Sentence Alignment Scoring for Parallel Corpora in Low-Resource Language Machine Translation

Authors: Li Linxia, Chen Bo, Zhou Maoke, Zhao Xiaobing, Zhao Xiaobing

Date: 2024-06-05T00:00:00+00:00

Abstract

Objective To quantify sentence alignment scoring for parallel corpora in low-resource languages, acquire high-quality parallel corpora, and improve machine translation performance.

Method We propose NeuroAlign, a neural network-based unsupervised sentence embedding method for scoring sentence alignment in bilingual parallel corpora: it embeds parallel sentence pairs into the same vector space, computes alignment scores for given candidate sentence pairs in the parallel corpus, and then filters out lower-scoring parallel sentence pairs according to the ranking to obtain high-quality bilingual parallel corpora for low-resource languages.

Results In the BUCC2018 parallel text mining task, the F1 score can be improved by 0.5-0.8; in the CCMT2021 low-resource language neural machine translation task, the BLEU score can be improved by 0.1-10.9; sentence alignment scoring can approach human evaluation.

Limitations Limited by the scarcity of low-resource bilingual parallel corpora, exploratory research has not been conducted on language pairs beyond Tibetan-Chinese, Uyghur-Chinese, and Mongolian-Chinese.

Conclusion It can be effectively applied to sentence alignment scoring for parallel corpora in low-resource languages, improving corpus quality from the data source end and thereby enhancing machine translation performance.

Full Text

Parallel Corpus Sentence Alignment Scoring for Low-Resource Language Machine Translation

Li Linxia[1],[3], Chen Bo[2],[3], Zhou Maoke[1],[3], Zhao Xiaobing[2],[3]

[1](School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, China)
[2](School of Information Engineering, Minzu University of China, Beijing 100081, China)
[3](National Language Resource Monitoring and Research Center of Minority Languages, Minzu University of China, Beijing 100081, China)

Abstract:
[Objective] This paper aims to quantify sentence alignment scores for low-resource parallel corpora to obtain high-quality data and improve machine translation performance. [Methods] We propose NeuroAlign, a neural network-based unsupervised sentence embedding method for scoring bilingual parallel sentence alignment. Parallel sentence pairs are embedded into the same vector space, alignment scores for candidate pairs are calculated, and low-scoring pairs are filtered out based on these scores to obtain high-quality bilingual parallel corpora for low-resource languages. [Results] In the BUCC2018 parallel text mining task, F1 scores improved by 0.5-0.8; in CCMT2021 low-resource neural machine translation, BLEU scores improved by 0.1-10.9; sentence alignment scores approach human evaluation levels. [Limitations] Due to the scarcity of low-resource bilingual parallel corpora, exploration has not been conducted beyond Tibetan-Chinese, Uyghur-Chinese, and Mongolian-Chinese language pairs. [Conclusions] This method can be effectively applied to sentence alignment scoring for low-resource language parallel corpora, improving data quality at the source to enhance machine translation performance.

Keywords: Machine Translation; Low-Resource Language; Parallel Corpus; Sentence Alignment Scoring
Classification Number: TP393, G250
DOI: 10.11925/infotech.2096-3467.2024.0065

1 Introduction

Machine translation system training requires large-scale bilingual parallel text corpora with identical semantics (referred to as parallel corpora), and it is commonly believed that larger corpora yield better translation quality. However, beyond data scale, other quality factors also affect machine translation performance [1], such as domain distribution and sentence alignment quality. Research shows that alignment errors in parallel corpora impact statistical machine translation (SMT) systems [2], while untranslated and misaligned segments in neural machine translation (NMT) have even greater effects [3]. Therefore, sentence alignment quality is a crucial factor influencing machine translation system performance.

