Evaluating Sentence Alignment Methods in a Low-Resource Setting: an English-Yorùbá study case

Edoardo Signoroni and **Pavel Rychlý**
NLP Centre
Faculty of Informatics
Masaryk University
e.signoroni@mail.muni.cz, pary@fi.muni.cz

Abstract

Parallel corpora are still crucial to train effective Machine Translation systems. This is even more true for low-resource language pairs, for which Neural Machine Translation has been shown to be less robust to domain mismatch and noise. Due to time and resource constraints, parallel corpora are mostly created with sentence alignment methods which automatically infer alignments. Recent work focused on state-of-the-art pre-trained sentence embeddings-based methods which are available only for a tiny fraction of the world's languages. In this paper, we evaluate the performance of four widely used algorithms on the low-resource English-Yorùbá language pair against a multidomain benchmark parallel corpus on two experiments involving 1-to-1 alignments with and without reordering. We find that, at least for this language pair, earlier and simpler methods are more suited to the task, all the while not requiring additional data or resources. We also report that the methods we evaluated perform differently across distinct domains, thus indicating that some approach may be better for a specific domain or textual structure.

1 Introduction

Parallel corpora are vital training data for Machine Translation (MT) systems, especially for low-resource languages where data is scarce (Steingrímsson et al., 2020). While unsupervised methods trained only on monolingual data have been proposed for Neural MT, they are still sensitive to noise and domain mismatch (Khayrallah and Koehn, 2018), and are outperformed by supervised and semi-supervised systems trained on relatively small parallel corpora (Kim et al., 2020). Collecting and curating data for the creation of a parallel corpus manually is a costly and time consuming task that requires expertise in the languages involved. It is even more difficult for low to no-resource languages, for which the number of speakers and research may be lower.

Thus, today parallel corpora are mostly created by employing automatic methods for sentence alignment. Sentence alignment is the task of taking parallel documents split into sentences and finding a bipartite graph which matches minimal groups of sentences that are translation of each other (Thompson and Koehn, 2019). In other words, to find target sentences with the same meaning to that of the source segments in multilingual texts (Steingrímsson et al., 2020). Several approaches have been proposed, from simple length-based algorithms (Gale and Church, 1993), to more complex methods employing multilingual sentence embeddings (Thompson and Koehn, 2019).

Our work evaluates four commonly used sentence alignment methods using Menyo20k (Adelani et al., 2021), a high-quality benchmark English-Yorùbá[1] parallel corpus, as reference. We experiment with 1-to-1 alignments with and without reordering. Our results show that, at least for this language pair, earlier, simpler systems may be more suited, as they perform better and do not require other data than the documents to be aligned, allowing them to be employed even when no other text or knowledge of the languages is available. Moreover, we leverage the domain annotation of the Menyo20k corpus to observe that the alignment methods in our evaluation perform differently across various domains. This indicates that some approach may be better suited to a specific domain or textual structure.

After this Introduction, we report on some recent work aimed at evaluating and improving sentence alignment for low-resource language pairs in Section 2. Then, in Section 3 we describe briefly

[1]Yorùbá is the third most spoken language in Africa, with 40 million native speakers. It is native to south-western Nigeria and the Republic of Benin and belongs to the Niger-Congo family.

the alignment methods in our evaluation, which methodology is outlined in Section 4. Section 5 reports the results of our experiments and Section 6 draws some conclusions.

2 Related Work

Some recent work deals with sentence alignment in a low-resource setting, focusing on evaluating and improving modern sentence embedding-based methods.

Tien et al. (2021) finds that Vecaling has several limitations: it performs poorly in aligning sentences which are located far apart in source and target documents, or it may align sentences which are not translations of one another, but have a high similarity score. Moreover, the error can be propagated to from one pair to another, since the sentence that should be in one alignment is moved to a further one. They propose a new method that overcomes these limitations by firstly aligning paragraphs and generating candidate sentence pairs only among the aligned paragraph's sentences. They work with the Vietnamese-Lao low-resource language pair by translating the Lao documents to Vietnamese, on which LASER has been trained. Then they find the sentence pairs with cosine similarity weighted for the ratio of text length and then retrieve the target sentence in the original language. They report significant improvements in precision and recall over Vecaling on their test set.

