SpanAlign: Sentence Alignment Method based on Cross-Language Span Prediction and ILP

Katsuki Chousa and **Masaaki Nagata** and **Masaaki Nishino**
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan
{katsuki.chousa.bg, masaaki.nagata.et, masaaki.nishino.uh}@hco.ntt.co.jp

Abstract

We propose a novel method of automatic sentence alignment from noisy parallel documents. We first formalize the sentence alignment problem as the independent predictions of spans in the target document from sentences in the source document. We then introduce a total optimization method using integer linear programming to prevent span overlapping and obtain non-monotonic alignments. We implement cross-language span prediction by fine-tuning pre-trained multilingual language models based on BERT architecture and train them using pseudo-labeled data obtained from unsupervised sentence alignment method. While the baseline methods use sentence embeddings and assume monotonic alignment, our method can capture the token-to-token interaction between the tokens of source and target text and handle non-monotonic alignments. In sentence alignment experiments on English-Japanese, our method achieved 70.3 F_1 scores, which are +8.0 points higher than the baseline method. In particular, our method improved by +53.9 F_1 scores for extracting non-parallel sentences. Our method improved the downstream machine translation accuracy by 4.1 BLEU scores when the extracted bilingual sentences are used for fine-tuning a pre-trained Japanese-to-English translation model. [1]

1 Introduction

Sentence alignment is the task that automatically extracts parallel sentences from noisy parallel documents. Parallel sentences are used to train cross-language models, especially for machine translation (MT) systems. Both the quantity and quality of the parallel sentences used for training are crucial for developing an accurate neural machine translation (NMT) system (Khayrallah and Koehn, 2018).

Recently, automatic sentence alignment methods using neural networks have gained popularity (Grégoire and Langlais, 2018; Artetxe and Schwenk, 2019a; Yang et al., 2019; Thompson and Koehn, 2019). Such systems have a scoring function to calculate how the two sentences are parallel from sentence embeddings and obtain an alignment hypothesis from these scores with an alignment algorithm. These embeddings are obtained by separately encoding each source and target sentence. However, these prior works do not utilize context information and the interaction between the tokens of the source and target text although it may be useful for calculating the score. Moreover, alignment algorithms of these works assumes monotonic alignments although not every alignment of noisy bilingual documents is monotonic. Quan et al. (2013) described a legislation corpus as an example of non-monotonic alignments. In such cases, the existing methods that assume monotonicity impair the accuracy.

In this paper, we propose a novel sentence alignment method, called SpanAlign. That can be decomposed into two methods: scoring and optimization.

We formalize the scoring method as a cross-language span prediction model, similar to models for the SQuAD 2.0 question answering task (Rajpurkar et al., 2016). Figure 1 shows an example.

[1] Our implementation and datasets will be available at https://github.com/nttcslab-nlp/spanalign.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4750–4761
Barcelona, Spain (Online), December 8-13, 2020

In meteorology, precipitation is any product of the condensation of atmospheric water vapour that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail...
Q. What causes precipitation to fall?
A. **gravity**

I said, "Would you go to governments and lobby and use the system?" He said, "No, I'd take to the individuals." **It's all about the individuals.** It's all about you and me. It's all about partnerships...
Q. 全ては個人についてのことであり
A. **It's all about the individuals.**

Figure 1: Example of SQuAD-style monolingual question answering (upper) and sentence alignment task from bilingual document based on cross-language span prediction (lower). In sentence alignment, Q. denotes a source sentence, and A. denotes a target text that corresponds the source sentence in a given target document.

In sentence alignment, given a target text and source sentences, the model predicts a translation of the source text as the answer, which is a span in the target text, and calculates its score. We used pre-trained multilingual language models based on BERT architecture (Devlin et al., 2019) and trained them using only pseudo-labeling data obtained from existing sentence alignment methods. One advantage of our approach is its adoption of cross-attention, which can capture context information and the token-to-token interaction between input sentences. Moreover, this approach could work well on various language pairs without human-annotated data.

Since this span prediction method independently predicts the target spans for source sentences, the target spans might overlap. Moreover, because this method is asymmetric, the source-to-target predictions might differ from the target-to-source ones. Therefore, we propose an optimization method that introduces a total optimization method using integer linear programming (ILP), inspired by DeNero and Klein (2008) and Nishino et al. (2016). This method can use the predictions of both directions and extracts non-monotonic alignments.