Early sentence alignment techniques based on feature engineering achieved some success, but this approach is relatively cumbersome and may capture inaccurate features, thereby affecting subsequent translation quality. With the develop-

ment of NMT, sentence alignment has received less attention than SMT and some supervised or semi-supervised NMT methods, though this does not diminish its impact on machine translation. Under the current popular "pre-training + fine-tuning" paradigm [4-5], while large-scale pre-trained language models perform well on machine translation tasks and seemingly reduce dependence on parallel corpora, most languages worldwide still lack large-scale, high-quality parallel corpora beyond resource-rich languages like Chinese and English. Most languages face a real dilemma of parallel corpus scarcity, and even existing low-resource languages that can be fine-tuned with small amounts of parallel data require guaranteed data quality.

Analysis of the CCMT2021 low-resource language parallel corpus reveals that even officially released high-quality corpora contain issues such as misalignment, untranslated segments, language errors, translation errors, sentence segmentation errors, and encoding errors. Inspired by parallel corpus mining tasks, this paper addresses the misalignment problem by proposing NeuroAlign (Neural Network-based Sentence Embedding Alignment Scoring Method for Bilingual Parallel Corpora), an unsupervised sentence embedding method based on neural networks for evaluating bilingual parallel corpus alignment quality. We aim to score and filter corpora for low-resource language machine translation.

This method first embeds bilingual parallel corpora into the same vector space, then calculates alignment scores by comparing the cosine similarity of given candidate sentence pairs (given candidate cosine) with the ratio difference to their k nearest neighboring candidates (nearby candidate cosine), while applying sentence length penalties to prevent overly short or long sentences from gaining undue advantage or disadvantage in scoring. Experiments demonstrate high performance on high-resource parallel corpus mining, low-resource NMT, and sentence alignment scoring tasks. Notably, this method requires no modifications to the translation system's model architecture—simply filtering high-quality parallel corpora at the data source improves machine translation performance.

2 Related Work

2.1 Parallel Corpus Alignment

Early parallel corpus acquisition relied on highly engineered systems [9], while later methods focused on text content through knowledge-based corpus collection [10] and parallel corpus mining [6], with sentence alignment being one such mining method.

Sentence alignment identifies correspondences between sentences in different languages within parallel corpora. Typically, an alignment scoring function is defined, and dynamic programming algorithms [11] maximize global alignment scores. Input comprises text pairs, outputting hypothesized sentence alignments. Early methods relied on statistical features such as sentence length [12-13], lexical information [14-15], or partial alignment [16]. Lexical-based methods are constrained by language-specific feature extraction, while length-

based methods perform poorly when sentence lengths are identical. With deep learning, neural network-based alignment methods emerged [17-20], achieving excellent results by leveraging vector similarity in embedding spaces [8,21-23], though specific embedding methods and alignment algorithms vary.

Melvin Johnson et al. [24] employed multilingual sentence embeddings to encode multiple languages for multilingual machine translation (one-to-many, many-to-one, many-to-many), requiring language tags before each language during encoding. Schwenk [8] learned joint multilingual sentence embeddings through a shared encoder without special tokens to indicate target languages, embedding full sentences from nine languages into a joint space and using distance thresholds between different language sentences to filter and mine parallel corpora. Unlike previous methods, this paper focuses on using sentence alignment methods to quantify alignment scores for bilingual parallel corpora, aiming to filter corpora for low-resource language machine translation and improve translation quality at the data source.

2.2 Sentence Alignment Scoring Metrics

Parallel corpus alignment scoring can employ human evaluation or automatic evaluation. Automatic evaluation falls into two categories:

First, when n pairs of translated words, sentences, or documents are known, alignment quality is typically measured by precision, recall, and F1-score.

Second, when the number of translated segments is uncertain, indirect metrics are used, such as translation evaluation metrics (e.g., Bilingual Evaluation Understudy, BLEU [25]) or cosine similarity.

Moore [14] and Varga [15] adopted translation alignment concepts, converting documents to the same language and introducing modified BLEU metrics to judge alignment quality or mine parallel corpora [5-6,19-20].

