# An Alternative to Thresholding for Margin-Based Bitext Mining

**Alex Jones**
Dartmouth College
alexander.g.jones.23@dartmouth.edu

**Derry Tanti Wijaya**
Boston University
wijaya@bu.edu

## Abstract

Obtaining high-quality parallel corpora is of paramount importance for training NMT systems. However, as many language pairs lack adequate gold-standard training data, a popular approach has been to mine so-called "pseudo-parallel" sentences from paired documents in two languages. In this paper, we outline some problems with current methods, propose computationally economical solutions to those problems, and demonstrate success with novel methods on the Tatoeba similarity search benchmark and on a downstream task, namely NMT. We uncover the effect of resource-related factors (i.e. how much monolingual/bilingual data is available for a given language) on the optimal choice of bitext mining approach, and echo problems with the oft-used BUCC dataset that have been observed by others. We make the code and data used for our experiments publicly available.[1]

## 1 Introduction

Mining so-called "pseudo-parallel" sentences from sets of similar documents in different languages ("comparable corpora") has gained popularity in recent years as a means of overcoming the dearth of parallel training data for many language pairs. With increasingly powerful computational resources and highly efficient tools such as FAISS (Johnson et al., 2017) at our disposal, the possibility of mining billions of pseudo-parallel bitexts for thousands of language pairs to the end of training a multilingual NMT system has been realized. In particular, Fan et al. (2020) perform global mining over billions of sentences in 100 languages, resulting in a massively multilingual NMT system containing supervised data for 2200 language pairs.

Despite these astounding breakthroughs in high-resource engineering, many questions remain to be answered about bitext mining from a research perspective, ones with particular relevance to the *low*-resource engineering case, i.e. contexts with limited computational resources. While Fan et al. (2020) yield impressive results using hundreds of GPUs, aggressive computational optimization, and a global bitext mining procedure (i.e. searching the entire target corpus for a source sentence match), how these results transfer to the low-computational-resource case is not clear. Moreover, the effect of circumstantial (e.g. the resources available for a given language or language pair) or linguistic (e.g. typological) factors on bitext mining performance remains highly understudied. In fact, we argue that efforts to scale this task have outpaced efforts to rigorously document and understand the factors which determine its outcome.

In this paper, we highlight the problematic nature of using similarity score thresholding (Artetxe and Schwenk, 2019b; Schwenk et al., 2019a,b; Fan et al., 2020) for mining both gold-standard and pseudo-parallel sentences, in the latter case focusing on document-level mining from the Wikipedia corpora in medium-low-resource languages (namely English-Kazakh and English-Gujarati). We propose a heuristic method involving pre-translation of source and/or target sentences before mining, and show that particular variations of this approach outperform various similarity thresholds for mining pseudo-parallel sentences, as well as gold-standard bitexts in certain cases. On the gold-standard mining task, we establish what are, to our knowledge, benchmarks on dozens of languages, and perform a comprehensive breakdown of results by language resource capacity, showing the optimal mining method to be partially dependent on this resource factor.

[1] https://github.com/AlexJonesNLP/alt-bitexts

## 2 Related Work

Mining pseudo-parallel sentences from paired corpora for the purposes of training NMT systems is a decades-old problem, and dozens of solutions have been tried, ranging from statistical or heuristic-based approaches (Zhao and Vogel, 2002; Resnik and Smith, 2003; Munteanu et al., 2004; Fung and Cheung, 2004; Munteanu and Marcu, 2006) to similarity-based, rule-based, and hybrid approaches (Azpeitia et al., 2017, 2018; Bouamor and Sajjad, 2018; Hangya et al., 2018; Schwenk, 2018; Ramesh and Sankaranarayanan, 2018; Artetxe and Schwenk, 2019a,b; Hangya and Fraser, 2019; Schwenk et al., 2019a,b; Wu et al., 2019; Keung et al., 2020; Tran et al., 2020; Kvapilíková et al., 2020; Feng et al., 2020; Fan et al., 2020). Benchmarks to measure performance on this task include the BUCC[2] '17/18 datasets (Zweigenbaum et al., 2017, 2018), whose task involves spotting gold-standard bitexts within comparable corpora, and the Tatoeba dataset (Artetxe and Schwenk, 2019b), whose task involves matching gold-standard pairs in truly parallel corpora.

