Parallel Sentence Mining by Constrained Decoding

Pinzhen Chen* **Nikolay Bogoychev*** **Kenneth Heafield** **Faheem Kirefu**
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
{pinzhen.chen, n.bogoych}@ed.ac.uk, {kheafiel, fkirefu}@inf.ed.ac.uk

Abstract

We present a novel method to extract parallel sentences from two monolingual corpora, using neural machine translation. Our method relies on translating sentences in one corpus, but constraining the decoding by a prefix tree built on the other corpus. We argue that a neural machine translation system by itself can be a sentence similarity scorer and it efficiently approximates pairwise comparison with a modified beam search. When benchmarked on the BUCC shared task, our method achieves results comparable to other submissions.

1 Introduction

Having large and high-quality parallel corpora is critical for neural machine translation (NMT). One way to create such a resource is to mine the web (Resnik and Smith, 2003). Once texts are crawled from the web, they form large collections of data in different languages. To find parallel sentences, a natural way is to score sentence similarity between all possible sentence pairs and extract the top-scoring ones. This poses two major challenges:

1. Accurately determining the semantic similarity of a sentence pair in two languages.
2. Efficiently scoring sentence similarity for all possible pairs across two languages.

Scoring each source sentence against each target sentence results in unaffordable quadratic time complexity. A typical workflow reduces the search complexity in a coarse-to-fine manner by aligning documents then aligning sentences within documents (Uszkoreit et al., 2010). However, translated websites may not have matching document structures.

More recent methods focus on direct sentence alignment. The results from Building and Using Comparable Corpora (BUCC) shared task show that direct sentence alignment can be done by sentence-level lexical comparison, neural comparison or a combination of the two (Zweigenbaum et al., 2017, 2018). A state-of-the-art method maps all sentences to multilingual sentence embeddings and compares them using vector similarity (Artetxe and Schwenk, 2019). Such sentence embeddings are produced by neural encoders, but the rise of the attention mechanism demonstrates that sentence embeddings alone are insufficient to obtain full translation quality (Bahdanau et al., 2015).

To exploit quality gains from the attention mechanism, we propose to use a full NMT system with attention to score potentially parallel sentences. The way we avoid pairwise scoring is inspired by constrained decoding in NMT, where the choice of output tokens is constrained to a predefined list (Hokamp and Liu, 2017). Our method works as follows: We designate one language as source and one language as target, and build a trie over all target sentences. Then we translate each source sentence to the target language, but constrain left-to-right beam search to follow the trie. In other words, every translation hypothesis is a prefix of some sentence in the target language. Rather than freely choosing which token to extend by, a hypothesis is limited to extensions that exist in the target language corpus. In effect, we are using beam search to limit target language candidates for each source sentence.

Our work makes two contributions to parallel sentence mining. First, instead of comparing translated text or neural similarity, we use an NMT model to directly score and retrieve sentences on-the-fly during decoding. Second, we approximate pairwise comparison with beam search, so only the top-scoring hypotheses need to be considered at each decoding step.

*Equal contribution.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1672–1678

2 Methodology

NMT systems can assign a conditional translation probability to an arbitrary sentence pair. Filtering based on this (Junczys-Dowmunt, 2018) won the WMT 2018 shared task on parallel corpus filtering (Koehn et al., 2018). Intuitively, we could score every pair of source and target sentences using a translation system in quadratic time, then return pairs that score highly for further filtering. We approximate this with beam search.

2.1 Trie-constrained decoding

We build a prefix tree (trie) containing all sentences in the target language corpus (Figure 1). Then we translate each sentence in the source language corpus using the trie as a constraint on output in the target language. NMT naturally generates translations one token at a time from left to right, so it can follow the trie of target language sentences as it translates.

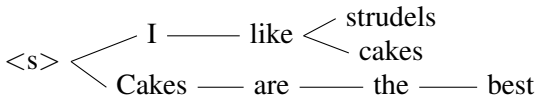


Figure 1: A monolingual trie storing three sentences.

Formally, translation typically uses beam search to approximately maximise the probability of a target language sentence given a source language sentence. We modify beam search to restrict partial translations to be a prefix of at least one sentence in the target language. The trie is merely an efficient data structure with which to evaluate this prefix constraint; partial translations are augmented to remember their position in the trie. We consider two places to apply our constraint.

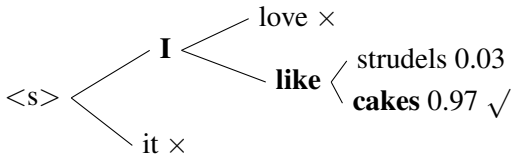
In post-expansion pruning, beam search creates hypotheses for the next word, prunes hypotheses to fit in the beam size, and then requires they be prefixes of a target language sentences. In practice, most sentences are do not have translations in the corpus and search terminates early if all hypotheses are pruned.