Sentence embedding methods typically use cosine similarity with fixed thresholds to measure alignment. Artetxe and Schwenk [6] noted this approach suffers from inconsistent cosine similarity score ranges, where incorrectly aligned sentence pairs may have higher cosine similarity than correctly aligned ones, making fixed-threshold filtering difficult. Consequently, researchers proposed various cosine similarity-based correction scores [6-7]. Building on Artetxe and Schwenk's algorithm [6], this paper applies smoothing to both source and target language sentence vectors, reducing the ratio difference between given candidate cosine and nearby candidate cosine in vector space for stricter alignment scoring.

3 NeuroAlign Methodology

This paper proposes NeuroAlign, a neural network-based unsupervised sentence embedding method for bilingual parallel corpus alignment scoring. First, parallel sentence pairs are embedded into the same vector space using neural sentence embedding methods. Then, alignment scores are calculated, and low-scoring

pairs are filtered to obtain relatively high-quality low-resource bilingual parallel corpora. The specific process is shown in Figure 1 [Figure 1: see original paper].

3.1 Neural Machine Translation

NMT is a sequence-to-sequence generation problem that automatically translates source language to target language through deep neural networks. NMT models consist of an encoder and decoder: the encoder converts source language sentences $x_1, ..., x_n$ into a continuous vector, while the decoder generates target language translations $y_1, ..., y_m$ from this vector. NMT training uses large-scale bilingual parallel corpora to optimize model parameters θ by maximizing the probability of target language sentences:

$$P(y_1, ..., y_m | x_1, ..., x_n; \theta) \tag{1}$$

Mainstream neural network models include Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and the Transformer [26] model based on attention mechanisms. This paper adopts the widely used Transformer model, which introduces self-attention and multi-head self-attention mechanisms to better capture contextual information. Transformer follows an encoder-decoder architecture with stacked layers, as shown in Figure 2 [Figure 2: see original paper].

Each encoder in Figure 2 contains two sub-layers: a multi-head self-attention layer and a feed-forward neural network layer, with residual connections and layer normalization applied to each sub-layer. The decoder includes the same two sub-layers plus an additional encoder-decoder attention layer that attends to encoder outputs, also using residual connections and layer normalization.

3.2 Multilingual Sentence Embedding

Multilingual sentence embedding maps full sentences from multiple languages into a joint vector space, using semantic distances between sentences to filter or mine parallel corpora. This paper uses the open-source tool LASER[2], which employs a single Bidirectional Long Short-Term Memory (BiLSTM) encoder unrestricted by language, without any input or output language signals, and shares a 40k Byte Pair Encoding (BPE) vocabulary [27] across all languages. The decoder receives embedded output language IDs at each time step. The encoder and auxiliary decoder jointly train a sequence-to-sequence system across multiple language pairs, as shown in Figure 3 [Figure 3: see original paper]. After training, the decoder is discarded, and fixed-length sentence representations are obtained through max-pooling over all encoder output states. This method uses fixed dimensions determined during pre-training, fine-tuning downstream tasks without backpropagation to the pre-trained model. The joint encoding across multiple languages provides cross-lingual consistency, making it suitable for parallel corpus mining.

3.3 Alignment Scoring

(1) Alignment Scoring

To overcome inconsistent cosine similarity score ranges, Artetxe and Schwenk [6] proposed a margin-based scoring method considering the ratio difference between a given candidate pair's cosine similarity (given candidate cosine) and its k nearest neighbors' cosine similarities (nearby candidate cosine). The calculation is as follows:

$$score(x, y) = \frac{cos(x, y)}{\sum_{i \in NN_k(x)} cos(x, i)} - \frac{cos(x, y)}{\sum_{j \in NN_k(y)} cos(j, y)} \tag{2}$$

where $NN_k(x)$ represents the k nearest target language sentence vectors to source vector x (excluding duplicate sentences), and $NN_k(y)$ represents the k nearest source language sentence vectors to target vector y, typically with $k = 4$.

