of 1×1	1×2	1×3	1×4	1×5	1×6	1×7	1×8	1×9	1×10	2×2	2×3	2×4	2×5	2×6	2×7	2×8	2×9	2×10	3×3	3×4	3×5	3×6	3×7	3×8	3×9	4×4	4×5	4×6	4×8	4×9	5×5	5×7	5×8	6×6
MultiMWA-MTRef	Train	2398	30371	59.50%	7956	3423	1648	530	198	112	48	0	10	1048	1536	936	270	204	84	16	18	20	549	612	360	198	63	24	27	176	180	72	32	72	25	70	80	72
	Dev	800	10350	65.65%	2434	876	408	100	30	14	16	9	10	308	474	144	60	48	0	0	0	0	135	156	105	0	0	0	0	48	40	0	0	0	0	0	0	0
	Test	800	10329	63.17%	2650	1092	404	175	36	28	24	0	0	312	516	200	50	72	0	0	0	0	81	144	75	36	0	0	0	64	40	24	0	0	0	0	0	0
MultiMWA-Newsela		500	8464	74.82%	838	483	180	80	48	35	8	36	0	152	276	144	40	24	28	16	18	0	117	180	45	18	63	0	0	0	20	0	0	0	0	0	0	0
MultiMWA-arXiv		200	5020	96.80%	78	30	8	0	0	0	0	0	0	8	12	0	0	0	0	0	18	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 9: Statistics of alignment shapes in each dataset. Each number represents how many word alignments are included for phrasal alignment with specific shape. For example, one 2×3 phrasal alignment will contribute to six word alignments. %of 1×1 is calculated by 1×1 over the sum of the row.

[11] http://www.cs.jhu.edu/~ccb/publications/paraphrase_guidelines.pdf