Relevant to similarity-based mining methods are well-aligned cross-lingual word and sentence embeddings, which are some of the oldest constructs in NLP and have been tackled using hundreds of diverse approaches. Even among relatively recent efforts, these approaches range from static, monolingual embeddings (Pennington et al., 2014; Mikolov et al., 2013; Arora et al., 2017; Kiros et al., 2015) to static, multilingual ones (Klementiev et al., 2012; Ammar et al., 2016; Schwenk and Douze, 2017) to contextualized, monolingual ones (Peters et al., 2018; Subramanian et al., 2018; Devlin et al., 2019; Liu et al., 2020; Conneau et al., 2017; Reimers and Gurevych, 2019) to contextualized, multilingual ones (Song et al., 2019; Conneau and Lample, 2019a; Conneau et al., 2020; Reimers and Gurevych, 2020; Feng et al., 2020; Wang et al., 2020), including efforts at cross-lingual alignment (Xu et al., 2018; Artetxe et al., 2018a; Schuster et al., 2019; Zhang et al., 2019; Cao et al., 2020). In this paper, our approach centers around using contextualized, multilingual sentence embeddings for the task of bitext mining, although we mention attempts at rule-and-similarity-based hybrid methods in Section B in the appendix.

For low resource languages where parallel training data is little to none, unsupervised NMT can play a crucial role (Artetxe et al., 2018b, 2019a,b, 2018c; Hoang et al., 2018; Lample et al., 2017; Lample et al., 2018b,c; Pourdamghani et al., 2019; Wu et al., 2019). However, previous works have only focused on high-resource and/or similar-to-English languages. Most recently, several works have questioned the universal usefulness of unsupervised NMT and showed its poor results for low-resource languages (Kim et al., 2020; Marchisio et al., 2020). They note the importance of linguistic similarity between source and target language, and domain proximity along with size and quality of the monolingual corpora, for good unsupervised NMT performance. They reason that since these conditions can hardly be satisfied in the case of low resource languages, they result in poor unsupervised performance for these languages. However, recently it has been shown that training a language model on monolingual data, followed by training with unsupervised MT objective and then training on mined comparable data (uns, 2021) can improve MT performance for low resource languages. In this work, we explore the usefulness of our mined bitext using a similar pipeline. We show an improvement over using only supervised training data for low resource language MT.

## 3 Model selection

### 3.1 Cross-lingual Sentence Embeddings

We initially experiment with XLM-RoBERTa (Conneau et al., 2020) for our bitext mining task, using averaged token embeddings or the [CLS] (final) token embedding as makeshift sentence embeddings. However, we replicate Reimers and Gurevych (2020)'s results in showing these ad-hoc sentence embeddings to have relatively poor performance on the BUCC '17/18 EN-FR train data (Zweigenbaum et al., 2017, 2018) compared to bona fide sentence embeddings like LASER (Artetxe and Schwenk, 2019b) and LaBSE (Feng et al., 2020). Thus, we opt to use LaBSE as our sentence embedding model, using its implementation in the Sentence Transformers [3] library, as LaBSE performs state-of-the-art (SOTA) or near-SOTA on the BUCC and Tatoeba datasets (Artetxe and Schwenk, 2019b)[4]. Moreover, being more recent than LASER, LaBSE has been investigated less thoroughly in the context of this task.

[2] Building and Using Comparable Corpora

[3] https://www.sbert.net

[4] https://github.com/facebookresearch/LASER/tree/master/data/tatoeba/v1

## 4 Methods

### 4.1 Margin-based Mining

For our primary mining procedure, we use margin-based mining as described in Artetxe and Schwenk (2019a). Seeking to mitigate the hubness problem (Dinu et al., 2014), margin scoring poses an alternative to raw cosine similarity in that it selects the candidate embedding that "stands out" the most from its $k$ nearest neighbors. We use the *ratio* margin score, as described in Artetxe and Schwenk (2019a) and defined below:

(1)

$$\text{score}(x, y) = \frac{\cos(x, y)}{\frac{1}{2k}\left(\sum_{z \in NN_k(x)} \cos(x, z) + \sum_{z \in NN_k(y)} \cos(y, z)\right)}$$

As in Artetxe and Schwenk (2019a), we use $k = 4$ for all our mining procedures. We acknowledge that $k$ is indeed a tuneable and important hyperparameter of KNN search, and that higher values of $k$ may work better for bitext mining in certain scenarios, depending on factors such as the size of the search space (Schwenk et al., 2019b). However, we don't make this hyperparameter a focus of this paper, instead addressing the problem of margin score thresholding and its relation to the size of the search space. We leave a thorough examination of $k$ and its effect on bitext mining performance for future work.

Additionally, Artetxe and Schwenk (2019a) describes four different "retrieval" techniques used to obtain sentence pairs after performing margin scoring, namely *forward*, *backward*, *intersection*, and *max score*. In the *forward* procedure, every sentence in the source corpus is matched with some sentence in the target corpus, with this mapping being possibly non-surjective (i.e. not every sentence in the codomain need be mapped to). The *backward* procedure is defined analogously, and the *intersection* method (INTERSECT in Algorithm 1) takes the intersection of the resulting sentence pairs from these two procedures. Following Artetxe and Schwenk (2019a), we find that *intersection* produces good results, and use it on all mining tasks. *Max score* takes the argmax sentence pair for any inconsistent alignments after bidirectional search (i.e. if sentences $x_j, y_k$ are paired in forward search and $x_l, y_k$ are paired in backward search, then take whichever has a higher associated margin score). Because *max score* yields little or no benefit over *intersection*, as shown in Artetxe and Schwenk (2019a), we decided not to use it. We also try taking the union (denoted UNION in Algorithm 1) of forward and backward searches to prioritize recall, but find that this harms overall F1 on the BUCC '17/18 EN-FR training set due to decreased precision, and abandon the technique in further experiments.