A larger difference in formula (2) indicates higher alignment scores, meaning the two language sentences are semantically closer. However, since nearby candidates are not always clustered in the same vector space (as shown in Figure 4 [Figure 4: see original paper] where vector y's target nearby candidate cosine includes $2y - x$), we improved formula (2), detailed in formulas (3)-(4).

To reduce the difference between given candidate cosine and nearby candidate cosine in the embedding space, we applied smoothing corrections to both source and target languages, incorporating cosine similarities between given source/target sentence vectors and their own nearby candidates (excluding the source and target vectors themselves) into the nearby candidate cosines, as shown in formulas (3) and (4). Candidate examples are illustrated in Figure 5 [Figure 5: see original paper]. Additionally, formula (3) includes length penalty (LP) for parallel sentence pairs.

(2) Length Penalty

Experiments revealed that after applying alignment scoring, extremely short sentences received relatively high scores while long sentences scored lower, causing excessive filtering of long sentences during score-based sorting. We believe imbalanced proportions of overly long or short sentences in datasets interfere with machine translation models.

Table 1 shows examples of Tibetan-Chinese parallel sentence pair alignment scores. Without considering sentence length, short sentences like "要发展高新技术产业" (Tibetan: Æ è È) scored high, while longer sentences scored lower. To eliminate length interference, we applied sentence length penalties to all given parallel pairs, preventing overly short or long sentences from gaining excessive advantage or disadvantage in alignment scoring, as detailed in formula (5).

$$LP = \frac{l_{pair}}{l_{corpus}} \tag{5}$$

where LP represents the ratio of the given parallel pair's average length to the overall corpus average length. The numerator l_{pair} is the average length of the candidate parallel pair, while the denominator l_{corpus} is the average length of all parallel pairs in the corpus. Table 1 shows alignment score examples after considering length factors, demonstrating that length penalty reduces short sentence scores and increases long sentence scores, making the scoring more objective and preventing high scores from concentrating on low-contribution short sentences. Note that this penalty was only used in machine translation-related experiments.

(3) Candidate Strategy

When generating target nearby candidate sentence pairs, we adopted the same four strategies as Artetxe and Schwenk [6]:

- **Forward retrieval**: Each source sentence aligns with exactly one highest-scoring target sentence; some target sentences may align with multiple source sentences or none.
- **Backward retrieval**: Same as forward but in the opposite direction.
- **Intersection retrieval**: Intersection of forward and backward candidates, discarding inconsistently aligned sentences.
- **Max retrieval**: Combination of forward and backward candidates, selecting the highest-scoring candidate pair.

4 Experiments and Results Analysis

We conducted experiments on BUCC2018 parallel corpus mining[3], CCMT2021 machine translation, and bilingual sentence alignment scoring tasks.

4.1 BUCC Parallel Corpus Mining

The Building and Using Comparable Corpora (BUCC) task evaluates parallel corpus mining by identifying translated sentence pairs from comparable corpora in two languages. The task involves mining parallel sentences between English and four languages (German, French, Russian, Chinese), with 150k-1.2M sentences per language divided into sample, training, and test sets, containing approximately 2-3% parallel sentences.

For comparison with Artetxe and Schwenk [6], we evaluated only on English-German and English-French parallel corpus mining. Table 2 shows NeuroAlign's precision, recall, and F1 scores on the BUCC2018 training set.

Experimental results demonstrate that in English-German mining, NeuroAlign maintains precision above 95%, improves recall by 1-1.5, and increases F1 by 0.6-0.8. In English-French mining, precision remains around 92%, recall improves by

1.3-1.7, and F1 increases by 0.5-0.6. NeuroAlign's scoring facilitates better parallel corpus mining and filtering, validating its application performance. Since the "max" retrieval method performed best in Table 2, we uniformly adopted it for generating target nearby candidates in subsequent experiments.

