Exploiting Sentence Order in Document Alignment

Brian Thompson
Johns Hopkins University
brian.thompson@jhu.edu

Philipp Koehn
Johns Hopkins University
phi@jhu.edu

Abstract

We present a simple document alignment method that incorporates sentence order information in both candidate generation and candidate re-scoring. Our method results in 61% relative reduction in error compared to the best previously published result on the WMT16 document alignment shared task. Our method improves downstream MT performance on web-scraped Sinhala–English documents from ParaCrawl, outperforming the document alignment method used in the most recent ParaCrawl release. It also outperforms a comparable corpora method which uses the same multilingual embeddings, demonstrating that exploiting sentence order is beneficial even if the end goal is sentence-level bitext.

1 Introduction

Document alignment is the task of finding parallel document pairs (i.e., documents that are translations of each other) in a large collection of documents, often crawled from the web. Aligned documents have historically been used to produce sentence-level machine translation (MT) data, but there is growing evidence that MT systems should be trained and evaluated using document-level context (Gong et al., 2011; Läubli et al., 2018; Voita et al., 2019; Junczys-Dowmunt, 2019).

We exploit the simple idea that two parallel documents should each contain approximately the same information, *in approximately the same order*. This idea can be traced back at least to the late 1990s, when STRAND (Resnik, 1998) measured how well linearized HTML tags from two documents could be aligned in order to judge whether two web pages were likely parallel. However, more recent work has primarily used unordered representations for documents, including bags of words or n-gram features and averages of sentence embeddings.

Our method consists of two main parts: First, we propose a simple method for candidate generation which embeds documents into a joint semantic embedding space (Berry and Young, 1995; Germann, 2016), in a way that preserves some order information in each document. This enables candidate generation via fast approximate nearest neighbor search. Second, we propose re-scoring those candidate pairs by performing sentence alignment and then scoring that alignment based on (1) the semantic similarity of the resulting aligned sentence pairs; (2) whether the sentence pairs are in the correct languages; and (3) the number of inserted/deleted sentences. Our re-scoring approach seeks to filter out documents pairs that contain similar information, but where the order of that information is not consistent between the two documents.

Our method results in a 61% relative reduction in the false positive rate on the WMT16 document alignment shared task versus the best previously reported method. Applied to web-scraped Sinhala–English data from ParaCrawl (Ban et al., 2020), it improves MT performance by 1.2 BLEU over the document alignment method used in the latest ParaCrawl release (Buck and Koehn, 2016b), when both are used with the Vecalign sentence alignment toolkit (Thompson and Koehn, 2019).

2 Method

We follow a 2-stage approach to consider the $D_S \times D_T$ possible alignments between D_S source documents and D_T target documents:[1]

1. **Candidate Generation**: We first find a fixed number, K, of target documents as potential matches for each source document.
2. **Candidate Re-scoring**: We re-score the $D_S \times K$ document pairs from part 1 using a more accurate but slower scoring method.

[1] We define the source/target such that $D_S > D_T$.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5997–6007,
November 16–20, 2020.

Both our candidate generation method and candidate re-scoring method explicitly account for the content of a document as well as the order of that content within the document.

2.1 Candidate Generation

We propose concatenating several sub-vectors—each emphasizing a different section of the document—to form a multilingual document vector. Each sub-vector is the sum of the sentence embeddings for the entire document, after embeddings are weighted to emphasize a given region of the document and to de-emphasize boilerplate text (e.g., from navigational buttons, pull-down menus, or headers).

Let S_n for $n \in \{0, ..., N-1\}$ be the N sentences in a given document. We compute sub-vectors V_j to emphasize uniformly spaced positions $j \in \{0, ..., J-1\}$ in the document:

$$V_j = \sum_{n=0}^{N-1} \text{emb}(S_n)\ H_j(n)\ B(S_n) \qquad (1)$$

where $\text{emb}(S_n)$ is the multilingual embedding of sentence S_n (see §2.1.1), $H_j(n)$ is a windowing function to emphasise the j^{th} region the document (see §2.1.2), and $B(S_n)$ down-weights boilerplate sentences (see §2.1.3).