We conducted sentence alignment experiments to evaluate the accuracy of our proposed method using an actual noisy newspaper dataset on English-Japanese articles, and our method significantly outperformed the previous works. We also evaluated on a downstream MT task and showed that the proposed method actually improved its performance.

2 Related Works

Previous sentence alignment methods are based on the context-independent similarity of source and target sentences, including sentence length (Gale and Church, 1993), bilingual dictionaries (Utsuro et al., 1994; Utiyama and Isahara, 2003; Varga et al., 2005), a machine translation system (Sennrich and Volk, 2011), and multilingual sentence embeddings (Thompson and Koehn, 2019). They usually use dynamic programming, which assumes that the alignments are monotonic. On the contrary, alignment methods using integer linear programming (ILP), which does not assume a monotonic alignment, have also been proposed. Nishino et al. (2016) formalized a sequence alignment problem as a set partitioning problem, which is a type of combinatorial optimization problem, and solved it by using ILP. DeNero and Klein (2008) showed that the problem of finding an optimal alignment can be cast as an ILP, which can quickly and reliably find the globally optimal alignment. Following these works, we use ILP for optimization to handle non-monotonic alignments.

Utiyama and Isahara (2003) proposed a method using the score of document alignment for sentence alignment. To obtain alignments for a document, they first translated a source document into a set of words from the target language using bilingual dictionaries. They then used each target document as a query and searched for the most similar source document in terms

of BM25 (Robertson and Walker, 1994). After that, they aligned the sentences in the aligned documents using DP matching (Gale and Church, 1993; Utsuro et al., 1994) based on similarity measure SIM, which is defined as the relative frequency of the one-to-one correspondence between the source and target words obtained from bilingual dictionaries. As a reliable measure for document alignment, they used AVSIM, which is the average of the SIMs obtained from the sentence pairs in the document pair. As a reliable measure for sentence alignment, they used the product of the document similarity AVSIM and the sentence similarity SIM. Use of document similarity for sentence alignment performed robustly for predicted documents alignments which are not always parallel. This method is commonly used for building publicly available English-Japanese parallel corpora, including the shared task data for the NTCIR Patent Translation[2] and the Workshop on Asian Translation (WAT)[3].

In a recent work, Thompson and Koehn (2019) proposed a sentence alignment method, called Vecalign, which uses bilingual sentence embeddings (Artetxe and Schwenk, 2019a) and recursive DP approximation. They used a German-French test set and achieved state-of-the-art results. Their method also effectively works for low- and medium-resource language pairs with the Bible dataset and used for building the ParaCrawl corpus, which is one of the largest parallel corpus across 23 EU languages with English by crawling the web (Bañón et al., 2020).

Since the targets of previous works on the sentence alignment task were mainly among European languages, it is unclear whether these methods are effective on such distant language pairs as English and Japanese. In this paper, we also explore how well previous methods work well for such language pair.

3 Proposed Method

3.1 Cross-language Span Prediction for Scoring Alignments

We first formalize the sentence alignment problem as a cross-language span prediction task from source sentences into spans in a target document. Figure 2a shows an example. The cross-language span prediction task is defined as follows: Suppose we have a source document with N tokens $F = \{f_1, f_2, \ldots, f_N\}$ and a target document with M tokens $E = \{e_1, e_2, \ldots, e_M\}$. Given consecutive source sentences $Q = \{f_i, f_{i+1}, \ldots, f_j\}$ that spans (i, j) in source document F, the task must extract target text $R = \{e_k, e_{k+1}, \ldots, e_l\}$ that spans (k, l) in target document E. For sentence alignment, it is necessary to handle many-to-many sentence alignments. Since we input consecutive source sentences as the input span, we can handle both 1-to-1 and many-to-many alignment in the proposed framework.