In pre-expansion pruning, a hypothesis in the beam generates a probability distribution over all tokens, but only the tokens corresponding to children of the trie node can be expanded by the hypothesis. The search process is guaranteed to find at least one target sentence for each source sentence. Downstream filtering removes false positives.

Algorithm 1 Trie-constrained beam search with maximum output length L, beam size B, vocabulary V and a pre-built trie $trie$

$beam_0 \leftarrow \{\text{<s>}\}$
$match \leftarrow \{\}$
for time step t in 1 to L **do**
  $beam_t \leftarrow \{\}$
  for hypothesis h in $beam_{t-1}$ **do**
    $V_t \leftarrow V$
    if pre-expansion **then** + v2
      $V_t \leftarrow V_t \cap \text{Children}(trie, h)$ + v2
    $beam_t \leftarrow beam_t \cup \text{Continue}(h, V_t, B)$
  $beam_t \leftarrow \text{NBest}(beam_t, B - |match|)$
  if post-expansion **then** + v1
    $beam_t \leftarrow beam_t \cap trie$ + v1
  Move full sentences from $beam_t$ to $match$.
  if $beam_t$ is empty **then**
    return $match$
return $match$

Algorithm 1 presents both variants of our modified beam search algorithm. Besides canonical beam search, "+ v1" indicates post-expansion pruning while "+ v2" indicates pre-expansion pruning. Figure 2 visualises trie-constrained beam search with post-expansion pruning.



- *Source*: Me gustan los pasteles (I like cakes)
- *Target trie*: as shown in Figure 1

Figure 2: Trie-constrained decoding with post-expansion pruning, using beam size 2. × denotes pruned hypotheses. √ denotes the retrieved sentence. Numbers denote translation probabilities.

The modified beam search algorithm allows us to efficiently approximate the comparison between a source sentence and M target sentences. We let B denote beam size and L denote maximum output length. Given each source sentence, our NMT decoder only expands the top B hypotheses intersecting with the trie, for at most L times, regardless of M. With N source sentences, our proposed method will reduce the comparison complexity from $O(MN)$ to $O(BLN)$, where $BL \ll M$.

2.2 Filtering

Pre-expansion pruning leaves each source sentence with an output, which needs to be filtered out if not parallel. We propose to use two methods. When NMT generates an output, a sentence level cross-entropy score is computed too. One way to perform filtering is to only keep sentences with a better per-word cross-entropy than a certain threshold. Another way is to use Bicleaner, an off-the-shelf tool which scores sentence similarity at sentence pair level (Sánchez-Cartagena et al., 2018). Filtering is optional for post-expansion pruning.

2.3 Trie implementation

The trie used in our NMT decoding should be fast to query and small enough to fit in memory. We use an array of nodes as the basic data structure. Each node contains a key corresponding to a vocabulary item, as well as a pointer to another array containing all possible continuations in the next level. Binary search is used to find the correct continuations to the next level. With byte pair encoding (BPE) (Sennrich et al., 2016), we can always keep the maximum vocabulary size below 65535, which allows us to use 2-byte integers as keys, minimising memory usage.

To integrate the trie into the decoder, we maintain external pointers to possible children nodes in the trie for each active hypothesis. When the hypotheses are expanded at each time step, the pointers are advanced to the next trie depth level. This ensures that cross-referencing the trie has a negligible effect on decoding speed.

3 Experiments

3.1 BUCC shared task

We evaluate our method on the BUCC shared task, which requires participants to extract parallel sentences from large monolingual data of English and other languages (Zweigenbaum et al., 2017, 2018). Monolingual and parallel sentences come from Wikipedia and News Commentary respectively. Data are divided into sample, train and test sets at a ratio of 1:10:10. The gold alignments for the test set are not public. Evaluation metrics adopted are precision, recall and F1 score.

When inspecting the BUCC shared task data, we discovered overlapping parallel sentences in the sample, train and test sets. For example, more than 60% of the German-English gold pairs in the test set appear in the train set too.[1]

3.2 Experiment details

We apply our methods on English (En) paired with German (De), French (Fr) and Russian (Ru) on BUCC sample data initially. We train separate translation models for each language into English. All models are Transformer-Base (Vaswani et al., 2017), trained using Marian (Junczys-Dowmunt et al., 2018) with BPE applied. We use parallel data from WMT news translation task (Bojar et al., 2015), excluding News Commentary to prevent our systems from memorising the gold parallel sentences given the overlap issue.

We choose beam size 90 by performing a grid search on De-En pair and keep it unchanged. Regarding the filtering for pre-expansion pruning, per-word conditional cross-entropy thresholds are tuned separately for each pair, because languages inherently have different (cross-)entropies. For Bicleaner, we stick to its default settings, except that we disable the language model filter. All our models translate into English, but our method is actually language-agnostic. Hence, we train a separate En→De model, which will allow us to compare our method in inverse translation directions.