We also perform all mining at the document level for the sake of computational thrift, and because recent approaches have targeted the global-level mining scenario but not verified the generalizability of the techniques used. The Primary mining procedures described above are also outlined in Algorithm 1.

**Algorithm 1:** Doc-level margin-based mining

**Given** $\mathcal{X}$, $\mathcal{Y}$, $k$, $t$ *JOIN_METHOD*
$\mathcal{X}$: Set of sentences in language X. May be grouped into documents or standalone sentences.
$\mathcal{Y}$: Set of parallel or comparable sentences in language Y.
$k$: Number of neighbors
JOIN_METHOD: Method of combining sentence pairs after mining in the forward and backward directions. One of either INTERSECT or UNION.
$t$: Margin similarity threshold

MINE SENTENCE PAIRS IN BOTH DIRECTIONS
**for** *document* $\mathcal{D} \in \mathcal{X}$ **do**
  **for** $x \in \mathcal{D}$ **do**
    $nn_x \leftarrow NN(x, \mathcal{Y}_\mathcal{D}, k)$ ; `// FAISS k-nearest neighbor search`
    $best_y = \text{argmax}_{y \in nn_x}\, score(x, y)$ ; `// Eq.(1)`
    **if** $score(x, best_y) > t$ **then**
      $fwd_D \leftarrow (x, best_y)$
    **end**
    $fwd \leftarrow fwd_D$
  **end**
**end**
**for** $\mathcal{D} \in \mathcal{Y}$ **do**
  **for** $y \in \mathcal{D}$ **do**
    $nn_y \leftarrow NN(y, \mathcal{X}_\mathcal{D}, k)$
    $best_x = \text{argmax}_{x \in nn_y}\, score(y, x)$
    **if** $score(best_x, y) > t$ **then**
      $bwd_D \leftarrow (best_x, y)$
    **end**
    $bwd \leftarrow bwd_D$
  **end**
**end**
**if** *INTERSECT* **then**
  $\mathcal{P} \leftarrow \{fwd\} \cap \{bwd\}$
**end**
**else if** *UNION* **then**
  $\mathcal{P} \leftarrow \{fwd\} \cup \{bwd\}$
**end**
**return** $\mathcal{P}$

**Algorithm 2:** Secondary retrieval procedures

**Given** $\mathcal{X}, \mathcal{Y}, k, \boldsymbol{\mathcal{M}}, JOIN\_METHOD$
$t$: Margin score threshold
$\boldsymbol{\mathcal{M}}$: An NMT model
**if** *TRANSLATE* **then**
  **if** *EN_TO_XX* **then**
    **for** $x \in \mathcal{X}$ **do**
      $\mathcal{X}_{trans} \leftarrow \boldsymbol{\mathcal{M}}(x \rightarrow lang_y)$
      $\mathcal{P}_{en\_xx} \leftarrow \boldsymbol{AlgorithmI}(\mathcal{X}_{trans}, \mathcal{Y}, k, JOIN\_METHOD, t)$
    **end**
    **if not** STRICT_INT *or* PAIRWISE_INT **then**
      **return** $\mathcal{P}_{en\_xx}$
  **end**
  **if** *XX_TO_EN* **then**
    **for** $y \in \mathcal{Y}$ **do**
      $\mathcal{Y}_{trans} \leftarrow \boldsymbol{\mathcal{M}}(y \rightarrow lang_x)$
      $\mathcal{P}_{xx\_en} \leftarrow \boldsymbol{AlgorithmI}(\mathcal{Y}_{trans}, \mathcal{X}, k, JOIN\_METHOD, t)$
    **end**
    **if not** STRICT_INT *or* PAIRWISE_INT **then**
      **return** $\mathcal{P}_{xx\_en}$
  **end**
**end**
$\mathcal{P}_{orig} \leftarrow \boldsymbol{AlgorithmI}(\mathcal{X}, \mathcal{Y}, k, JOIN\_METHOD, t)$
**if** *STRICT_INT* **then**
  **return** $\mathcal{P}_{orig} \cap \mathcal{P}_{en\_xx} \cap \mathcal{P}_{xx\_en}$
**end**
**else if** *PAIRWISE_INT* **then**
  **return** $\mathcal{P}_{orig} \cap \mathcal{P}_{en\_xx} \bigcup \mathcal{P}_{orig} \cap \mathcal{P}_{xx\_en} \bigcup \mathcal{P}_{en\_xx} \cap \mathcal{P}_{xx\_en}$
**end**
**else**
  **return** $\mathcal{P}_{orig}$
**end**

### 4.2 Filtering Procedures

#### 4.2.1 Thresholding

The most straightforward measure for filtering mined sentence pairs is setting a similarity score threshold, as shown in Artetxe and Schwenk (2019a). Of course, there is a precision-recall tradeoff inherent to adjusting this threshold, and we show it is problematic in other ways for our document-level approach on a noisy corpus as well. We argue that choosing this threshold is an expensive and ambiguous process, one which has not been addressed with much rigor or been show to generalize to diverse mining scenarios.

