Leveraging CoHere Multilingual Embeddings and Inverted Softmax Retrieval for Automatic Parallel Sentence Alignment in Low-Resource Languages

Auwal Abubakar Khalid
Bayero University, Kano
aka2000078.mcs@buk.edu.ng

Salisu Musa Borodo
Bayero University, Kano
salisuborodo@gmail.com

Amina Imam Abubakar
University of Abuja
amina.imam@uniabuja

Abstract

We present an improved method for automatic parallel sentence alignment in low- resource languages. We used CoHere multilingual embeddings and inverted softmax retrieval. Our technique achieved a higher F1-score of 78.30% on the MAFAND-MT test set, compared to the existing technique's 54.75%. Precision and recall have shown similar performance. We assessed the quality of the extracted data by demonstrating that it outperforms the existing technique in terms of low-resource translation performance.

1 Introduction

Because of a lack of large parallel datasets, neural machine translation systems trained on low-resource languages produce subpar results (Fernando et al., 2023; Haddow et al., 2022). Parallel corpora continue to be required for effective training of machine translation systems (Althobaiti, 2021; Yousef et al., 2022; Kaufmann, 2012; Paetzold et al., 2017; Resnik and Smith, 2003). Signoroni and Rychlỳ (2023); Haddow et al. (2022) found that neural machine translation is less reliable for language pairs with minimal resources. Even when parallel data is accessible, it is usually of lower quality or obtained from highly specialized sources, such as IT documentation or religious literature (Jaworski et al., 2023; Haddow et al., 2022; Ling et al., 2016). As a result, it cannot be used alone to accurately train general-purpose translation systems. To address this issue, parallel corpora can be mined from the internet (Fernando et al., 2023; Ling et al., 2016; Resnik and Smith, 2003). There are several plausible parallel sentences online, especially on multilingual news and instructional websites (Zhao et al., 2021; Riesa and Marcu, 2012; Makazhanov et al., 2018). Parallel corpora are typically constructed using automated sentence alignment approaches due to time and resource restrictions (Signoroni and Rychlỳ, 2023; Althobaiti, 2021; Hameed et al., 2016). Automatic sentence alignment is the technique of determining which sentences in a source text match to which sentences in a target text, allowing for the extraction of probable parallel sentences from big corpora (Chousa et al., 2020; Yousef et al., 2022; Brown et al., 1993). Several solutions have been presented. Multilingual sentence embedding-based methods have shown advantage in extending Natural Language Processing (NLP) tasks to a large number of languages, without the need to train a language-specific model (Signoroni and Rychlỳ, 2023; Heffernan et al., 2022; Chousa et al., 2020). Multilingual embeddings provide a universal foundation for sentence alignment that crosses linguistic boundaries (Artetxe and Schwenk, 2019). Using embedding-based methods, sentences in both languages can be represented in a single vector space so that sentences with semantic similarity are adjacent to each other in the vector space (Althobaiti, 2021). A recent study by Abdulmumin et al. (2023) shows that using closed-access CoHere multilingual embeddings resulted in a considerable improvement over earlier state-of-the-art LASER embeddings in parallel sentence alignment for low-resource languages. However, the authors implemented their alignment model using the standard nearest neighbor retrieval method. Although it is a simple and intuitive method for finding similar instances, standard nearest neighbor algorithm can only works well for relatively small datasets. It may suffer from hubness-where certain sentences tend to appear overly similar to many others in an embedding space (Dinu et al., 2015), and may not be effective in capturing complex relationships between sentences in different languages. Therefore, this research aims to propose an improved retrieval method—inverted softmax (Smith et al., 2017)—to enhance the alignment accuracy of the model, and consequently, improve translation quality.

2 Related Works

Research on automatic sentence alignment has evolved from early heuristic-based methods to modern neural approaches, especially in the context of low-resource languages. The methods can be broadly categorized into: (1) length and statistical-based approaches, (2) lexical and dictionary-based methods, (3) hybrid and alignment-tool frameworks, and (4) neural embedding-based approaches.

2.1 Length and Statistical-based Methods

Initial efforts relied heavily on sentence length as a proxy for alignment probability. Church (1993) proposed a character-length-based probabilistic model using dynamic programming. Brown et al. (1991) extended this with token counts and anchor points. Chen (1993); Papageorgiou et al. (1994) incorporated both sentence length and word identity. These methods assumed monotonic alignment and performed well in structured, clean data. Fung and McKeown (1994) introduced DK-Vec, targeting noisy parallel corpora using frequency and position heuristics. These early models laid the foundation for fast and language-independent alignment tools.