Table 1 reports the performance of our systems on the sample data. Our method exhibits a much higher precision than recall. We hypothesise that if the systems in inverse directions retrieve different sentence pairs, then taking a union will sacrifice some precision for recall, consequently a higher F1. Thus, we present in the same table the results of taking the union of outputs from En→De and De→En systems, labelled as "(3) ∪ (4)". Likewise, we also take the union of the results from cross-entropy and Bicleaner filtering and report scores in the same table.

It turns out that pre-expansion works better than post-expansion. In order to directly compare with previous work, we tune parameters of its filtering thresholds on train data for De-En pair, and apply the pre-expansion variant on the test data. Our results, evaluated by the BUCC organisers, are reported in Table 2 together with other submissions.

Finally, we conduct an add-on experiment to see how our system would perform with in-domain

[1]The shared task organisers confirmed the issue after we pointed it out. They re-evaluated previous submissions without overlapping parallel sentences. On average, recall drops by 2% with the largest being 4%.

	(1) Fr→En			(2) Ru→En			(3) De→En			(4) En→De			(3) ∪ (4)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
(v1) post-expansion	92	62	74	**99**	61	75	88	61	72	96	59	73	81	75	81
(v2) pre-expansion															
+ cross-entropy (CE)	**97**	72	83	98	**84**	**90**	**96**	73	83	**98**	79	**88**	**96**	87	**91**
+ Bicleaner (BC)	86	77	81		n/a*		93	81	86	91	82	86	86	87	87
+ CE ∪ BC	93	**81**	**86**		n/a		91	**84**	**87**	90	**86**	**88**	91	**91**	**91**

* Bicleaner does not have a published classifier model for Ru-En.

Table 1: Precision, recall and F1 of our methods on BUCC sample set.

data. We fine-tune our De→En and En→De systems on News Commentary, excluding the sentence pairs which appear in BUCC train or test sets. As BUCC submissions are asked not to use News Commentary, this is only used to contrast with our own results on the train set.

	Train	Test
Azpeitia et al. (2018)	84.3	85.5
Wieting et al. (2019)	77.5	n/a*
Artetxe and Schwenk (2019)	91.9	95.6
(v2) pre-expansion + CE ∪ BC	83.0	83.9
+ fine-tuning	85.5	n/a

* Wieting et al. directly evaluated on the public train set.

Table 2: F1 scores of our method and other methods on BUCC De-En train and test sets.

4 Results and Analysis

Experiments on the sample data in Table 1 show that pre-expansion pruning outperforms post-expansion by about 10 F1 points. This can be explained by the fact that the decoder has a better chance to generate the correct target sentence if the available vocabulary is constrained. For both variants, the high precision reflects the effectiveness of using NMT as a sentence similarity scorer. Regarding filtering methods, we notice that Bicleaner achieves a more balanced precision and recall, while filtering by per-word cross-entropy leads to very high precision but lower recall. Generally, the latter does better in terms of F1. Taking a union of the output from the two filtering methods results in a even more balanced precision and recall, without damaging F1. This implies that the two filtering techniques keep different sentence pairs.

Table 2 shows that our method achieves comparable performance to other methods. Moreover, our models are trained using a vanilla Transformer-Base architecture on WMT data. Without data or model wise techniques (e.g. in-domain fine-tuning), they are nowhere close to state-of-the-art NMT systems (Barrault et al., 2019). Contrasting Table 1 and Table 2 reveals a discrepancy between our method's F1 scores on the sample and train sets. We suspect that when there are more possible target sentences, our model will have more choices, leading to a lower performance. The same behaviour is also observed in other BUCC 2018 submissions which report their scores on the sample data (Azpeitia et al., 2018; Leong et al., 2018).

Overall our method does not outperform state-of-the-art which leverages neural embeddings. We identify several weaknesses: beam search can only find local optima, and a genuine parallel sentence cannot be recovered once it is pruned. Thus the method is vulnerable when parallel sentences have different word ordering. For example, "Por el momento, estoy bebiendo un café" (English: "At the moment, I am drinking a coffee") can hardly match "I am drinking a coffee at the moment", because an NMT system will have very low probability of generating a reordered translation, unless using an undesirably large beam size. Moreover, compared to methods that consider textual overlap, NMT is sensitive to domain mismatch and rare words (Koehn and Knowles, 2017). When a system is confused by rare words in the source, we observe that the overly zealous language model in the decoder generates a fluent sentence in the trie rather than a translation. This problem is alleviated when our systems are fine-tuned on in-domain data, as shown in Table 2 that there is a gain in F1.