Chimoto and Bassett (2022) experiment with LASER and LaBSE (Feng et al., 2021) to extract bitext for two unseen low-resource African languages, Luhya and Swahili. They find that both pre-trained models perform poorly at zero-shot alignment on Luhya. They thus fine-tune the embeddings on a small set of parallel Luhya sentences and report significant gains, with the accuracy of LaBSE increasing from 22% to 53.3%. This is further improved to over 85% by restricting the datastet to sentence embedding pairs with cosine similarity above 0.7.

Fernando et al. (2022) evaluates the effectiveness of pre-trained language models such as LASER, XLM-R (Conneau et al., 2020), and LaBSE on document and sentence alignment in the context of the low-resource languages of Sinhala, Tamil, and English. They introduce a weighting mechanism based on small-scale bilingual lexicons to improve the semantic similarity measure used by the methods they evaluate, thus improving the resulting alignments. Contrarily to our work, they find that the multilingual sentence embeddings-based methods significantly outperform the Hunalign baseline on their test language pairs. This discrepancy should be investigated in future work.

3 Baselines

The works introduced in Section 2 deal mostly with sentence embeddings-based methods.[2] While these methods were shown to be effective and able to generalize to unseen languages in some instances (Thompson and Koehn, 2019; Conneau et al., 2020), this did not hold in other low-resource test cases, for which further work on both the system and the model was needed to reach a satisfactory performance (Chimoto and Bassett, 2022). Moreover, they are still not free from issues in handling sentences which are found far apart in the documents and employ non-optimal scoring functions, such as raw sentence embedding cosine similarity (Tien et al., 2021). Lastly, they still require resource-heavy pre-trained models which are available only for a tiny fraction of the world's languages. Thus, we also include earlier methods in our evaluation. Table 1 summarizes the methods we have taken into account.

The earliest widely documented statistical-based methods were explored by Gale and Church (1993) on the assumption that the length of a sentence is highly correlated with the length of its translation. Moreover, they concluded that there is a stable ratio between the sentence lengths in any two language. Their method assigns a probabilistic score to each correspondence of sentences, based on the scaled difference of lengths, in number of characters, of the two sentences and the variance of this difference. The score is used in a dynamic programming framework to find the maximum likelihood alignment of sentences. They worked on a trilingual corpus of economic reports issued by the Union Bank of Switzerland (UBS) in English, French, and German, and a bilingual sample of 90 million words from proceedings of the Parliament of Canada in English and French.

Varga et al. (2007) describe *Hunalign*, a hybrid method, combining a dictionary with a length-based approach. Hunalign starts by producing a crude word to word translation for each word token in the dictionary according to the token with the

[2] Fernando et al. (2022) cites Hunaling as one of their baselines, however.

Baselines	Scoring Function	Reference
Gale-Church	Sentence length	Gale and Church (1993)
Hunaling	Sentence length + dictionary	Varga et al. (2007)
Bleualign	Machine translation metric (BLEU)	Sennrich and Volk (2010, 2011)
Vecalign	Sentence embedding cosine similarity	Thompson and Koehn (2019)

Table 1: Summary of the sentence alignment methods in our evaluation.

highest frequency in the target corpus. The pseudo target language is then compared to the actual target text on a sentence to sentence basis with a similarity score based on the number of shared words, which is the heaviest component of the scoring, and the sentence length in characters based on the ratio of longer to shorter. Once the similarity matrix is obtained for the relevant pairs, dynamic programming is used to find the optimal alignment with penalties for skipping and coalescing sentences. The algorithm works even in the absence of a dictionary in which case the texts are first aligned with the source text as the crude translation of itself and then a simple dictionary can be bootstrapped by collecting source-target token pairs with an association score higher than 0.5. They mainly experiment on Hungarian, but cite also Romanian and Slovenian, motivated by the need to build parallel corpora for "medium density languages".

Sennrich and Volk (2010) presents *Bleualign*, an automatic alignment method based on MT. They propose to use automatically translated text and a measure of the quality of this translation, in this case BLEU (Papineni et al., 2002), as a similarity score to find reliable alignments to be used as anchor points. Sennrich and Volk (2011) details an iterative approach for Bleualign. They build a rough alignment using the Gale-Church algorithm and then train a first MT system on these aligned data. They then use the generated translations to compute the sentence level BLEU score and employ it as a measure of alignment. They work on a corpus of French and German text obtained by OCR from the yearbooks of the Swiss Alpine Club between 1864-1982. They claim the system to be more resilient to noise and fairly language independent, despite depending heavily on the provided translation, and thus on a MT system with reasonable performance for their language pair. This is problematic in resource-poor conditions due to the need for enough data to train the MT system and it is computationally more demanding due to the need of an automatic translation.