2.2 Lexical and Dictionary-based Methods

To address alignment ambiguity, lexical resources became central. McEnery et al. (1994) used approximate string matching, improving results from Kay and Roscheisen (1993). Hunalign (Varga et al., 2007) combined sentence-length statistics with bilingual dictionaries, refining alignments through iterative dictionary induction. Ma (2006) proposed Champollion, which weighted rare words more heavily for alignment but remained dependent on dictionary quality. Chen and Du (2003) tackled one-to-many word alignments in spoken corpora, while Resnik and Smith (2003) used web mining to extract parallel data. Other notable works include Melamed (2001) on idiom-aware alignment, and Semmar and Fluhr (2007) on cross-language information retrieval.

2.3 Hybrid and Tool-based Approaches

Several toolkits emerged combining statistical, linguistic, and rule-based methods. Deng and Byrne (2006) introduced MTTK, a language-independent toolkit for SMT training. JMaxAlign (Kaufmann, 2012) applied maximum entropy classifiers, while MASSAlign (Paetzold et al., 2017) targeted monolingual simplification. Efforts also explored specific domains: Ohmori and Higashida (1999) for Japanese-English collocations, and Volk et al. (2008) for treebank alignment. Some methods incorporated alignment correction and visualization tools (Macdonald, 2001; Cardon and Grabar, 2019). Recently, SpanAlign (Chousa et al., 2020) applied integer linear programming for non-monotonic alignments, and Stodden and Kallmeyer (2022) developed TS-ANNO for simplified corpora.

2.4 Neural Embedding-based Approaches

Recent advances have seen a shift toward embedding based models. VecAlign (Thompson and Koehn, 2019) combined dynamic programming with LASER embeddings (Artetxe and Schwenk, 2019) to align low-resource pairs such as Sinhala–English and Nepali–English. While effective, it suffered from misalignment in distant sentences. Paragraph-level filtering was later introduced to reduce this issue, as applied in Vietnamese–Lao alignment. Fine-tuning multilingual sentence embeddings has further improved performance. Chimoto and Bassett (2022) enhanced zero-shot alignment for Luhya–Swahili by fine-tuning LaBSE on a small Luhya parallel corpus, boosting accuracy from 22% to 85% with cosine similarity filtering. Abdulmumin et al. (2023) demonstrated that CoHere's closed-access multilingual embeddings significantly outperformed LASER on Hausa–English. Their method improved downstream MT quality, though it relied on basic nearest-neighbor retrieval, which the authors acknowledged as a limitation. We therefore, propose a better retrieval technique–inverted softmax (Smith et al., 2017)–due to its ability to mitigate hubness and capture complex patterns between sentences in different languages

3 Methodology

3.1 Alignment Workflow

Given parallel documents in two languages –source (s) and target (t), the task is to match sentences that are translations of each other. The parallel documents, are the ones in two languages containing similar information. Our sentence alignement workflow is shown in figure 1. Sentences are aligned in three (3) steps:

i. Sentence Vectorization: The first step is sentence representation. Each sentence in both the source (e.g., Hausa) and target (e.g., English) documents is mapped into a shared vec-

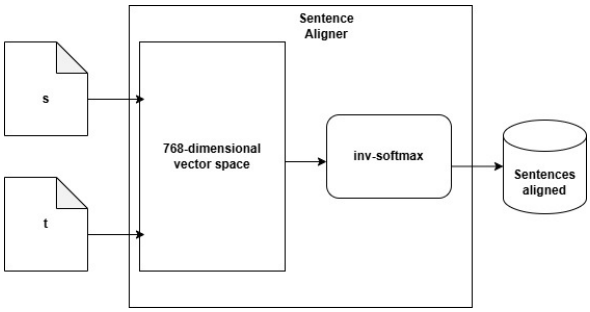

Figure 1: Sentence alignment workflow

tor space that captures their semantic features, using a pretrained multilingual sentence embedding model (CoHere multilingual embedding). This means that a Hausa sentence and its correct English translation should be close together in that space.

ii. Similarity measure: For each sentence pair, the cosine similarity is computed next. This measures how close two vectors are in direction. Cosine similarity is high (near 1) if sentences are likely translations of each other.

iii. Translation Retrieval: For each source sentence vector, retrieval algorithm (inverted softmax) searches for the most similar target sentence vector based on cosine similarity. It's like saying: "For this Hausa sentence, which English sentence is closest in meaning according to their vector representations?"

3.2 CoHere Multilingual Embedding Model

The 768-dimensional CoHere multilingual embedding model was developed to support a number of tasks, including cross-lingual zero-shot content moderation, multilingual semantic search, and customer feedback compilation, in more than 100 languages, including Hausa. This model can only be accessed via an API, which re quires authentication with an API key. This key is available for users to generate at their website [1].