The final document vector V is a concatenation of normalized position-weighted sub-vectors V_j. Candidate document pairs are found by searching for pairs using cosine distance and approximate nearest neighbor search. We compare all documents from a given webdomain.[2]

2.1.1 Sentence Embeddings

Function $\text{emb}(S_n)$ maps sentence S_n into a multilingual vector space. In this work we use LASER embeddings (Artetxe and Schwenk, 2019b), as the authors provide a pretrained model that works in 93 languages.[3] LASER embeddings require a significant amount of storage space, so for all experiments in this work so we project them from their native size of 1024 down to 128 dimensions using Principal Component Analysis (PCA), as we find this results in a good performance/space trade-off (see Appendix A).

2.1.2 Windowing Function

$H_j(n)$ is a windowing function to emphasize the j^{th} region of a document. If we were to use a simple rectangular window, then our method would be equivalent to splitting the document into sections and computing the average sentence embedding for each section. However, we instead use many smoothed overlapping windows in an effort to encode more fine-grained position information into the final vector document vector, while also making the document alignment process more robust to offsets between parallel sentences, such as in a document pair with a boilerplate header or advertisement present in one document but not the other.

For our windowing function $H_j(n)$ we select a modified PERT distribution (Vose, 2000) with support over $[0, J]$ and mode $\left(\frac{j+0.5}{J}\right)N$. Modified PERT is based on the PERT (Malcolm et al., 1959; Clark, 1962) distribution, but adds a parameter γ to control peakedness of the distribution. PERT is a re-parameterization of the Beta distribution that is defined by the minimum, most likely and maximum values a variable can take.

We select J=16 and γ=20 to produce windows that look reasonable to the authors (see Appendix B). We do not sweep J or γ, as we are concerned about overfitting given our small development set (see Table 1).

2.1.3 Boilerplate Down-weighting

Many 'sentences' in web-crawled data are not true sentences, but boilerplate text such as text of navigational buttons, headers, or pull-down menus. We explore three methods for down-weighting such boilerplate text:

1. Scaling by the inverse of the log of number of the documents containing a given sentence, inspired by IDF (Sparck Jones, 1988; Buck and Koehn, 2016b)
2. A more aggressive variant of IDF which scales sentences by the inverse of the (linear, as opposed to log) number of documents containing a given sentence, which we denote 'LIDF'
3. Scaling each sentence by its length, in characters, as boilerplate lines are often very short (Kohlschütter et al., 2010).

We find that all three boilerplate methods improve candidate generation performance, but select LIDF as it resulted in the best recall performance on our development set in preliminary experiments.

2.2 Candidate Re-scoring

To re-score a document pair proposed by candidate generation, we perform sentence alignment and

[2] A webdomain is a top-level website (e.g., acted.org).
[3] github.com/facebookresearch/LASER

score the quality of the resulting sentence alignment in order to judge whether the proposed document pair appears to be a good translation pair. Our goal is to filter out documents pairs that may contain similar information, but where the order of that information is not consistent between the two documents, indicating they are not parallel.

Our proposed document pair scoring function is:

$$S(E,F) = \frac{1}{|a(E,F)|} \sum_{e,f \in a(E,F)} \text{sim}(e,f) p(L_E|e) p(L_F|f) \quad (2)$$

where $a(E,F)$ is the sentence alignment (see §2.2.1) of documents E and F; $\text{sim}(e,f)$ is the cosine similarity between sentences e and f; and $p(L_e|e)$, $p(L_f|f)$ are the probabilities that sentences e, f are in the correct languages L_E, L_F (see §2.2.2). To penalize unaligned sentences, $a(E,F)$ includes insertions/deletions but we define $\text{sim}(e,f)$ to be zero in such cases.

2.2.1 Sentence Alignment

To perform sentence alignment, we use Vecalign (Thompson and Koehn, 2019).[4] Vecalign uses multilingual sentence embeddings to judge sentence similarity, in conjunction with a dynamic programming approximation based on fast dynamic time warping (Salvador and Chan, 2007) to approximate a search over the full space of possible sentence alignments in linear time complexity with respect to document length. We follow Thompson and Koehn (2019) and again use LASER embeddings, except we project all embeddings down to size 128.