Finally we discuss the limitation of evaluating our method on the BUCC task. First, our method based on NMT can be liable to favour

machine-translated texts, whereas the BUCC data is unlikely to contain those. Next, we notice that some parallel sentences in BUCC data are not included in the gold alignments. For instance, in De-En train set, "de-000081259" and "de-000081260" are the same German sentence, and so are "en-000036940" and "en-000036941" on the English side. Gold alignments only include (de-000081259, en-000036940) and (de-000081260, en-000036941), but not the other two. Lastly, it still remains unknown if a system optimised for F1 will produce the sentences that can truly improve NMT performance.

5 Related Work

A typical parallel corpus mining workflow first aligns parallel documents to limit the search space for sentence alignment. Early methods rely on webpage structure (Resnik and Smith, 2003; Shi et al., 2006). Later, Uszkoreit et al. (2010) translate all documents into a single language, and shortlist candidate document pairs based on TF-IDF-weighted n-grams. Recently, Guo et al. (2019) suggest a neural method to compare document embeddings obtained from sentence embeddings .

With the assumption that matched documents are parallel (no cross-alignment), sentence alignment can be done by comparing sentence length in words (Brown et al., 1991) or characters (Gale and Church, 1993), which is then improved by adding lexical features (Varga et al., 2005). After translating texts into the same language, BLEU can also be used to determine parallel texts, by anchoring the most reliable alignments first (Sennrich and Volk, 2011). Most recently, Thompson and Koehn (2019) propose to compare bilingual sentence embeddings with dynamic programming in linear runtime.

There are also research efforts on parallel sentence extraction without the reliance on document alignment. Munteanu and Marcu (2002) acquire parallel phrases from comparable corpora using bilingual tries and seed dictionaries. Azpeitia et al. (2018) computes Jaccard similarity of lexical translation overlap. Leong et al. (2018) use an autoencoder and a maximum entropy classifier. Bouamor and Sajjad (2018) consider cosine similarity between averaged multilingual word embeddings. Guo et al. (2018) design a dual encoder model to learn multilingual sentence embeddings directly with added negative examples. Wieting et al. (2019) obtain sentence embeddings from sub-word embeddings and train a simpler model to distinguish positive and negative examples. Artetxe and Schwenk (2019) refine Guo et al. (2018)'s work and achieve state-of-the-art by looking at the margins of cosine similarities between pairs of nearest neighbours.

In our work, using NMT as a similarity scorer relies on constrained decoding (Hokamp and Liu, 2017), which has been applied on image captioning (Anderson et al., 2017) and keyword generation (Lian et al., 2019).

6 Conclusion and Future Work

We bring a new insight into using NMT as a similarity scorer for sentences in different languages. By constraining on a target side trie during decoding, beam search can approximate pairwise comparison between source and target sentences. Thus, overall we present an interesting way of finding parallel sentences through trie-constrained decoding. Our method achieves a comparable F1 score to existing systems with a vanilla architecture and data.

Maximising machine translation scores is biased towards finding machine translated text produced by a similar model. More research is needed on this problem given the prevalent usage of NMT. We hypothesise that part of the success of dual conditional cross-entropy filtering (Junczys-Dowmunt, 2018) is checking that scores in both directions are approximately equal, whereas a machine translation would be characterised by a high score in one direction.

Finally, scalability is a key issue in large-scale mining of parallel corpora, where both quantity and quality are of concern. The scalability of direct sentence alignment without a document aligner has not been thoroughly investigated in our work, as well as other related work.

Acknowledgments

This work has received funding from the European Union under grant agreement INEA/CEF/ICT/A2017/1565602 through the Connecting Europe Facility. This paper reflects the authors' views; INEA is not responsible for any use that may be made of the information contained in this paper.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3197–3203. Association for Computational Linguistics.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2018. Extracting parallel sentences from comparable corpora with STACC variants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Houda Bouamor and Hassan Sajjad. 2018. H2@BUCC18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, pages 169–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

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, Belgium, Brussels. 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.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Chongman Leong, Derek F. Wong, and Lidia S. Chao. 2018. UM-pAligner: Neural network-based parallel sentence identification model. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Yijiang Lian, Zhijie Chen, Jinlong Hu, Kefeng Zhang, Chunwei Yan, Muchenxuan Tong, Wenying Han, Hanju Guan, Ying Li, Ying Cao, Yang Yu, Zhigang Li, Xiaochun Liu, and Yue Wang. 2019. An end-to-end generative retrieval method for sponsored search engine–decoding efficiently into a closed target domain. *arXiv preprint arXiv:1902.00592*.

Dragos Stefan Munteanu and Daniel Marcu. 2002. Processing comparable corpora with bilingual suffix trees. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 289–295. Association for Computational Linguistics.

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

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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).

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A DOM tree alignment model for mining parallel data from the web. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 489–496, Stroudsburg, PA, USA. 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.

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.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. *Proceedings of the RANLP 2005 Conference*, pages 590–596.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4602–4608, Florence, Italy. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).