To solve the span prediction task, the model chooses a span (k, l) of target text R that corresponding to source sentences Q in target document E. The score of span ω is the product of its start position probability p_1 and end position probability p_2, defined as:

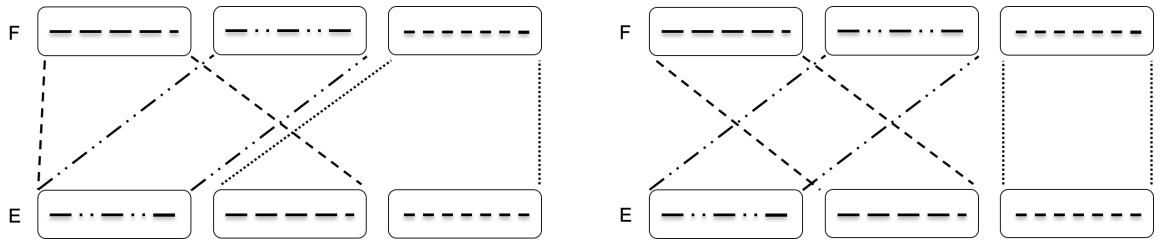
$$\omega_{ijkl} = \text{softmax}\left(p_1(k \mid E, Q) \cdot p_2(l \mid E, Q)\right). \quad (1)$$

Here, we apply softmax function for 10-best scores to fix the scale of score since these scores tend to be small values without softmax.

For calculating p_1 and p_2, we apply the pre-trained multilingual language representation models based on BERT architecture (Devlin et al., 2019). Even though these models were designed for monolingual language understandings tasks for several languages, we found that it also works surprisingly well for this cross-language span prediction task. In the model, the source sentences and the target document are concatenated to generate a sequence like "[CLS] source sentences [SEP] target document [SEP]" as input, where '[CLS]' and '[SEP]' are respectively classification and separator tokens. We add two independent output layers to the pre-trained model. These layers predict the probability of token indexes p_1 and p_2 in the target document that become the start and end of an output span.

[2] http://ntcir.nii.ac.jp/PatentMT-2/

[3] http://lotus.kuee.kyoto-u.ac.jp/WAT/



(a) Independent span prediction from each sentence in F. (b) ILP for consistent alignments.

Figure 2: Illustration of proposed method. (a) shows independent span prediction on scoring module and (b) shows global optimal alignments which are solved inconsistencies using ILP.

The best span $(\hat{k}, \hat{l})$ is chose by maximizing the score of a span ω_{ijkl}, as follows:

$$(\hat{k}, \hat{l}) = \underset{(k,l):1 \leq k \leq l \leq M}{\arg\max} \omega_{ijkl}. \tag{2}$$

Furthermore, we need to determine whether the target text corresponding to the source sentences exists since the actual noisy parallel corpora contain non-parallel sentences as noise. We address this problem with the values predicted at the position of '[CLS]' given source span (i,j) to calculate non-parallel score ϕ_{ij}. If this score exceeds the scores of the predicted spans, we assume that no corresponding target text exists.

Since the predicted spans are not necessarily agreed with sentence boundaries, we have to convert it into sentence sequence for sentence-level optimization and evaluation. Therefore, we identify the longest sentence sequence which is completely included in the predicted span, and regarded it as a sentence-level prediction.

3.2 ILP for Predicted Spans Optimization

In this section, we explain how to obtain many-to-many alignments using the scores that we obtained in the previous section. First, we assume a simple method that splits a source document into appropriate sentences, and we want to decide the best spans for them as predicted alignments. However, this method has three problems: (1) many of the predicted alignments overlap since the span prediction model independently predicts the target spans (like Figure 2a); (2) it remains unclear how to select appropriate input spans in the source document; (3) prediction from a source document are probably different from those from a target document because the span prediction model is asymmetric. Therefore, we formalize this problem as a total optimization method that can prevent spans from inconsistencies and maximize the sum of predicted scores in both the uni-directional and bi-directional scores. Symmetrizing the span scores of two directions makes the span prediction more reliable and bi-directional scores are expected to improve the alignment accuracy.

For the total optimization of sentence alignment, dynamic programming (Gale and Church, 1993) is commonly used although it assumes monotonic alignment between source and target sentences. However, not all alignments are monotonic, especially for noisy bilingual documents. Therefore, to obviate this assumption, we use a modified version of a previous method (DeNero and Klein, 2008; Nishino et al., 2016) because it can handle both the non-monotonic alignment and the null alignment of continuous segments using ILP. We formalize this problem for predicting a corresponding target span for a given source span using a neural network and find the best non-overlapping collection of spans using ILP (see Figure 2b). We convert the predicted scores into costs and minimize the sum of these costs instead of maximizing the sum of the scores because it achieved higher accuracy in the preliminary experiments.