2.2.2 Language ID

One artifact of using multilingual sentence embeddings is that they give perfect alignment scores to exact, un-translated sentence copies. Since automatic language identification (LID) of web data is often erroneous and not well defined,[5] this can result in un-translated, (near) duplicate documents being found as document pairs. We propose to use all sentences (regardless of language) in sentence alignment, as we hypothesize that copies provide a strong signal for sentence alignment. However, when scoring the alignment we introduce sentence-level LID probabilities to penalize sentence pairs that are not in the correct languages.

[4] github.com/thompsonb/vecalign

[5] We observe numerous mixed-language documents (e.g., main body in one language and the boilerplate in another).

	WMT16 train	WMT16 test	ParaCrawl test
English Docs.	349k	682k	9.68M
French Docs.	225k	522k	-
Sinhala Docs.	-	-	1.49M
Webdomains	49	203	1721
Gold Pairs	1624	2402	0

Table 1: Counts for WMT16 and ParaCrawl data used in this work.

3 Experiments and Results

We evaluate our document alignment method in both high- and low-resource settings. Note that our method is not trained on any parallel documents, and is designed to be as language agnostic as possible. However, it relies on LASER embeddings, which are trained on bitext. Thus we expect performance to be at least partially a function of the quantity of data that LASER is trained on.[6] For high-resource, we use the publicly available French–English data released for the WMT 2016 shared task on document alignment (Buck and Koehn, 2016a) and evaluate document recall following the shared task. The shared task provides a strong set of baselines, as 13 different teams contributed at least one submission. For low-resource, we experiment with Sinhala–English documents extracted from ParaCrawl. In this setting we do not have gold document alignments, so we instead evaluate the quality of MT systems trained on the data extracted via document alignment.

We develop and set all parameters using the training data from WMT16 ('WMT16-train') and then test on the WMT16 test data ('WMT16-test') and the Sinhala–English ParaCrawl data. Basic statistics for each dataset are shown in Table 1.

3.1 Candidate Generation

We find that encoding order in document vectors substantially reduces the number of candidates, K, that must be searched to find the correct document: see Figure 1. The improvement is largest when a small number of candidates are considered—the proposed method approximately halves the num-

[6] For the two languages considered here, LASER was trained on much more French–English data (8.8M) than Sinhala–English data (796k) (Artetxe and Schwenk, 2019b). This comparison is likely complicated by data quality (which we generally expect to be higher in higher-resources languages) and benefits of training in related languages.

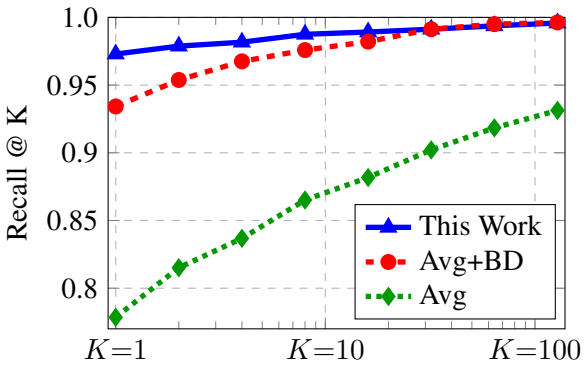

Figure 1: Fraction of the time that a correct document (or near duplicate of it) is found in the top K candidates, as a function of K, found by searching document vectors made from average sentence vectors ('Avg'), average sentence vectors with boilerplate down-weighting ('Avg+BD'), and the proposed method incorporating document order. Results shown on WMT16-test.

ber of false positives between K=1 and K=10 compared to the stronger of the two baselines.

3.2 Document Alignment Recall

Within each webdomain, we embed documents as described in §2.1. For each French document, we find the top K=32 candidate translations via approximate nearest neighbor search using FAISS (Johnson et al., 2017). We then re-score each candidate pair with Equation 2. Language ID probabilities are estimated using fastText (Joulin et al., 2016).[7] We extract the highest scoring document pairs via the greedy search method described in Buck and Koehn (2016b).[8]

We evaluate document pairs following Buck and Koehn (2016a).[9] The proposed method has a recall of 98.5%, compared to the previous best of 96.2% (see Table 2); this corresponds to a 61% relative reduction in false positive rate. We also try our candidate generation method without rescoring (i.e., K=1) and find that it outperforms prior work, but is not as strong as our candidate generation method in conjunction with our candidate re-scoring method. For a description of the contrastive methods, see Buck and Koehn (2016a).