Thompson and Koehn (2019) presents *Vecalign*. They propose a sentence alignment scoring function based on the similarity of bilingual sentence embeddings, which has been shown to be effective in related tasks such as filtering non-parallel sentences and locating parallel sentences in comparable corpora. Moreover, blocks of sentences can be represented as the average of their embeddings, which does not depend on the number of sentences being compared, thus reducing the computational complexity. They use the LASER multilingual sentence embeddings (Artetxe and Schwenk, 2019) and compute similarity as cosine similarity, normalized with randomly selected embeddings to avoid hubness (Radovanovic et al., 2010; Lazaridou et al., 2015), i.e. the tendency of some vectors ("hubs") to appear in the top neighbour lists of many items. To align the text, they start by creating an approximate sentence alignment using the average embeddings of adjacent sentences. Then they refine this alignment with the original sentence vectors, limiting the search in a small window around the approximate alignment. They claim state-of-the-art results on the Bleualign dataset and on Bible test sets (Christodouloupoulos and Steedman, 2015). In this low-resource setting, they work on Arabic, Turkish, Somali, Afrikaans, Tagalog, and Norwegian. All these languages but Norwegian appear in the training data for LASER, albeit in different sizes. They consider *verse-alignments* as their gold-standard, for which they report an average improvement of 28 verse-level F_1 score on Hunalign in bootstrap mode. As we will show in Section 5, this improvement in performance does not hold in our experiments on English-Yorùbá.

4 Methodology

The objective of our work is to evaluate the widespread sentence alignment methods briefly described in Section 3 in a low-resource setting. To achieve this, we carry out two experiments on Menyo20k (Adelani et al., 2021), a high quality English-Yorùbá multidomain parallel corpus. Over-

Shorthand	N of sentences	Data source
book	2014	"Out of His Mind" Book
cc	193	Creative Commons license
digital	941	ICT/digital & Kolibri Tech sentences
jw	3508	JW news
misc	687	Short text from various domains
movie	774	Movie transcript
news	5980	News articles
proverbs	2700	Yoruba proverbs
radio	258	Radio transcripts
tedTalks	2945	Ted Talks transcripts
udhr	100	Universal Declaration of Human Rights
menyo	**20100**	**TOTAL**

Table 2: Domains of the Menyo20k corpus and their sizes in number of sentences.

all the dataset contains 20.100 sentences gathered from various domains such as news articles, TED talks, movie and radio transcripts, science and technology text, Yorùbá proverbs, books, and short articles curated from the web. Monolingual text crawled from the web were professionally translated and verified by native speakers. We thus assume the corpus as a a gold-standard for our experiments.

For our purposes we concatenate the train, dev, and test splits in which the corpus is divided into one text file containing 1-to-1 alignments. Table 2 gives the sizes of the corpus and its different domain splits.[3]

The first experiment, dubbed *NATURAL-ORDER*, is straightforward: we apply the alignment methods mentioned in Section 3 to each section of the corpus and on the corpus as a whole. We then evaluate the resulting alignments against the reference with an algorithm that iterates over both the proposed alignments and the reference to return a pair as correct only when the candidate alignment is identical to the one in the reference.

The second experiment, *SHUFFLED-ORDER*, is similar to *NATURAL-ORDER*, with the addition of reordering: we artificially shuffle the target side by randomly scrambling the sentences in a window of 3. More precisely, we start from sentences at lines 1 to 3 and we randomly shuffle them in this group; we then move on to sentences at lines 4 to 6, and scramble them as well. We continue in this manner until the end of the document is reached. This is done to avoid creating unrealistic data, since it is not usual for sentences that should be aligned to be very far apart in the translation of same text. We then proceed as in *NATURAL-ORDER*, by applying the alignment methods and evaluating their outputs against the gold standard.