We defined score ω_{ijkl} for the target span (k, l) given source span (i, j), which is obtained from the scoring method. By exchanging the source and target documents in the model, we also define score ω'_{ijkl} for the same span pairs. We respectively define non-parallel scores ϕ_{ij} for the

source span (i, j) and ϕ'_{kl} for the target span (k, l), which mean the spans are not translated in the other language. Let a_{ijkl} be a pair of span (i, j) in source document F and span (k, l) in target document E. We can then define the bilingual alignments as a set of span pairs, where there is no overlap for any span pairs in the set. The following is the ILP formalization:

$$\text{Minimize} \quad \sum_{ijkl} c_{ijkl}\ y_{ijkl} + \sum_{ij} \phi_{ij}\ b_{ij} + \sum_{kl} \phi'_{kl}\ b'_{kl} \tag{3}$$

$$\text{Subject to} \quad y_{ijkl}, b_{ij}, b'_{kl} \in \{0, 1\} \tag{4}$$

$$\sum_{i \le x \le j} \left[b_{ij} + \sum_{kl} y_{ijkl} \right] = 1, \qquad \forall x : 1 \le x \le N \tag{5}$$

$$\sum_{k \le x \le l} \left[b'_{kl} + \sum_{ij} y_{ijkl} \right] = 1, \qquad \forall x : 1 \le x \le M \tag{6}$$

where c_{ijkl} is the costs obtained from ω_{ijkl} and ω'_{ijkl}. y_{ijkl} is a binary variable that indicates whether span pair a_{ijkl} is included in the alignment with $y_{ijkl} = 1$ (Eq.4). b_{ij}, b'_{kl} is a binary variable that indicates whether source span (i, j) or target span (k, l) is non-parallel with $b_{ij} = 1$ or $b'_{kl} = 1$. Eq. 5 guarantees that for each sentence in source document F, there is at most one span pair a_{ijkl} in the alignment that includes the source sentence. Eq. 6 guarantees the same constraints for target document E. By combining the above two constraints, each sentence in E and F is guaranteed to be included at most once in the alignment.

As discussed in Thompson and Koehn (2019), sentence alignment should seek a minimal parallel pair. But we found that the optimization which directly uses ω and ω' tend to select many-to-many alignments. To relax this problem, we penalize the cost of many-to-many alignment pair by multiplying the average of the number of source and target sentences. We defined c_{ijkl} as follows:

$$c_{ijkl} = \frac{\text{nSents}(i, j) + \text{nSents}(k, l)}{2}(1 - \Omega_{ijkl}), \tag{7}$$

where $\text{nSents}(i, j)$ denotes the number of sentences in the span (i, j) $(= j - i + 1)$. Ω_{ijkl} is introduced to weighted average ω and ω' for bi-directional optimization:

$$\Omega_{ijkl} = \lambda\omega_{ijkl} + (1 - \lambda)\omega'_{ijkl}. \tag{8}$$

where λ is a hyperparameter that define the relative importance of the source-to-target and target-to-source scores. By setting $\lambda = 1$ or $\lambda = 0$, the optimization becomes uni-directional; by setting them to 0.5, the optimization becomes bi-directional.

The spans from one collection of source sentences that we obtain this way are close to a square of the number of sentences. To reduce the computational cost, we use the best spans for each input and filter out the others. In preliminary experiments, we found that the accuracy was not improved even if we used more than the 1-best spans for each input. This may be because the spans other than the 1-best one become noisy for optimization.

4 Experiments on Sentence Alignment Accuracy

We conducted the experiment to evaluate sentence alignment accuracy using an actual noisy parallel newspaper article dataset on English-Japanese. We evaluate two variants of our method exploiting different optimization methods: monotonic DP (Gale and Church, 1993) and ILP (§3.2) As a baseline, we adopted Vecalign (Thompson and Koehn, 2019), which achieved state-of-the-art results. We also compared our method with Utiyama and Isahara (2003) because it is the defacto standard approach for building English-Japanese parallel corpora.