[7] dl.fbaipublicfiles.com/fasttext/supervised-models/lid.176.bin

[8] Buck and Koehn (2016b) found that in practice the greedy search outperformed the theoretically optimal Kuhn–Munkres algorithm (Munkres, 1957).

[9] We use their "soft" recall, which gives credit to document pairs for which the English or French document (but not both) differed from a gold document pair by less than 5%, as measured by text edit distance.

Method	Recall
Azpeitia and Etchegoyhen (2016)	93.1%
Germann (2016)	95.0%
Gomes and Pereira Lopes (2016)	95.9%
Dara and Lin (2016)	96.0%
Buck and Koehn (2016b)	96.2%
This Work: Without Re-Scoring	97.1%
This Work: With Re-Scoring	**98.5%**

Table 2: Document recall on WMT16-test, compared to previous best reported results. The proposed method outperforms prior work, even before re-scoring.

3.3 Impact on Downstream MT

We perform document alignment on Sinhala–English documents web-scraped by ParaCrawl. We apply the same method as in French–English, using the same parameters. We compare to document alignment via Buck and Koehn (2016b), followed by sentence alignment using both Vecalign and Hunalign (Varga et al., 2007), as the latter was used for the most recent ParaCrawl release.

Our document alignment method and Vecalign both use LASER embeddings. The use of LASER embeddings has been proposed for finding parallel sentences in comparable corpora (i.e., without doing document alignment), using a margin-based criterion (Artetxe and Schwenk, 2019a). Since both methods use the same multilingual embeddings (LASER), this allows us to determine whether using document-level information (i.e., performing document alignment and then sentence alignment) provides better data than simply treating the data as comparable corpora and searching for sentence pairs. We refer to this method 'LASER-cc.' For a fair comparison with our document alignment method, we search for sentence pairs within each webdomain.

For each method of finding parallel sentences, evaluation is the same: Since the true amount of parallel data is unknown, we rank the data from highest to lowest quality following Chaudhary et al. (2019) and train systems on a number of different data amounts, as measured by the number of English words. We train NMT systems following the WMT19 sentence filtering shared task (Koehn et al., 2019). Following Thompson and Koehn (2019), we train 5 systems per setting and report both mean and standard deviation BLEU scores. We report BLEU scores using sacreBLEU (Post, 2018).

Method	BLEU
Buck + Hunalign	8.74 +/- 0.20
Buck + Vecalign	10.46 +/- 0.13
LASER-cc	10.40 +/- 0.15
This Work + Vecalign	**11.62 +/- 0.09**

Table 3: Downstream BLEU (+/- standard deviation for 5 runs) for the three document alignment + sentence alignment methods compared in this work, plus the comparable corpora method LASER-cc. 'Buck' denotes Buck and Koehn (2016b). BLEU shown for best filtering threshold for each method; see Appendix C for the results over the entire range of threshold values.

Results at the best threshold for each method are shown in Table 3, and results for the full sweep over all thresholds are provided in Appendix C. The proposed method improves downstream MT performance by 1.2 BLEU over Buck and Koehn (2016b), when both are used in conjunction with Vecalign, and 2.9 BLEU over Buck and Koehn (2016b) with Hunalign (used in the most recent Paracrawl release).

The proposed method also outperforms the LASER-cc baseline by 1.2 BLEU. As LASER-cc and the proposed method use the exact same sentence embeddings, this result shows that incorporating sentence order not only produces documents that can be used for document-level MT training, but also results in higher quality sentence pairs.

4 Related Work

There is a large amount of prior work in document alignment. One of the simplest methods is URL similarity (Resnik, 1998; Chen and Nie, 2000), although this has been shown to be brittle (Tiedemann, 2011). HTML structure (Resnik and Smith, 2003; Shi et al., 2006) or metadata such as publication date (Munteanu and Marcu, 2005) is often similar between parallel websites. However, most more recent work has focused on content similarity via bag-of-words or bag-of-ngrams, using bilingual lexicon (Ma and Liberman, 1999; Fung and Cheung, 2004; Ion et al., 2011; Esplà-Gomis et al., 2016; Etchegoyhen and Azpeitia, 2016; Azpeitia and Etchegoyhen, 2019), machine translation (Uszkoreit et al., 2010), or phrase tables (Gomes and Pereira Lopes, 2016).