#### 4.2.2 Pre-Translation

Our approach capitalizes on multiple similarity-related signals by first translating either the source texts (i.e. en→xx), target texts (xx→en), or both. In our experiments on the Tatoeba dataset (Artetxe and Schwenk, 2019b), we translate with Google Translate / GNMT (Wu et al., 2016) using Cloud Translation API. We also experimented with using Tiedemann and Thottingal (2020), but observed poor performance (e.g. poor coverage) for multiple language pairs. Due to the cost of using this API on large bodies of text, when mining on the English-Kazakh and English-Gujarati comparable corpora, we train a supervised system on WMT'19 data (Barrault et al., 2019), with training corpora sizes given in Table 1. When translating in either direction, we translate the entire corpus, e.g. translating all English sentences in the Wikipedia corpus to Kazakh.

#### 4.2.3 Strict & Pairwise Intersection

We also experiment with combining sentence pairs after mining using all three procedures described above and in Algorithm 2. We first mine using three approaches:

1. Mine sentence pairs using margin-based scoring (Algorithm 1) with the original en, xx sentences
2. Mine pairs with the original en and translated xx→en sentences
3. Mine pairs with the original xx and translated en→xx sentences

After doing so, we either perform a "strict intersection" (STRICT_INT in Algorithm 2)—keeping only sentence pairs which appear in all three sets of pairs—or "pairwise intersection" (PAIRWISE_INT), a voting approach that keeps any pairs occurring in $\geq 2$ of the sets above.

### 4.3 Supervised and Unsupervised NMT

We follow the same pipeline for training MT in (uns, 2021) that is based on XLM (Conneau and Lample, 2019b). Following their pipeline, we first pretrain a bilingual Language Model (LM) using the Masked Language Model (MLM) objective (Devlin et al., 2019) on the monolingual corpora of two languages (e.g. Kazakh and English for en-kk) obtained from Wikipedia, WMT 2018/2019[5] and Leipzig corpora (2016)[6]. For both the LM pretraining and NMT model fine-tuning, unless otherwise noted, we follow the hyper-parameter settings suggested in the XLM repository[7]. For every language pair we extract a shared 60,000 subword vocabulary using Byte-Pair Encoding (BPE) (Sennrich et al., 2016). After pretraining the LM, we train a NMT model in an unsupervised manner following the setup recommended in Conneau and Lample (2019b), where both encoder and decoder

[5] http://data.statmt.org/news-crawl/
[6] https://wortschatz.uni-leipzig.de/en/download/
[7] http://github.com/facebookresearch/XLM

are initialized using the same pretrained encoder block. For training unsupervised NMT, we use back-translation (*BT*) and denoising auto-encoding (*AE*) losses (Lample et al., 2018a), and the same monolingual data as in LM pretraining. Lastly, to train a supervised MT using our mined comparable data, we follow *BT+AE* with *BT+MT*, where *MT* stands for supervised machine translation objective for which we use the mined data. We stopped the training when the validation perplexity (LM pre-training) or BLEU (translation training) was not improved for ten checkpoints. We run all our experiments on 2 GPUs, each with 12GB memory.

We compare the performance in terms of BLEU score of our MT model with a model that follows the same pipeline (LM pre-training, unsupervised MT training, followed by supervised MT training) but that uses (human translation) training data from WMT19 (Table 1). The size of the monolingual data we use for LM pretraining are also shown in Table 1.

| **Train data** | **Number of sentences** | |
|---|---|---|
| | **en-kk** | **en-gu** |
| **Monolingual** | 9.51M | 1.36M |
| **Supervised** | | |
| WMT'19 | 222,165 | 22,321 |
| **Comparable** | | |
| **1** LaBSE (threshold = 1.06) | 430,762 | 120,989 |
| **2** LaBSE (pairwise intersection, doc-level, all) | 154,679 | 113,955 |
| **3** LaBSE (pairwise intersection, doc-level, all, threshold = 1.20) | 55,765 | — |
| **4** LaBSE (pairwise intersection, doc-level, all, threshold = 1.35) | 19,099 | — |

Table 1: Sizes (in number of sentences) of training corpora used in training supervised and semi-supervised NMT. The comparable/pseudoparallel sentences are mined using margin-based scoring with LaBSE, with secondary retrieval procedures given in parentheses. These procedures are described in Section 4.