For evaluation metrics, we used the F_1 score of sentence alignment, which is one of the standard metrics for sentence alignment[4]. The score was calculated by the correct and predicted alignment

[4]We used *strict* score from this script: `https://github.com/thompsonb/vecalign/blob/master/score.py`

pairs. However, this metric didn't directly evaluate the accuracy of extracting non-parallel sentences even though there are a lot of these non-parallel sentences in noisy bilingual documents. Therefore, for further detailed analysis, we also calculated the Precision/Recall/F_1 score for each number of source and target sentences in alignment pairs based on our implementation.

4.1 Implementation Details

All of our models described above were implemented using huggingface/transformers (Wolf et al., 2019). We used the base setting (12-layer, 768-hidden, 12-heads) of XLM-RoBERTa (Conneau et al., 2019) [5] for the span prediction and ILOG CPLEX 12.8.0.0 as an ILP solver. The parameters of the span prediction model are shared for both directions: source-to-target and target-to-source. The hyperparameters were set as follows: The learning rate was $3e-5$, the batchsize was 20, the number of training epochs was 5, the maximum sequence length was 384, the maximum length of the source sentences was 158, and the doc stride was 64. Since the number of input tokens for the models is limited, we adopted a sliding window approach to handle the long documents. In this approach, a window for the target document input, whose length is maximum sequence length minus length of source sentences minus three, slides with a stride of the doc stride. We then chose the best parameters with the highest F_1 score on the development set. We consider the alignment is composed of up to 4 total sentences for each language; that is we limited the number of source sentences to 4.

For Vecalign, we used the implementation provided by the authors[6] and set the alignment max size to 8 and the maximum number of allowed overlaps to 10. To obtain multilingual sentence embeddings, we used LASER (Artetxe and Schwenk, 2019b), which was pre-trained on 93 languages[7].

For Utiyama and Isahara (2003), we used our implementation. Sentences were split using sentence boundary symbols[8] with additional rules and tokenized by MeCab-UniDic[9] for Japanese and TreeTagger[10] for English. We used the following dictionaries for our experiment: an EDR Japanese-to-English dictionary, an EDR English-to-Japanese dictionary, and an EDR Technical Term dictionary[11]. The number of entries was 483,317 for the Japanese-to-English dictionaries and 367,347 for English-to-Japanese dictionaries.

4.2 Dataset

For the experiments on English and Japanese, we used a collection of newspaper articles from the Yomiuri Shimbun and their translations published in The Japan News (formerly the Daily Yomiuri), which is the newspaper's English edition. We purchased the newspaper's CD-ROMs for research purposes[12], and created automatically and manually aligned datasets.

We created 2,989 bilingual documents from 317,491 Japanese and 3,878 English articles published in 2012 and automatically extracted them using Utiyama and Isahara (2003) The number of English articles were 1.5% of Japanese articles and 60% of the English articles were aligned as bilingual documents. We used the sentence alignment as the training data, which were obtained using our implementation of Utiyama and Isahara (2003)[13].

As the development/test set, the manually aligned dataset consists of 157 bilingual document pairs obtained by manually searching through 182 English documents for the corresponding Japanese documents during two one-week periods: 2013/02/01-2013/02/07 and 2013/08/01-

[5] https://huggingface.co/xlm-roberta-base

[6] https://github.com/thompsonb/vecalign

[7] https://github.com/facebookresearch/LASER

[8] ！, ？, and 。 for Japanese and !, ?, :, ;, and . for English

[9] https://taku910.github.io/mecab/

[10] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[11] http://www2.nict.go.jp/ipp/EDR/ENG/indexTop.html

[12] https://database.yomiuri.co.jp/about/glossary

[13] In preliminary experiments, we found that using Utiyama and Isahara (2003) achieved better results than using Vecalign.

	English		Japanese	
	sentences	tokens	sentences	tokens
train	19.7	30.6	23.9	30.2
dev	17.9	32.2	18.9	29.9
test	23.8	31.5	26.2	28.0

Table 1: Average number of sentences and tokens for the dataset.

Methods	Direction	Precision	Recall	F_1
Vecalign	-	.591	.658	.623
Utiyama and Isahara (2003)	-	.604	.603	.604
SpanAlign (monotonic DP)	En-Ja	.644	.560	.599
	Ja-En	.666	.551	.603
	Bi-direction	.682	.611	.644
SpanAlign (ILP)	En-Ja	.690	.594	.638
	Ja-En	.720	.611	.661
	Bi-direction	**.734**	**.675**	**.703**

Table 2: Results of sentence alignment accuracies in English-Japanese.