Some work has considered high-level order as a filtering step after using a unordered representation to generate candidates: Ma and Liberman (1999) and Le et al. (2016) discard n-gram pairs outside a fixed window, while Uszkoreit et al. (2010) filters out documents that have high edit distance between sequences of corresponding n-gram pairs. Utiyama and Isahara (2003) and Zhang et al. (2006) use sentence similarity and/or number of aligned sentences after performing sentence alignment to score candidate documents. Guo et al. (2018) score document pairs using the sentence-level nearest neighbor as well as the absolute difference in sentence position between sentence pairs. In contrast to these methods, our work considers high-level order in both candidate generation and re-scoring.

Guo et al. (2019) demonstrated neural document embeddings are effective representations for document alignment. They trained on millions of document pairs in each specific language pair of interest; in contrast, this work is much simpler and does not require document-level training data.

5 Conclusion

We present a simple but effective method for document alignment. Our method uses multilingual sentence embeddings and explicitly models the *order* of sentences in documents, in both candidate generation and candidate re-scoring. Our method outperforms all published results on the dataset released for the WMT16 shared task on document alignment. It also increases downstream MT performance in a low-resource setting over prior work, including a margin-based comparable corpora method (Artetxe and Schwenk, 2019a). We use the same embeddings as the comparable corpora method, thus the improvement over that method demonstrates the importance of including sentence order in document alignment, even when document-level alignments are not required.

Acknowledgments

Brian Thompson is supported through the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program.

References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

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.

Andoni Azpeitia and Thierry Etchegoyhen. 2016. DOCAL - vicomtech's participation in the WMT16 shared task on bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 666–671, Berlin, Germany. Association for Computational Linguistics.

Andoni Azpeitia and Thierry Etchegoyhen. 2019. Efficient document alignment across scenarios. *Machine Translation*, pages 1–33.

Marta Ban, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Espl-Gomis, Mikel Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz-Rojas, Leopoldo Pla, Gema Ramrez-Snchez, Elsa Sarras, 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*. Association for Computational Linguistics.

Michael W Berry and Paul G Young. 1995. Using latent semantic indexing for multilanguage information retrieval. *Computers and the Humanities*, 29(6):413–429.

Christian Buck and Philipp Koehn. 2016a. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Christian Buck and Philipp Koehn. 2016b. Quick and reliable document alignment via TF/IDF-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 672–678, Berlin, Germany. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmn, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 263–268, Florence, Italy. Association for Computational Linguistics.

Jiang Chen and Jian-Yun Nie. 2000. Parallel web text mining for cross-language ir. In *Content-Based Multimedia Information Access-Volume 1*, pages 62–77. Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.

Charles E Clark. 1962. Letter to the editor – the pert model for the distribution of an activity time. *Operations Research*, 10(3):405–406.

Aswarth Abhilash Dara and Yiu-Chang Lin. 2016. YODA system for WMT16 shared task: Bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 679–684, Berlin, Germany. Association for Computational Linguistics.

Miquel Esplà-Gomis, Mikel Forcada, Sergio Ortiz-Rojas, and Jorge Ferrández-Tordera. 2016. Bitextor's participation in WMT'16: shared task on document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 685–691, Berlin, Germany. Association for Computational Linguistics.

Thierry Etchegoyhen and Andoni Azpeitia. 2016. A portable method for parallel and comparable document alignment. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 243–255.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and e. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 57–63, Barcelona, Spain. Association for Computational Linguistics.

Ulrich Germann. 2016. Bilingual document alignment with latent semantic indexing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 692–696, Berlin, Germany. Association for Computational Linguistics.

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 697–702, Berlin, Germany. Association for Computational Linguistics.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Mandy Guo, Yinfei Yang, Keith Stevens, Daniel Cer, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Hierarchical document encoder for parallel corpus mining. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 64–72, Florence, Italy. Association for Computational Linguistics.