3.3 Retrieval Algorithms

Retrieval algorithms are a key component of automatic parallel sentence alignment systems. They determine how semantically similar sentences across languages are identified in a shared embedding space. In this work, we compare the standard nearest neighbor retrieval with our proposed inverted softmax retrieval, which aims to improve alignment quality, especially in low-resource and noisy settings.

3.3.1 Standard Nearest Neighbor Retrieval

Nearest Neighbor (NN) retrieval is one of the simplest and most widely used methods for sentence alignment. For each source sentence embedding $\mathbf{x}_i$, the most similar target sentence $\mathbf{y}_j$ is selected based on cosine similarity:

$$s(\mathbf{x}_i, \mathbf{y}_j) = \frac{\mathbf{x}_i \cdot \mathbf{y}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{y}_j\|}$$

$$\mathbf{y}_j^* = \arg\max_{\mathbf{y}_j} s(\mathbf{x}_i, \mathbf{y}_j)$$

This approach is efficient and easy to implement, which makes it a popular baseline in multilingual alignment systems. However, it has notable limitations. In high-dimensional embedding spaces, certain sentences—often short or generic ones—tend to appear overly similar to many others, a phenomenon known as *hubness*. As a result, these "hub" sentences are frequently retrieved even when they are not true translations. Moreover, the search is one-directional and does not account for similarity normalization across the dataset.

3.3.2 Inverted Softmax Retrieval

To address these issues, we replace the nearest neighbor search with **Inverted Softmax (ISF)** (Smith et al., 2017) retrieval, a probabilistic method that introduces normalization over source embeddings. This helps counteract the hubness effect and yields more reliable alignments. Given a source embedding $\mathbf{x}_i$ and a target embedding $\mathbf{y}_j$, the probability that $\mathbf{x}_i$ corresponds to $\mathbf{y}_j$ is computed as:

$$P(\mathbf{x}_i \mid \mathbf{y}_j) = \frac{\exp\left(\beta \cdot s(\mathbf{x}_i, \mathbf{y}_j)\right)}{\sum_k \exp\left(\beta \cdot s(\mathbf{x}_k, \mathbf{y}_j)\right)}$$

where β is a temperature parameter controlling how sharply the similarities are weighted. Unlike the standard approach, ISF conditions on the target sentence by normalizing over all source embeddings. This means that a target sentence similar to many sources (a potential hub) receives lower relative scores, reducing its chance of being incorrectly aligned.

Advantages Over Nearest Neighbor

1. Reduces hubness: ISF penalizes overly generic targets that appear similar to many sources. **2. Improves precision:** It favors more distinctive and

[1] https://dashboard.cohere.com/api-keys

contextually relevant sentence pairs. **3. Remains efficient:** Despite the added normalization, ISF can be implemented efficiently with matrix operations. Overall, the inverted softmax retrieval provides a more balanced and accurate alignment mechanism, particularly useful when working with low-resource or noisy parallel data.

4 Experiment

4.1 Datasets

4.1.1 Crawled Data

We used the 1,000 most recent news stories from the Premium Times News [2] [3] website that had been crawled in both Hausa and English (Abdulmumin et al., 2023). To preprocess these data, we separated each collected document into a list of sentences using the Natural Language Tool Kit (NLTK) sentence tokenizer. The target and source files were then created by combining these sentences. Following tokenization using the NLTK's word tokenizer, we removed blank lines and sentences that were either shorter than five words or more than eighty words. Table 1 shows the statistics of the crawled sentences.

Language	Crawled Sentences	Cleaned Sentences
Hausa	13,916	13,560
English	23,148	22,671

Table 1: Statistics of Monolingual Hausa and English Sentences

4.1.2 MAFAND-MT

We evaluate the proposed and baseline aligners on the MAFAND-MT [4] (Adelani et al., 2022) dataset, a multilingual benchmark for African machine translation and alignment tasks. The dataset contains high-quality parallel sentences covering several African–English language pairs, professionally curated and cleaned from news and general-domain sources. In this work, we focus on the *English–Hausa* and *Hausa–English* subsets. Each direction contains parallel text divided into training, development, and test splits, following the official partitioning. The Hausa–English portion consists of several tens of thousands of aligned sentence pairs, with development and test sets typically around 1–2K examples each. This dataset provides a realistic evaluation setting for low-resource alignment due to the moderate corpus size, linguistic diversity, and domain variation.