Whenever possible, we used the implementations available online[4] with the configuration that required the least amount of pre-existing resources or further work, such as fine-tuning. For the Gale-Church method we employed the implementation provided with Bleualign. We use LASER (Artetxe and Schwenk, 2019) to compute the sentence embeddings for Vecalign. While the encoder for Yorùbá was provided in the library as part of the LASER3 extension (NLLB Team, 2022), we had to train our own `sentencepiece` (Kudo, 2018) model.[5] Hunalign was run without a precompiled dictionary. Since no end-to-end iterative implementation of Bleualign was found, we applied the method without a reference translation. We also attempted to train a NMT model only on the Menyo20k corpus aligned with the Gale-Church algorithm, as the Bleualign authors suggest in their second paper. For this, we trained a standard transformer (Vaswani et al., 2017) using `fairseq` (Ott et al., 2019) with the following parameters: vocabulary size 2000, *adam* optimizer, dropout 0.1, label smoothing 0.1, max tokens 4096, and optimizing for BLEU. After 60 epochs, however, the model failed to reach more than 5 BLEU in both translation directions, with its output hallucinated and noisy, and was thus deemed not useful to further alignment steps with Bleualign.

5 Results

Table 3 summarize the results of our evaluation.

The upper rows of the table report the results for *NATURAL-ORDER*, the simple 1-to-1 alignment without reordering. The Gale-Church baseline perform best in 8 out of 12 domains, with the percentage of correct alignments between 82.95% for the *radio* domain, and the 100% of *udhr*. It scores 99.96% on *menyo-all*.

Bleualign is the least performing method in 9 out of 12 domains, sharing its only 100% on *udhr* with all the other methods. It fares particularly badly for the literary domain, getting just 37.87% of alignments correctly for *book* and 56.07% on *proverbs*. On *menyo-all*, it returns 79.16% of correct align-

[3]For a full breakdown on the sources and data collection of the Menyo20k corpus, we defer to their paper.

[4]Hunalign: https://github.com/danielvarga/hunalign; Bleualign: https://github.com/rsennrich/Bleualign; Vecalign: https://github.com/thompsonb/vecalign

[5]We used the Yorùbá Wikipedia as training data and the same parameters for the other models in LASER3. Limiting the training data to the Menyo20k corpus failed to achieve the necessary vocabulary size needed by LASER3.

Split			book	cc	digital	jw	misc	movie	news	proverbs	radio	tedTalks	udhr	menyo-all	avg
Score	NAT	*bleu*	**37.87%**	**85.49%**	**91.73%**	**86.55%**	**84.86%**	**66.54%**	**90.72%**	56.07%	80.23%	**92.87%**	**100.0%**	**79.16%**	**79.34%**
		ga	99.6%	**100.0%**	**100.0%**	**99.43%**	**99.71%**	**100.0%**	**99.92%**	**99.93%**	82.95%	99.29%	**100.0%**	**99.96%**	**98.40%**
		hun	**99.85%**	**100.0%**	99.36%	97.86%	90.6%	97.67%	99.31%	**35.83%**	**100.0%**	**99.56%**	**100.0%**	90.6%	92.55%
		vec	97.72%	95.85%	94.17%	96.29%	97.82%	99.1%	97.98%	89.12%	**78.68%**	98.64%	**100.0%**	94.4%	94.99%
	SHF	*bleu*	**11.07%**	35.75%	28.21%	38.0%	**27.07%**	**22.09%**	41.94%	18.84%	35.27%	31.23%	**21.0%**	31.36%	28.48%
		ga	24.22%	33.68%	25.77%	**24.2%**	30.28%	32.04%	**24.52%**	**25.24%**	**21.71%**	**25.7%**	**21.0%**	**25.26%**	**26.15%**
		hun	**34.99%**	**47.15%**	**36.9%**	**42.76%**	**38.86%**	**35.01%**	**45.95%**	**9.88%**	**39.92%**	**39.88%**	**48.0%**	**38.53%**	**38.15%**
		vec	26.1%	**32.12%**	**25.66%**	28.56%	29.69%	32.04%	29.07%	24.02%	**21.71%**	31.4%	41.0%	28.63%	29.17%

Table 3: Percentage of correct 1-to-1 alignments for each method and domain in *NORMAL-ORDER* (NAT) and *SHUFFLED-ORDER* (SHF). The abbreviations for the alignment methods are the following: *bleu* : Bleualign, *ga* : Gale-Church, *hun* : Hunalign, *vec* : Vecalign. The last column reports the average score for each method.

ments.