Radu Ion, Alexandru Ceauşu, and Elena Irimia. 2011. An expectation maximization algorithm for textual unit alignment. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 128–135, Portland, Oregon. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmn, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 56–74, Florence, Italy. Association for Computational Linguistics.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 441–450, New York, NY, USA. ACM.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Thanh C. Le, Hoa Trong Vu, Jonathan Oberländer, and Ondřej Bojar. 2016. Using term position similarity and language modeling for bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 710–716, Berlin, Germany. Association for Computational Linguistics.

Xiaoyi Ma and Mark Y. Liberman. 1999. Bits: A method for bilingual text search over the web. In *In Proceedings of the Machine Translation Summit VII*.

Donald G Malcolm, John H Roseboom, Charles E Clark, and Willard Fazar. 1959. Application of a technique for research and development program evaluation. *Operations research*, 7(5):646–669.

James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

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, Brussels, Belgium. Association for Computational Linguistics.

Philip Resnik. 1998. Parallel strands: A preliminary investigation into mining the web for bilingual text. In *AMTA*.

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

Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A dom tree alignment model for mining parallel data from the web.

Karen Sparck Jones. 1988. Document retrieval systems. chapter A Statistical Interpretation of Term Specificity and Its Application in Retrieval, pages 132–142. Taylor Graham Publishing, London, UK, UK.

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.

Jörg Tiedemann. 2011. Bitext alignment. *Synthesis Lectures on Human Language Technologies*, 4(2):1–165.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109, Beijing, China. Coling 2010 Organizing Committee.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational*

Linguistics, pages 72–79, Sapporo, Japan. Association for Computational Linguistics.

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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef. 2010. Challenges in building a multilingual alpine heritage corpus. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

D Vose. 2000. *Risk analysis: a quantitative guide*. John Wiley & Sons.

Ying Zhang, Ke Wu, Jianfeng Gao, and Phil Vines. 2006. Automatic acquisition of Chinese-English parallel corpus from the web. In *ECIR*.

A Vecalign Speed/Space/Accuracy Trade-off

We experiment with projecting the 1028-dimension LASER embeddings into a lower dimensional space using PCA prior to use in Vecalign. We evaluate sentence alignment accuracy following Thompson and Koehn (2019), on the German–French test set released with Bleualign (Sennrich and Volk, 2010), consisting of manually aligned yearbook articles published in both German and French by the Swiss Alpine Club from the Text+Berg corpus (Volk et al., 2010). Accuracy and alignment time for a range of embedding sizes are shown in Figure 2. Timing is measured on a laptop with a 1.80GHz i7-8550 CPU. We see strong performance ($F_1 > 0.85$) for embeddings down to size 32, in conjunction with up to a 70% reduction in runtime and 97% reduction in disk space required to store the embeddings. However, we select a slightly larger dimension of 128 for use in this work. This projection has minimal impact on sentence alignment accuracy, which we expect to have a direct impact on candidate re-scoring performance. We do not explore the relationship between projected size and candidate generation performance in this work.