4.2 Implementation of Sentence Aligners

In accordance with the baseline–nearest neighbor aligner (Abdulmumin et al., 2023), the evaluation script of vecmap[5] was modified to formulate the source-target sentence aligner. Employing inverted softmax retrieval, the aligner was created by utilizing the CoHere multilingual embedding model to transform both source and target sentences into a 768-dimensional vector. The CoHere embedding API, available for free, imposes a limit of approximately 6,000 sentence conversions to embeddings per minute. Consequently, to address this limitation, the CoHere sentence aligner was designed to pause for 61 seconds after processing a batch of source and target sentences. The batch size was set at 2,000 (or the remaining number of sentences), encompassing both the source and target sentences (totaling 4,000) at each iteration until obtaining embeddings for every sentence. To preserve the generated embeddings for potential future use, we save them to a file and upload it whenever the embedding of a previously converted sentence is required.

4.3 Evaluation

We use the MAFAND-MT (Adelani et al., 2022) datasets, which provide gold-standard target sentences, to compare the performance of the proposed inverted softmax aligner against the existing nearest neighbor aligner. This setup enables the use of precision, recall, and F1-score to objectively measure the quality of the aligned sentence pairs. For the evaluation, we focused on the English–Hausa subset of the MAFAND-MT train, development, and test sets. Furthermore, we utilized the labeled MAFAND-MT training data to train machine translation models in a semi-supervised manner, incorporating the automatically aligned crawled sentences. Specifically, we fine-tuned a publicly available checkpoint of the M2M-100[6] sequence-to-sequence model on the MAFAND-MT development set. The M2M-100 transformer architecture was designed to enable direct translation across 100 languages without requiring English as an intermediary. Following training, model perfor-

[2] https://www.premiumtimesng.com/
[3] https://hausa.premiumtimesng.com/
[4] https://github.com/masakhane-io/lafand-mt
[5] https://github.com/artetxem/vecmap
[6] https://huggingface.co/docs/transformers/model_doc/m2m_100

mance was evaluated using the sacreBLEU [7] metric on the MAFAND-MT test set.

5 Results and Discussion

5.1 Sentence Aligners

As shown in Table 2 the proposed sentence aligner outperforms the existing method across all datasets and evaluation metrics. Notably, it achieves substantial gains in precision (e.g., 74.9% vs. 46.8% on Dev) and recall (e.g., 80.5% vs. 55.4% on Dev), leading to consistently higher F1-scores. These improvements indicate the proposed method's effectiveness in retrieving more accurate and comprehensive parallel sentence pairs, thereby enhancing overall alignment quality.

Dataset	Existing Aligner			Proposed Aligner		
	P	R	F1	P	R	F1
Mafand-Dev	46.8	55.4	49.0	**74.9**	**80.5**	**76.7**
Mafand-Test	57.7	61.1	54.8	**76.6**	**81.8**	**78.3**
Mafand-Train	37.4	44.8	39.2	**67.6**	**73.8**	**73.7**

Table 2: Performance comparison of the existing and proposed sentence aligners on the Mafand dataset in terms of precision (P), recall (R), and F1 score (%). Bolded values indicate better performance.

5.2 Machine Translation

Table 3 displays the performances of the models that were trained for English to Hausa and Hausa to English translation directions. In both directions, it is evident that the proposed aligner aligned sentences are more advantageous to the translation models than the existing aligner aligned sentences, with an increase of 0.8 bleu points in English-Hausa translation, and 1.4 bleu points in Hausa-English translation.

Training Data	En→Ha	Ha→En
Existing Aligned Data	15.49	12.55
Proposed Aligned Data	**16.91**	**13.44**

Table 3: BLEU scores of translation models trained on data generated by the existing and proposed sentence aligners.

Limitations

In this study, we proposed an improved sentence alignment method for low-resource languages, leveraging CoHere's multilingual embeddings and inverted softmax retrieval. The proposed aligner consistently outperformed the existing method in precision, recall, and F1-score across all datasets, demonstrating robust and balanced performance. It proved more effective in identifying accurate parallel sentences, which translated to improved BLEU scores in English–Hausa and Hausa–English machine translation tasks. These results highlight the importance of retrieval strategies in sentence alignment quality. Future work will explore extending this technique to other low-resource African languages to support broader multilingual NLP efforts.

[7] https://huggingface.co/spaces/evaluate-metric/sacrebleu

References

Idris Abdulmumin, Auwal Abubakar Khalid, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Lukman Jibril Aliyu, Babangida Sani, Bala Mairiga Abduljalil, and Sani Ahmad Hassan. 2023. Leveraging closed-access multilingual embedding for automatic sentence alignment in low resource languages. *arXiv preprint arXiv:2311.12179*.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, and 26 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *Preprint*, arXiv:2205.02022.