2013/08/07. It consists of 131 articles and 26 editorials. We manually aligned the sentences for the 157 document pairs and obtained 2,243 many-to-many alignments. Among the manually aligned data, we used 15 articles for the development set, the next 15 articles for the test set, and the remaining articles and editorials as future reserve.

Table 1 shows the average number of sentences and tokens for each set. The sentences were tokenized using xlm-roberta-base model[14] from transformers.

4.3 Results

Table 2 shows the F_1 scores through whole alignment pairs. Our proposed method with both optimizations outperformed the baseline method. Bi-directional optimization using ILP achieved 70.3 F_1 scores, which are 8.0 points higher than Vecalign and 9.9 points higher than Utiyama and Isahara (2003). It shows our proposed scoring method works more effectively than the previous works. Bi-directional optimization using ILP also outperformed both the uni-directional optimizations and the method using monotonic DP. This indicates that our proposed optimization method can leverage both uni-directional predictions and works more effectively than monotonic DP.

Table 3 shows the results for each number of source and target sentences in alignment pairs. Our proposed method outperformed the previous work on all number of sentences pairs and the method using ILP achieved the highest F_1 scores except for the 1-to-2 alignment pairs. In particular, the F_1 scores of correctly identifying non-parallel source and target sentences are 80.0 and 95.1, which is significantly higher than those of Vecalign (26.1 and 79.5). We conjectured this is because our model explicitly classifies whether the sentence is parallel, and both directions are trained in one model. These results indicate that our proposed model can handle even non-parallel sentences with high accuracies and effectively work on noisy bilingual documents.

We used four NVIDIA Tesla K80 (12GB) for training the model, which took about 53 hours. Predicting spans for each input took about 1.9 seconds on a Tesla K80 GPU and 0.39 seconds was the average time of optimizing the predicted spans with ILP for a document in test set.

[14] `https://huggingface.co/xlm-roberta-base`

# of Japanese sentences	Vecalign Utiyama SpanAlign (DP, bidi) SpanAlign (ILP, bidi)	# of English sentences: 0	1	2	3
	0	-	1.00/.150/.261 1.00/.100/.182 .833/.750/.789 .933/.700/**.800**	-	-
	1	.967/.674/.795 1.00/.884/.938 1.00/.907/**.951** 1.00/.907/**.951**	.685/.739/.711 .652/.765/.704 .864/.661/.749 .881/.774/**.824**	.786/.702/.742 .667/.596/.629 .786/.702/.742 .795/.745/**.769**	.600/.300/.400 .167/.100/.125 .857/.600/**.706** .857/.600/**.706**
	2	-	.190/.444/.267 .125/.222/.160 .333/.333/**.333** .273/.333/.300	1.00/.167/.286 .400/.333/.364 .250/.333/.286 .333/.500/**.400**	-

Table 3: Experimental results for each number of source and target sentences in alignment pairs. Values in N-th row and M-th column denote Precision/Recall/F_1 scores of N-to-M alignment pairs. Hyphens indicate no alignment pairs in the test set.

5 Experiments on Machine Translation Task

Parallel sentences, which were extracted in the sentence alignment task, are important to train cross-language models, especially for building machine translation systems. To evaluate the effectiveness of our method on downstream MT tasks, we automatically and manually aligned the sentences on actual noisy newspaper article data in a distant language pair: Japanese-to-English. We compared our method to Vecalign (Thompson and Koehn, 2019) and Utiyama and Isahara (2003). We extracted parallel sentences using each method and randomly sample 300,000 sentence pairs to train the NMT models. Since this newspaper dataset is noisy and the model that only uses these datasets works poorly, we created two types of NMT models: one is only trained on extracted parallel sentences, the other is fine-tuned by the pre-trained model using the extracted dataset.

Moreover, to confirm the reliability of the sentence alignment costs c_{ijkl}, we took out various amounts of parallel corpora from the extracted sentence alignment in ascending order of their costs. We then fine-tuned the pre-trained model by using these parallel corpora and compared the translation accuracies of each method.