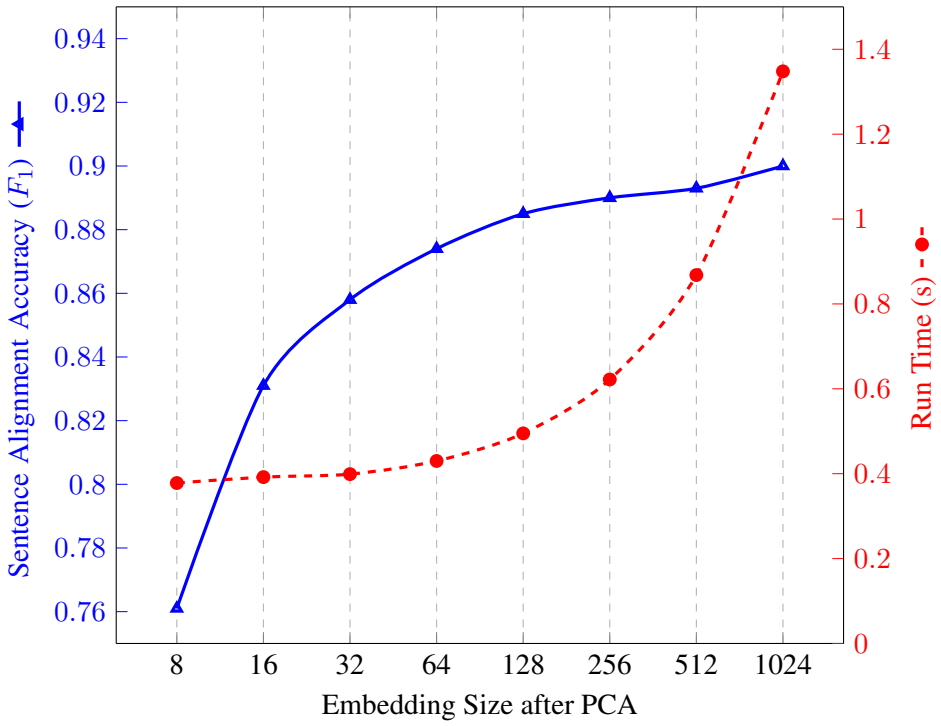

Figure 2: F_1 (solid blue line) vs time to align (dashed red line) the German–French test set after projecting LASER embeddings to various dimensions using PCA.

B Modified PERT Window Illustration

Figure 3 shows the 16 modified PERT windows used in this work, for an example document. We select J=16 and γ=20 to produce windows that look reasonable to the authors, but do not explore sweeping either parameter due to concerns about overfitting on the development set.