Hunalign and Vecalign perform similarly, with scores over 90% for most domains, and 90.6% for *menyo-all*. Hunalign fails for *proverbs*, correctly aligning only 35.83% of the sentences. The lowest score for Vecalign is on *radio*, with 78.68%.

It is apparent that the structured nature of the Universal Human Rights Declaration generally favours alignment. Conversely, the more fluid nature of *proverbs* may hamper methods such as Hunalign, which rely on lexical information for alignment. This domain, however, seems to be better handled using just length information.

The lower half of Table 3 reports the results for *SHUFFLED-ORDER*, 1-to-1 alignments with reordering. All methods fail to reach 50% of found correct alignments. Hunalign scores highest in 11 domains out of 12, achieving its best score of 48.0% on *udhr*. It also detains the lowest score of the experiment, 9.88% on *proverbs*. Hunalign correctly aligns 38.53% of *menyo-all*. The other methods all perform inadequately, with values close to random for the window of 3 chosen for reordering. Apart from the aforementioned Hunalign on *proverbs* other low outliers are the Bleualign scores on *book* and *proverbs*. Again, these domains seem to be more problematic, significantly hampering the systems. Moreover, reordering appears to invalidate the accuracy even on the highly structured text in *udhr*.

6 Conclusions

In this paper we presented an evaluation of four commonly used sentence alignment methods when applied to a low-resource language pair, such as English-Yorùbá.

While working well for high resource languages and domains, more recent sentence embedding-based alignment methods do not perform similarly for a low-resource pair such as the one in our study. Earlier methods, based on sentence length statistics and bootstrapped dictionaries, returned better alignments on the Menyo20k corpus. All of these methods, however, do not seem suitable when sentence reordering is involved. Some methods appear to perform better for specific domains, as shown by the difference in scores for the literary domain, such as with the *book* and *proverbs* splits where text is less structured and translations may not be literal. Conversely, all methods return perfect alignments on the highly structured text of the *udhr*.

Even without these results, one may argue that simpler methods, which do not require a huge amount of resources, both in term of computation and data, and are mostly language-independent, are better suited to the low-resource setting. Bleualign assumes the use of machine translated data, and thus a MT system, which has to be trained to satisfactory quality. This is usually not possible in a low- and no resource settings. Vecalign requires multilingual sentence embeddings, in our case LASER, which in turn need language specific encoders and a sentencepiece model. In turn, these components need further data than simply the documents to be aligned.

Limitations and Future Work

One obvious limitation of the present work is given by its testing dataset, which includes just one corpus and one low-resource language pair. Future work may expand the study to further language pairs, leveraging other benchmark parallel corpora such as FLORES (Goyal et al., 2022), which would allow to explore other variables, e.g. the effect of typological differences.

Another limitation is that the experiments and their evaluation is currently confined to 1-to-1 alignments. Moving to more complex combinations would require costly manual intervention. However, a qualitative analysis of peculiar cases could be undertaken.

Acknowledgements

We thank the reviewer for their useful inputs. The work of the author is supported by the Internal Grant Agency of Masaryk University, Lexical Computing, and the Ministry of Education of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2018101.

References

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

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.

Everlyn Chimoto and Bruce Bassett. 2022. Very low resource sentence alignment: Luhya and Swahili. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 1–8, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The bible in 100 languages. *Lang. Resour. Eval.*, 49(2):375–395.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1479–1491, Online. Association for Computational Linguistics.

Aloka Fernando, Surangika Ranathunga, Dilan Sachintha, Lakmali Piyarathna, and Charith Rajitha. 2022. Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. *Knowledge and Information Systems*, 65:1–42.

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

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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. Association for Computational Linguistics.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China. Association for Computational Linguistics.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*,

pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

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. Northern European Association for Language Technology (NEALT).

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2020. Effectively aligning and filtering parallel corpora under sparse data conditions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 182–190, Online. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Ha Nguyen Tien, Dat Nguyen Huu, Huong Le Thanh, Vinh Nguyen Van, and Minh Nguyen Quang. 2021. KC4Align: Improving sentence alignment method for low-resource language pairs. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 354–363, Shanghai, China. Association for Computational Lingustics.

Daniel Varga, Péter Halácsy, Andras Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. *Parallel corpora for medium density languages*, pages 247–258.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.