Maha Jarallah Althobaiti. 2021. A simple yet robust algorithm for automatic extraction of parallel sentences: A case study on arabic-english wikipedia articles. *IEEE Access*, 10:401–420.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *29th annual meeting of the association for computational linguistics*, pages 169–176.

Rémi Cardon and Natalia Grabar. 2019. Parallel sentence retrieval from comparable corpora for biomedical text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 168–177.

Boxing Chen and Limin Du. 2003. Preparatory work on automatic extraction of bilingual multi-word units from parallel corpora. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 2, August 2003*, pages 77–92.

Stanley F Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

Everlyn Asiko Chimoto and Bruce A Bassett. 2022. Very low resource sentence alignment: Luhya and swahili. *arXiv preprint arXiv:2211.00046*.

Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. 2020. Spanalign: Sentence alignment method based on cross-language span prediction and ilp. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761.

Kenneth Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 1–8.

Yonggang Deng and Bill Byrne. 2006. Mttk: An alignment toolkit for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*, pages 265–268.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. *Preprint*, arXiv:1412.6568.

Aloka Fernando, Surangika Ranathunga, Dilan Sachintha, Lakmali Piyarathna, and Charith Rajitha. 2023. Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. *Knowledge and Information Systems*, 65(2):571–612.

Pascale Fung and Kathleen McKeown. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. *arXiv preprint cmp-lg/9409011*.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Riyafa Abdul Hameed, Nadeeshani Pathirennehelage, Anusha Ihalapathirana, Maryam Ziyad Mohamed, Surangika Ranathunga, Sanath Jayasena, Gihan Dias, and Sandareka Fernando. 2016. Automatic creation of a sentence aligned sinhala-tamil parallel corpus. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 124–132.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *Preprint*, arXiv:2205.12654.

Rafał Jaworski, Sanja Seljan, and Ivan Dunđer. 2023. Four million segments and counting: Building an english-croatian parallel corpus through crowdsourcing using a novel gamification-based platform. *Information*, 14(4):226.

Max Kaufmann. 2012. Jmaxalign: A maximum entropy parallel sentence alignment tool. In *Proceedings of COLING 2012: Demonstration papers*, pages 277–288.

Martin Kay and Martin Roscheisen. 1993. Text-translation alignment. *Computational linguistics*, 19(1):121–142.

Wang Ling, Luis Marujo, Chris Dyer, Alan W Black, and Isabel Trancoso. 2016. Mining parallel corpora from sina weibo and twitter. *Computational linguistics*, 42(2):307–343.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *LREC*, pages 489–492.

Kirsty Macdonald. 2001. Improving automatic alignment for translation memory creation. In *Proceedings of Translating and the Computer 23*.

Aibek Makazhanov, Bagdat Myrzakhmetov, and Zhenisbek Assylbekov. 2018. Manual vs automatic bitext extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Anthony M McEnery, Michael P Oakes, and Roger G Garside. 1994. The use of approximate string matching techniques in the alignment of sentences in parallel corpora. In *Proceedings of the Second International Conference on Machine Translation: Ten years on*.

I Dan Melamed. 2001. *Empirical methods for exploiting parallel texts*. MIT press.

Kumiko Ohmori and Masanobu Higashida. 1999. Extracting bilingual collocations from non-aligned parallel corpora. In *Proceedings of the 8th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. Massalign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4.

Harris Papageorgiou, Lambros Cranias, and Stelios Piperidis. 1994. Automatic alignment in parallel corpora. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 334–336.

Philip Resnik and Noah A Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Jason Riesa and Daniel Marcu. 2012. Automatic parallel fragment extraction from noisy data. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 538–542.

Nasredine Semmar and Christian Fluhr. 2007. Arabic to french sentence alignment: Exploration of a cross-language information retrieval approach. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 73–80.

Edoardo Signoroni and Pavel Rychlỳ. 2023. Evaluating sentence alignment methods in a low-resource setting: an english-yorùbá study case. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 123–129.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *Preprint*, arXiv:1702.03859.

Regina Stodden and Laura Kallmeyer. 2022. Ts-anno: an annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th annual meeting of the association for computational linguistics: system demonstrations*, pages 145–155.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1342–1348.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Martin Volk, Torsten Marek, and Yvonne Samuelsson. 2008. Human judgements in parallel treebank alignment.

Tariq Yousef, Chiara Palladino, David J Wright, and Monica Berti. 2022. Automatic translation alignment for ancient greek and latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107.

Shiyu Zhao, Xiaopu Li, Minghui Wu, and Jie Hao. 2021. The mininglamp machine translation system for wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 260–264.