## 5 Experiments

### 5.1 Gold-standard Bitext Retrieval

In gold-standard bitext retrieval tasks, the goal is to mine gold-standard bitexts from a set of parallel or comparable corpora. We use the common approach of finding k-nearest neighbors for each sentence pair (in both directions, if using INTERSECT in Algorithm 1), then choosing the sentence that maximizes the ratio margin score (Equation 1 in Section 4.1).

**Tatoeba Dataset**[8] The Tatoeba dataset, introduced by Artetxe and Schwenk (2019b), contains up to 1,000 English-aligned, gold-standard sentence pairs for 112 languages. In light of our focus on lower-resource languages, we experiment only on the languages listed in Table 10 of Reimers and Gurevych (2020), which are languages without parallel data for the distillation process they undertake. This heuristic choice is supported by relative performance against languages *with* parallel data for distillation: the average raw cosine similarity baseline with LaBSE for the latter was 96.3, in contrast with 73.7 for the former. Specifically, the ISO 639-2 codes[9] for the languages we use are as follows: afr, amh, ang, arq, arz, ast, awa, aze, bel, ben, ber, bos, bre, cbk, ceb, cha, cor, csb, cym, dsb, dtp, epo, eus, fao, fry, gla, gle, gsw, hsb, ido, ile, ina, isl, jav, ksb, kaz, khm, kur, kzj, lat, lfn, mal, mhr, nds, nno, nov, oci, orv, pam, pms, swg, swh, tam, tat, tel, tgl tuk, tzl, uig, uzb, war, wuu, xho, yid.

**BUCC Dataset** The BUCC '17/18 dataset (Zweigenbaum et al., 2017, 2018), provided by the Workshop for Building and Using Comparable Corpora, features English-aligned comparable corpora from Wikipedia in French, German, Chinese, and Russian, with gold-standard bitexts from News Commentary inserted randomly throughout. The goal of the task is extract these gold-standard pairs, with performance measured using standard F1-score. This task has been tackled with a variety of approaches (Bouamor and Sajjad, 2018; Etchegoyhen and Azpeitia, 2016; Azpeitia et al., 2018, 2017; Hangya et al., 2018; Artetxe and Schwenk, 2019a,b; Hangya and Fraser, 2019; Keung et al., 2020; Feng et al., 2020; Reimers and Gurevych, 2020).
We use only the publicly available[10] EN-FR train data in our experiments, and initially experiment using rule-based metrics on top of margin-based mining, similar to Keung et al. (2020). However, we note major problems with the BUCC data, which are discussed in Section B, and for this reason—coupled with the lackluster performance of these rule-based metrics—do not report results on

[8] https://github.com/facebookresearch/LASER/tree/master/data/tatoeba/v1

[9] https://www.loc.gov/standards/iso639-2/php/code_list.php

[10] https://comparable.limsi.fr/bucc2017/bucc2017-task.html

this dataset, though the methods we try are described in Section B.

### 5.2 Pseudo-parallel Sentences From Comparable Corpora

In addition to gold-standard bitext mining, we also mine pseudo-parallel sentences from so-called comparable corpora. The aim of this task is as follows: given two sets of similar documents in different languages, find sentence pairs that are close enough to being translations to act as training data for an NMT system. Of course, unlike the gold-standard mining task, there are not ground-truth labels present for this task, and so evaluation must be performed on a downstream task like NMT.

**Comparable Corpora** Our comparable data is mined from comparable documents, which are linked Wikipedia pages in different languages obtained using the langlinks from Wikimedia dumps. For each sentence in a foreign language Wikipedia page, we use all sentences in its corresponding linked English language Wikipedia page as potential comparable sentences.

**Pre-processing** Since our comparable corpora for both EN-KK and EN-GU are grouped into documents, the most important pre-processing step we perform is eliminating especially short documents before similarity search. The motivation for this is that since we search at document-level, the quality of the resulting pairs could be highly degraded in particularly small search spaces, in a way that neither thresholding nor voting could mitigate. Note that average document length was much shorter for both Gujarati and Kazakh than for English, due simply to shorter Wikipedia articles in those languages. For the EN-KK corpus, we omitted any paired documents whose English version was $< 30$ words or whose Kazakh version was $< 8$ words, which we determined somewhat arbitrarily by seeing what values allowed for a sufficient number of remaining sentences. For the EN-GU corpus, we take a more disciplined approach and lop off the bottom 35% of shortest document pairs, which happened to be $document\_length = 21$ sentences for English and 5 sentences for Gujarati. This step accounted for the large number of documents in each corpus that contained very few sentences (see Figure 1 for an example).

We performed additional more-or-less standard pre-processing, such as removing URLs, non-standard characters, and superfluous white space, as well as recurrent noise that we spotted in the corpora (such as "href" in the English part of the EN-KK corpus).