4.2 Low-Resource Language Machine Translation

(1) Experimental Data and Model Parameters

We trained machine translation systems on Tibetan-Chinese, Uyghur-Chinese, and Mongolian-Chinese datasets from CCMT2021 bilingual translation tasks, using CCMT2017, CCMT2018, and CCMT2017 as validation sets respectively. Data scales are detailed in Table 3 . Tibetan sentences were syllable-segmented, Chinese sentences were character-segmented, and all language pairs underwent BPE processing.

For experimental settings, we used the Transformer NMT model from Facebook's open-source toolkit Fairseq[4], with 6-layer encoders and decoders, Adam optimizer (betas: 0.9, 0.98), learning rate 0.0005, dropout 0.3, batch size 4096, maximum 100 training epochs, beam size 4 for decoding, and other default Fairseq parameters.

(2) Baseline System Comparison

We compared four baseline systems representing different sentence alignment scoring methods for low-resource language pairs (Tibetan-Chinese, Mongolian-Chinese, Uyghur-Chinese), filtering low-scoring parallel pairs based on scores before training translation systems on the same NMT architecture:

- **Baseline**: Randomly sampled 95% of parallel pairs from deduplicated corpora
- **Cos**: Scored alignment using cosine similarity
- **Margin**: Used Artetxe and Schwenk's [6] margin-based scoring
- **NeuroAlign**: Scored using formula (3) without length penalty

To maintain corpus scale for low-resource NMT, we filtered only the bottom 5% lowest-scoring pairs. Results are shown in Table 4 .

Results show that corpus filtering improves NMT performance, confirming alignment quality's importance. Our proposed scoring method significantly enhances translation, particularly for Tibetan-Chinese and Uyghur-Chinese. Specifically:

- vs. Baseline: NeuroAlign improved BLEU by 0.31 and 10.87 on Tibetan-Chinese, 0.66 on Uyghur-Chinese, and 0.07 on Mongolian-Chinese CCMT2019.
- vs. Cos: NeuroAlign improved by 0.52 and 3.62 on Tibetan-Chinese, and 0.28 on Uyghur-Chinese.
- vs. Margin: NeuroAlign improved by 0.37 and 5.94 on Tibetan-Chinese, 0.07 on Uyghur-Chinese, and 0.05 on Mongolian-Chinese CCMT2017.

However, Baseline achieved better results on Mongolian-Chinese, with other methods causing performance degradation. Investigation revealed that after alignment scoring, short sentence pairs (length 1-5) scored high and were retained, while long pairs scored low and were filtered. These short pairs contribute insufficient semantic value and underutilize computational resources.

Table 5 shows that in the first 10k pairs of Mongolian-Chinese training data, Baseline's average short pair length exceeded other systems. After adding length penalty, Margin+LP and NeuroAlign+LP increased average short pair length by 1.16 and 1.15 respectively.

Figure 6 [Figure 6: see original paper] illustrates short pair distribution indices in the first 10k pairs, showing random uniform distribution in Baseline, but concentrated at the front after Cos/Margin scoring, preventing effective filtering. Length penalty reduced short sentences in Margin+LP and NeuroAlign+LP.

Thus, alignment scoring tends to give high scores to short sentences and low scores to long ones, excessively filtering long sentences and reducing semantic richness. To verify this, we conducted ablation studies with length penalty added to Margin and NeuroAlign.

(3) Ablation Study Comparison

Table 6 shows ablation results. Adding length penalty improved translation quality across all languages:

- Margin+LP vs. Margin: +0.5 and +2.66 on Tibetan-Chinese, +0.64 on Uyghur-Chinese, +0.92 and +0.3 on Mongolian-Chinese.
- NeuroAlign+LP vs. NeuroAlign: +0.61 and +0.53 on Tibetan-Chinese, +0.2 on Uyghur-Chinese, +0.34, +0.13, and +0.35 on Mongolian-Chinese.