5.1 Dataset

We collected our next dataset from Yomiuri Shimbun and The Japan News, as in the previous experiments. For the training dataset, we used articles published from 1989 to 2015 (except for articles used in manually aligned dataset) and created 110,821 bilingual documents and, automatically aligned them by Utiyama and Isahara (2003). The settings or hyperparameters are the identical as those of the sentence alignment experiments. For our proposed method, we only used bi-directional optimization because it achieved the best results in the previous experiments. The manually aligned dataset, for the development and test set, is the same dataset used in the previous section. We created 162 parallel sentences from 15 articles for the development set and 238 parallel sentences from the next 15 articles for the test set.

Methods	Extracted Only	Fine-tuned
Utiyama and Isahara (2003)	0.8	14.2 (+2.7)
Vecalign	0.8	12.2 (+0.7)
SpanAlign (ILP)	4.1	18.3 (+6.8)

Table 4: BLEU scores for actual newspaper articles in Japanese-to-English. The values in parentheses are difference of BLEU gains against the pre-trained model.

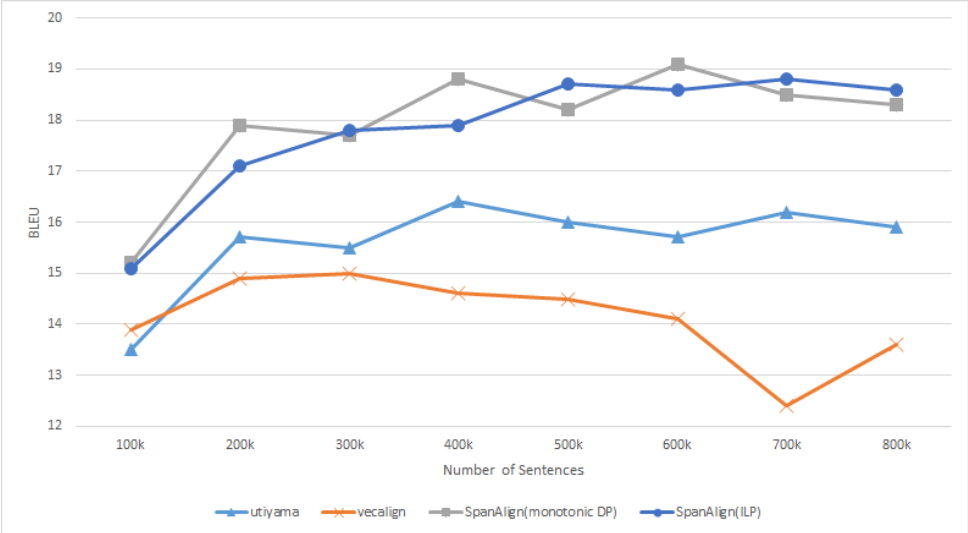


Figure 3: Comparing translation accuracies on various amounts of parallel sentence pairs.

5.2 Implementation Details

We preprocessed the data with SentencePiece (Kudo and Richardson, 2018) to create subword vocabularies and split the sentences into subword tokens. We set the vocabulary size to 16,000, which is shared between the source and target languages.

We created NMT models that used only extracted data using fairseq (Ott et al., 2019). These models are based on Transformer (Vaswani et al., 2017). We set the hyperparameters as follows: The size of the word embedding dimension was 512, the size of the feed-forward embedding was 1,024, the number of attention heads was 4, and the layers of encoder and decode was 6. For the training settings, we used Adam optimizer with $\alpha = 5e-4$, $\beta_1 = 0.9$, and $\beta_2 = 0.98$. The decoder's output embedding matrices are shared with its input embedding matrices. We adopted Inverse Square Root scheduler with a linear warmup of 4,000 steps to modify the learning rate. The dropout rate was set to 0.3 and the number of tokens in each mini-batch was set to 4,096 with accumulating the gradients of 8 mini-batches for update.

We used the pre-trained model using JParaCrawl (Morishita et al., 2020), whose corpus was created by largely crawling the web and automatically aligning parallel sentences. The authors of JParaCrawl released the pre-trained model which was trained using 8,763,995 parallel sentences. We fixed the models following the training settings and hyperparameters from their SentencePiece models and scripts for fine-tuning[15].

For the evaluation metrics, we calculated the BLEU scores (Papineni et al., 2002) with sacreBLEU (Post, 2018) to measure the translation accuracy.