**Document-level mining vs. global mining** Due to the sizes of the comparable corpora and our computational resources, we perform document-level mining (described in Algorithm 1) when retrieving pseudo-parallel sentence pairs for NMT training and global mining (mining over all sentence pairs in each corpus) when experimenting on the Tatoeba and BUCC corpora. Like Schwenk et al. (2019b); Fan et al. (2020), we speculate that global mining yields better results than document-level mining, all else being equal. However, like Schwenk et al. (2019a), we note that this conjecture has yet to be rigorously examined, and that we don't boast the resources to do so meaningfully.

### 5.3 NMT

We conduct experiments on Kazakh and Gujarati. They are spoken by 22M and 55M speakers worldwide, respectively, and are distant from English, in terms of writing scripts and alphabets. Additionally, these languages have few parallel but some comparable and/or monolingual data available, which makes them ideal and important candidates for our low-resource unsupervised NMT research.

Our monolingual data for LM pre-training of these languages (shown in Table 1) are carefully chosen from the same topics (for Wikipedia) and the same domain (for news data). For the news data, we also select data from similar time periods (late 2010s) to mitigate domain discrepancy between source and target languages as per previous research (Kim et al., 2020). We also downsample the English part of WMT NewsCrawl corpus so that our English and the corresponding foreign news data are equal in size.

## 6 Results & Analysis

### 6.1 Tatoeba Dataset[11]

We mine bitexts on the Tatoeba test set in 64 different languages (listed in Section 5.1) using the primary mining procedure described in Algorithm 1 with *intersection* retrieval, in addition to seven different secondary mining procedures. The methods and corresponding results are reported in Table 3 in terms of F1, and are summarized as follows, in the order in which they appear in the table:

[11] https://github.com/AlexJonesNLP/alt-bitexts/blob/main/source/retrieve_tatoeba_results.ipynb

1. Raw cosine similarity (Reimers and Gurevych, 2020): find closest sentence pair using cosine similarity only
2. "Vanilla" margin scoring: perform forward and backward searches and take intersection
3. Margin scoring, threshold=1.06: margin scoring with a threshold of 1.06, à la Schwenk et al. (2019b) (Method **1** in Table 1)
4. . . . threshold=1.20: optimal BUCC threshold
5. Margin scoring using EN sentences translated to XX (Method
6. . . . using XX sentences translated to EN
7. The strict intersection of pairs generated by methods 2, 5, and 6
8. The pairwise intersection of pairs generated by method 2, 5, and 6 (Method **2**)

We report F1 instead of accuracy because the intersection methods (in both primary and secondary procedures) permit less than 100% recall.

The results are broken down across languages by resource availability (as in "high-resource" or "low-resource"), as ranked on a 0-5 scale[12], and summarized in Table 4. Language-specific results are given in Table 5.

Because many of the languages in Table 3 lack support in GNMT, the dominant method overall is vanilla margin scoring (Method 2 above), being the best-performing method on 28/64 languages[13] and seeing an average gain over the baseline (Method 1) of +5.2 for all languages and +6.9 for languages on which it was the best-performing method. However, for languages with translation support, the pairwise intersection method (Method 8) won out, with an average gain over the baseline of +4.0, in contrast to vanilla margin scoring (+3.6). Moreover, pairwise intersection increased F1-score over vanilla margin scoring for 26/38 of these languages. In fact, among these 38 languages, vanilla margin scoring outperformed translation-based or hybrid (intersection) methods on only 11 languages, five of which were translated zero-shot (e.g. substituting Standard German for Low German or Esperanto for Ido when translating).

Simply translating non-English sentences into English before mining (Method 6) also performed well, netting best results on 18 languages and outperforming other methods on resource level 3 (+4.3 F1 over baseline) and level 4 (+1.8) languages. Meanwhile, pairwise intersection performed best on level 0 (+7.3) and level 2 (+2.6) languages, with vanilla margin scoring taking home the bread on level 1 (+5.2). Notably, thresholding (Methods 3&4) almost exclusively did more harm than good (Method 3 achieved best results on only 3 languages, and Method 4 on none), and though reporting this may be viewed as a strawman attack on thresholding in the context of this task—identifying bitexts in gold-standard parallel corpora, as opposed to noisy comparable corpora—we note that gold-standard bitexts simply don't reliably lie beyond some set threshold, as shown in the right-two graphs in Figure 3. Additionally, performing strict intersection (Method 7) led to decreased F1 due to dampened recall, suggesting majority voting is a better way to combine signals from similarity searches than all-or-nothing voting. We note as well that 6 languages on which vanilla margin scoring performs best are constructed (e.g. Esperanto, Ido) and 2 are extinct (Old English and Old Russian), inflating those results somewhat from a natural/living-language-focused perspective.

### 6.2 NMT[14]

In Table 2, we show the performance in terms of BLEU scores of various NMT training schemes on the same WMT'19 test set. We train the supervised MT part of our pipeline system with gold-standard data (human translation WMT'19 data), our mined comparable/pseudoparallel ("silver-standard") data, and combinations of both i.e., training with comparable data followed by training with gold-standard data. We also provide Google Massively Multilingual MT performance on the same WMT'19 test set (Wu et al., 2016).