Length penalty positively impacts alignment scoring. Figure 7 [Figure 7: see original paper] shows Mongolian-Chinese length distribution changes: after adding length penalty to NeuroAlign, short pairs (length 1-10) decreased by 2,422 while other length intervals increased by 1,471, 718, 128, 51, 25, 8, and 20 respectively. Length penalty effectively filters short sentences while retaining more semantically complete long sentences.

Table 7 shows filtered short sentence examples. These scored high under NeuroAlign but low after NeuroAlign+LP, as they carry insufficient semantic richness and can be safely filtered.

Table 8 compares translations from different ablation systems. With length penalty, NeuroAlign+LP produces translations closer to references, such as translating "原动力" as "首动力" (prime mover) and "务实合作" as "有成果合作" (fruitful cooperation) in Tibetan-Chinese, and "恐怖袭击" as "恐袭" (terror attack) in Uyghur-Chinese. Semantic capture also improves, as seen in Mongolian-Chinese where the preposition "从" (from) is accurately translated versus the possessive relationship captured by Margin. These results confirm length penalty's positive contribution.

4.3 Low-Resource Language Sentence Alignment Scoring

This task evaluated sentence alignment scoring for low-resource NMT training corpora. Table 9 compares automatic alignment scores with human evaluation on CCMT2021 Tibetan-Chinese, Mongolian-Chinese, and Uyghur-Chinese data using the "max" retrieval strategy.

Cosine similarity scores are generally low, Margin scores increase overall, while our proposed method scores decrease, indicating stricter evaluation closer to human judgment.

We manually extracted 100 test pairs from CCMT2021 Tibetan-Chinese data containing six misalignment types: misalignment, untranslated segments, language errors, translation errors, segmentation errors, and encoding errors, with each pair containing at least one error type. Native speakers annotated scores (6 points total, -1 per error type), with final scores averaged across two Tibetan native speakers. Table 10 shows automatic and human scores, demonstrating our method's scores are closer to human evaluation.

5 Conclusion

This paper proposes NeuroAlign, a neural network-based unsupervised sentence embedding method for bilingual parallel corpus alignment scoring. Experiments on parallel corpus mining, low-resource NMT, and alignment scoring demonstrate effective evaluation of low-resource bilingual parallel corpus quality. Filtering low-quality pairs based on scores significantly improves NMT performance, with applicability to high-resource parallel corpus mining as well. Limited by low-resource parallel corpus scarcity, we have not explored beyond Tibetan-Chinese, Uyghur-Chinese, and Mongolian-Chinese pairs. Future work will explore more language pairs and additional corpus quality metrics beyond alignment, as we believe alignment is only one important quality indicator among others that also affect machine translation.

References

[1] Koehn P, Knowles R. Six Challenges for Neural Machine Translation[C]//Proceedings of the First Workshop on Neural Machine Translation. 2017: 28-39.
[2] Goutte C, Carpuat M, Foster G. The Impact of Sentence Alignment Errors on Phrase-based Machine Translation Performance[C]//Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers. 2012.
[3] Khayrallah H, Koehn P. On the Impact of Various Types of Noise on Neural Machine Translation[C]//2nd Workshop on Neural Machine Translation and Generation. Association for Computational Linguistics, 2018.
[4] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805,