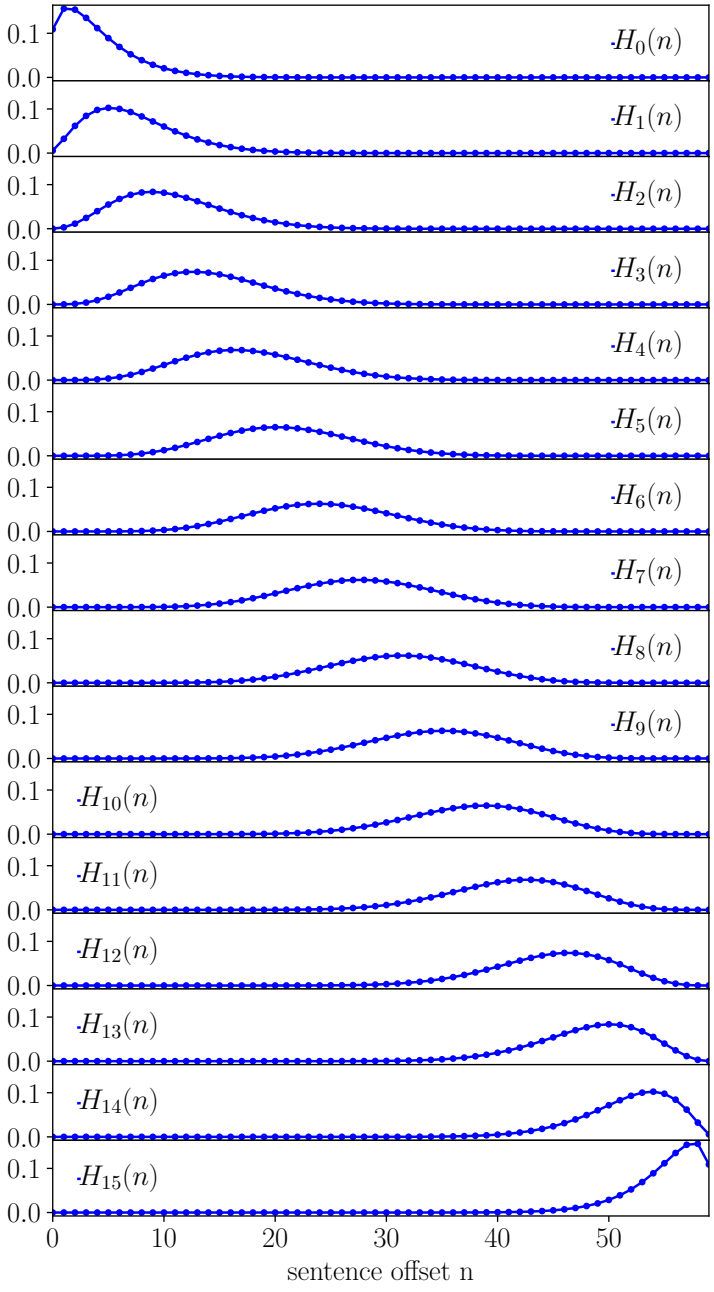

Figure 3: The 16 modified PERT windows used in this work, for an example document containing 60 sentences. Each window emphasizes a different region of the document, but the regions have substantial overlap in an effort to make the final document vector robust to alignment noise, such as offsets caused by a boilerplate header or advertisement present in one document but not the other.

C Downstream MT Performance for All Thresholds

Since the underlying amount of aligned Sinhala–English documents from ParaCrawl is unknown, in order to evaluate downstream MT performance we rank the sentence pairs produced by each method from highest to lowest quality following (Chaudhary et al., 2019) and train each system on many different thresholds. The thresholds for each method are selected to produce different amounts of data, which we measure in English words. Results are shown in Figure 4.

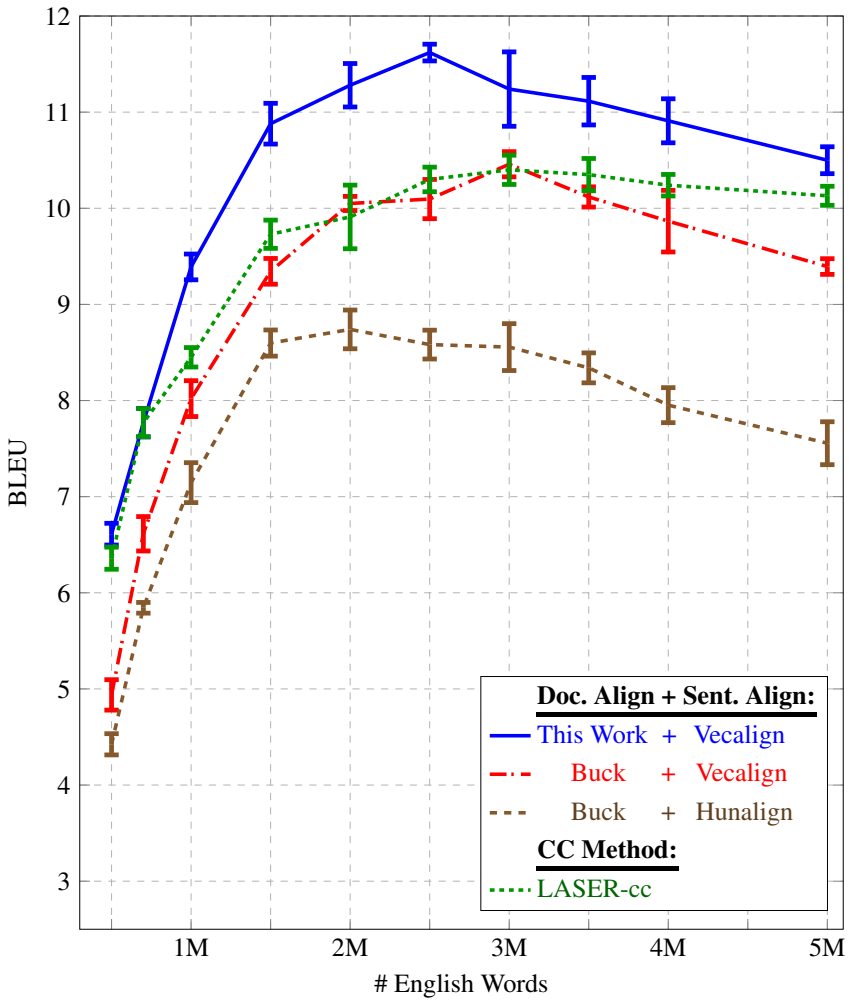

Figure 4: BLEU scores (mean +/- standard deviation for 5 training runs) for systems trained on parallel sentences extracted via several methods, over a range of different filtering thresholds. 'Buck' denotes Buck and Koehn (2016b). LASER-cc denotes the comparable corpora method of Artetxe and Schwenk (2019a).