As we can see in Table 2, our method of mining bitext without thresholding (Method **2**) results in higher BLEU performance than bitext mined using margin scoring with a threshold of 1.06 (Method **1**), which is a commonly used threshold recommended by previous works for mining bitext using margin scoring. Method **2** also results in the best en→gu performance, which outperforms previous unsupervised or supervised works. It outperforms the best previous work that uses WMT'19 data and iterative bitext mining by +3.3 BLEU. Since we do not perform iterative mining, if we consider the same previous work without iterative mining i.e., Tran

[12] rb.gy/psmfnz

[13] Note that 6/64 languages lack a resource categorization, so we report results on the remaining 58

[14] https://github.com/AlexJonesNLP/alt-bitexts/tree/main/source

et al. (2020) Iter 1, ours outperforms that model by +12.1 BLEU in en→gu direction and by +8.3 BLEU in gu→en direction.

When combined with supervised i.e., gold-standard data for training, our method for mining bitext which does not use any thresholding (Method **2**+WMT'19) also outperforms the same model which uses bitext mined using margin scoring with a threshold of 1.06 (Method **1**+WMT'19). Method **2**+WMT'19 also results in the best en→kk performance, which outperforms previous unsupervised or supervised works. It outperforms the best previous work that uses WMT'19 data and iterative bitext mining by +4.7 BLEU. Since we do not perform iterative mining, if we consider the same previous work without iterative mining i.e., Tran et al. (2020) Iter 1, ours outperforms that model by +5.6 BLEU in en→kk direction and by +2.8 BLEU in kk→en direction. It is also worth noting that for training our pipeline model we use fixed hyperparameter settings suggested in the XLM repository while previous works perform extensive hyperparameter tuning. We believe our performance can be improved further by tuning our hyperparameter settings.

These results on low resource MT further demonstrate the superiority of our method for mining bitext without thresholding compared to margin scoring with thresholding for downstream low resource MT applications.

### 6.3 The Problem with Thresholding

One benefit of our proposed approach is that it is threshold-agnostic, unlike previous approaches (Artetxe and Schwenk, 2019a; Schwenk et al., 2019a,b). Furthermore, our results on semi-supervised (really, supervised+unsupervised+semi-supervised) MT show that adding data mined using the pairwise intersection method (Method **2** in Table 2) improves over the WMT'19 baseline, while adding data mined using a threshold of 1.2 actually *hurts* performance considerably. These results are in line with the somewhat arbitrary nature of margin score thresholding observed elsewhere. Figure 3 shows distributions of margin scores on sentence pairs mined on our English-Kazakh comparable corpus (using document-level mining), on the BUCC English-French training

| Corpus | kk→en | en→kk | gu→en | en→gu |
|---|---|---|---|---|
| **Unsupervised** | | | | |
| Kim et al. (2020) | 2.0 | 0.8 | 0.6 | 0.6 |
| **Supervised** | | | | |
| WMT'19 (Kim et al., 2020) | 10.3 | 2.4 | 9.9 | 3.5 |
| WMT'19 (Tran et al., 2020) Iter 1 | 9.8 | 3.4 | 8.1 | 8.1 |
| WMT'19 (Tran et al., 2020) Iter 3 | **13.2** | 4.3 | **18.0** | 16.9 |
| Google MT (Wu et al., 2016) | **28.9** | **23.1** | **26.2** | **31.4** |
| **Our pipeline: unsup.+Sup.** | | | | |
| WMT'19 | 11.2 | 7.3 | 5.7 | 10.2 |
| Method **1** | 6.6 | 4.1 | 16.2 | 19.8 |
| Method **2** | 8.6 | 6.1 | 16.4 | **20.2** |
| Method **1**+WMT'19 | 11.8 | 7.9 | 15.4 | 18.5 |
| Method **2**+WMT'19 | 12.6 | **9.0** | 15.8 | 19.1 |
| **Previous training procedure** | | | | |
| Method **2**+WMT'19 | 11.8 | 7.9 | — | — |
| Method **3**+WMT'19 | 11.8 | 8.1 | — | — |
| Method **4**+WMT'19 | 12.2 | 8.5 | — | — |
| LaBSE (threshold=1.20)+WMT'19 | 8.9 | 6.6 | — | — |

Table 2: NMT training schemes and corresponding BLEU scores on WMT'19 test set. We train supervised systems with gold-standard data, comparable/pseudoparallel ("silver-standard") data, and combinations of both. We also try supplementing unsupervised training with each of these three types of supervised data, providing full supervision, weak supervision, or both. We also provide a benchmark from Wu et al. (2016). The methods listed in the table are given in Table 1.