5.3 Results

Table 4 shows translation accuracies of two types of models for each method. Our methods for both types of models outperformed the prior works. In particular, on fine-tuned models,

[15] https://github.com/MorinoseiMorizo/jparacrawl-finetune

our proposed method achieved 18.3 BLEU scores, which is 4.1 points higher than the next best score. These results indicate that our proposed method works well on automatically aligned documents and it is more useful for downstream tasks.

Figure 3 shows the results on various amounts of parallel sentence pairs. We can also see our methods outperformed the prior works on every amount of data. Focusing on small data sizes, BLEU scores of our methods increased sharply. This is probably because the alignment cost is reliable and the misaligned data are filtered out.

6 Discussion

We first discuss about the usefulness of the token-level interaction for sentence alignment. In the context of an information retrieval (IR) task, neural models are characterized as either representation-based or interaction-based, according to their architecture (Guo et al., 2016). The representation-based model maps a query and a document into low-dimensional spaces and calculates its similarity. The interaction-based model calculates the similarity through the interaction between the elements of a query and a document. Nie et al. (2018) shows that the interaction-based model generally outperforms the representation-based model on IR task. For sentence alignment, which is one of the cross-language IR tasks, the prior works resemble a representation-based model, and our scoring method resembles an interaction-based model. Our experiments, which show that our method outperformed the representation-based baseline models, indicate that the interaction between the source and target tokens is also useful for sentence alignment, as well as the IR task.

The proposed cross-language span prediction method can be used for any alignment scoring between two sequences. We have already applied it to word alignment (Nagata et al., 2020), and we would like to extend the method to other related problems.

7 Conclusion

We presented a novel method of automatic sentence alignment from noisy parallel documents based on cross-language span prediction and optimization using integer linear programming. Experimental results showed that our method significantly outperformed previous works and improved the performance of downstream MT tasks in English-Japanese.

In future work, we will apply our proposed method to other language pairs, use other multilingual pre-training models that support long sequence input, and incorporate with document alignment methods. Current sentence alignment methods assume sentence boundaries are given in advance and depend on the accuracy of sentence segmentation. However, sentence segmentation often creates ambiguity, especially for low-resource languages and those whose words are not delimited by white spaces, such as Chinese and Japanese. We are considering an extension of our method so that it doesn't use sentence boundaries.

References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In Proceedings of the ACL-2019, pages 3197–3203.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7:597–610.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4555–4567, Online, July. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

John DeNero and Dan Klein. 2008. The complexity of phrase alignment problems. In Proceedings of ACL-08: HLT, Short Papers, pages 25–28, Columbus, Ohio, June. 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. In Proceedings of the NAACL-2019, pages 4171–4186.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1):75–102.

Francis Grégoire and Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In Proceedings of COLING-2018, pages 1442–1453.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pages 55–64.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 74–83, Melbourne, Australia, July. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the EMNLP-2018, pages 66–71.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 3603–3609, Marseille, France, May. European Language Resources Association.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), Online, November. Association for Computational Linguistics.

Yifan Nie, Yanling Li, and Jian-Yun Nie. 2018. Empirical study of multi-level convolution models for ir based on representations and interactions. In Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '18, page 59–66, New York, NY, USA. Association for Computing Machinery.

Masaaki Nishino, Jun Suzuki, Shunji Umetani, Tsutomu Hirao, and Masaaki Nagata. 2016. Sequence alignment as a set partitioning problem. Journal of Natural Language Processing, 23(2):175–194.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of NAACL-HLT 2019: Demonstrations.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Xiaojun Quan, Chunyu Kit, and Yan Song. 2013. Non-monotonic sentence alignment via semisupervised learning. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 622–630, Sofia, Bulgaria, August. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of EMNLP-2016, pages 2383–2392.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Proceedings of the SIGIR-1994, pages 232–241.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011), pages 175–182, Riga, Latvia, May. Northern European Association for Language Technology (NEALT).

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In Proceedings of EMNLP-2019, pages 1342–1348.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In Proceedings of the ACL-2003, pages 72–79.

Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. 1994. Bilingual text, matching using bilingual dictionary and statistics. In Proceedings of the COLING-1994.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. 2005. Parallel corpora for medium density languages. In Proceedings of the RANLP-2005, pages 590–596.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the NIPS 2017, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In Proceedings of the IJCAI-2019, pages 5370–5378.