2018.
[5] Liu Y, Gu J, Goyal N, et al. Multilingual Denoising Pre-training for Neural Machine Translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.
[6] Artetxe M, Schwenk H. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
[7] Artetxe M, Schwenk H. Massively Multilingual Sentence Embeddings for Zero-shot Cross-lingual Transfer and Beyond[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 597-610.
[8] Schwenk H. Filtering and Mining Parallel Data in a Joint Multilingual Space[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 228-234.
[9] Shi L, Niu C, Zhou M, et al. A Dom Tree Alignment Model for Mining Parallel Data from the Web[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 2006: 489-496.
[10] Ma S, Zhang C. Automatic Collection of the Parallel Corpus with Little Prior Knowledge[C]//International Symposium on Natural Language Processing Based on Naturally Annotated Big Data. Cham: Springer International Publishing, 2014: 95-106.
[11] Richard Bellman. An Introduction to the Theory of Dynamic Programming[J]. Technical Report, RAND Corporation, Santa Monica, CA. 1953: 0104.
[12] Brown P F, Lai J C, Mercer R L. Aligning Sentences in Parallel Corpora[C]//29th Annual Meeting of the Association for Computational Linguistics. 1991: 169-176.
[13] Gale W A, Church K W. A Program for Aligning Sentences in Bilingual Corpora[J]. Computational Linguistics, 1994, 19(1): 75-102.
[14] Moore R C. Fast and Accurate Sentence Alignment of Bilingual Corpora[C]//Conference of the Association for Machine Translation in the Americas. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002: 135-144.
[15] Varga D, Halácsy P, Kornai A, et al. Parallel Corpora for Medium Density Languages[J]. Amsterdam Studies in the Theory and History of Linguistic Science Series 4, 2007, 292: 247.
[16] Simard M, Foster G F, Isabelle P. Using Cognates to Align Sentences in Bilingual Corpora[C]//Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages.
[17] Sennrich R, Volk M. MT-based Sentence Alignment for OCR-generated Parallel Texts[C]//In 9th Conference of the Association for Machine Translation in the Americas. 2010.
[18] Sennrich R, Volk M. Iterative, MT-based Sentence Alignment of Parallel Texts[C]//Proceedings of the 18th Nordic conference of computational linguistics. 2011: 175-182.
[19] Schwenk H, Chaudhary V, Sun S, et al. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational

Linguistics: Main Volume. 2021: 1351-1361.
[20] Guo M, Shen Q, Yang Y, et al. Effective Parallel Corpus Mining using Bilingual Sentence Embeddings[C]//Proceedings of the Third Conference on Machine Translation: Research Papers. 2018: 165-176.
[21] Hangya V, Braune F, Kalasouskaya Y, et al. Unsupervised Parallel Sentence Extraction from Comparable Corpora[C]//Proceedings of the 15th International Conference on Spoken Language Translation. 2018: 7-13.
[22] Thompson B, Koehn P. Vecalign: Improved Sentence Alignment in Linear Time and Space[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 1342-1348.
[23] Zhang W. Improve Sentence Alignment by Divide-and-conquer[J]. arXiv preprint arXiv:2201.06907, 2022.
[24] Johnson M, Schuster M, Le Q V, et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-shot Translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 339-351.
[25] Papineni K, Roukos S, Ward T, et al. Bleu: A Method for Automatic Evaluation of Machine Translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[J]. Advances in neural information processing systems, 2017, 30.
[27] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units[C]//54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2016: 1715-1725.
[28] Ott M, Edunov S, Baevski A, et al. FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling[C]//Proceedings of NAACL-HLT 2019: Demonstrations, 2019.

Corresponding author: Zhao Xiaobing, E-mail: nmzxb_{cn}@163.com.

Funding: This work is supported by The National Social Science Fund of China (Grant No. 22&ZD035).

Author Contributions:
Li Linxia: Designed research, experiments, paper writing;
Chen Bo: Proposed research ideas, paper revision;
Zhou Maoke: Corpus verification, data analysis;
Zhao Xiaobing: Final paper revision.

Conflict of Interest: All authors declare no conflict of interest.

Supporting Data:
[1] Building and Using Comparable Corpora dataset (BUCC2018). https://comparable.limsi.fr/bucc2018/bucc2
task.html
[2] Li Linxia. Low-resource Language Parallel Corpus Sentence Alignment Scoring Dataset. DOI:10.57760/sciencedb.j00133.00298.

Note: Figure translations are in progress. See original paper for figures.