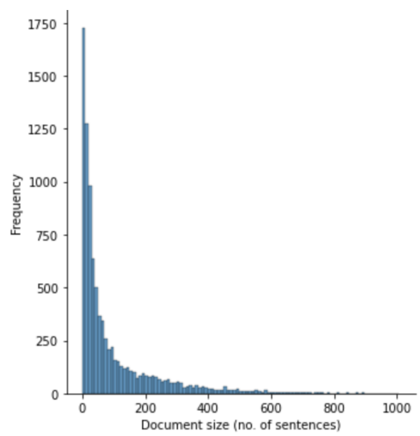



Figure 1: Distribution of document sizes in English component of EN-GU Wikipedia corpus.

data (globally mined) and on two Tatoeba test sets, namely English-Maltese and English-Telugu (also globally mined).

First, we note that the margin distributions on the latter two datasets—for which we've plotted 99%+ ground-truth pairs—appear approximately normally distributed over a significantly large range (around size 0.7-1 for both), rendering it impossible to choose a single threshold that catches all pairs. This is in line with the much more extensive results displayed in Table 3, in which only a few of the 64 language pairs aren't harmed by even a low threshold of 1.06. On the BUCC data, the margin scores appear almost perfectly normally distributed, seeming to belie our critique. However, a close analysis of this distribution reveals a small local maximum around 1.3, most likely representing the gold-standard pairs that were injected into the BUCC corpus (Zweigenbaum et al., 2017, 2018). This may explain the success of this threshold in others' studies using this dataset [15].

Another issue is that the optimal margin threshold appears dependent on the size of the search space, posing a particular issue for document-level mining in which this size differs from document to document. The choice of margin threshold is discussed in both Schwenk et al. (2019b) and Schwenk et al. (2019a), but neither address the topic with much rigor. Schwenk et al. (2019a) examines a very narrow range of margin thresholds for only two language pairs (four directions) on bitexts mined from Wikipedia, but yield no truly conclusive results, nor any disciplined method for selecting the optimal threshold. The same may be said of Schwenk et al. (2019b), in which the optimal threshold is justified by BLEU evaluation on a single language pair. The margin score Schwenk et al. (2019b) uses to mine globally over millions or billions of sentences performs terribly on our corpus using document level mining, with higher thresholds yielding only marginal improvements. To the contrary, we speculate that the various signals in our voting-based approach provide the same sort of denoising effect as other voting-based approaches (e.g. voting models), and while Table 2 shows that the generated bitexts still benefit from thresholding (see results under "Previous training procedure," which simply involved training for less time on less GPUs), voting alone acts as a sufficient heuristic to produce reasonably good precision and recall. As can be seen by Figure 2, the proposed approach does *not* perform a sort of implicit thresholding.

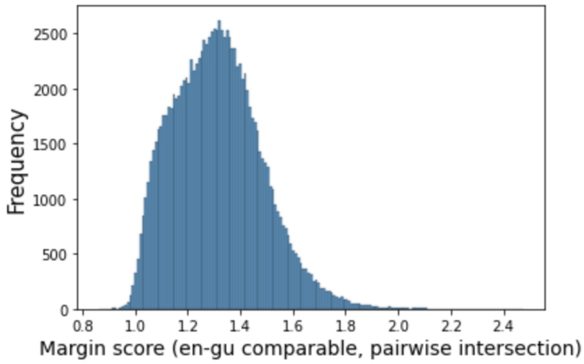



Figure 2: Margin scores on EN-GU pairs mined using the pairwise intersection method.

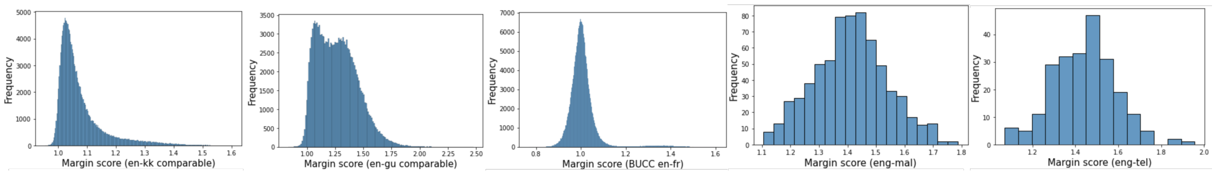
# 7 Discussion

## 7.1 Cross-lingual Alignment in Multilingual Sentence Embedding Models

One upshot of the results of our approach is that evidently, there is something to be gained from translating texts before performing similarity search, that there is something salient in this signal that is not in the signal generated by the original text's embedding. What this points to is some deficiency in the cross-lingual alignment in LaBSE (and likely in other cross-lingual sentence embedding models as well).

Cross-lingual alignment has been investigated in the context of monolingual embeddings rather rigorously. Vulić et al. (2020) investigates the causes of cross-lingual misalignment between monolingual embedding spaces, and pinpoints language model training data size and training regimes as the main culprits, to the exclusion of typological factors such as morphology and word order. Furthermore, Pires et al. (2019) and Wu and Dredze (2019) investigate the cross-lingual alignment ability of mBERT Devlin et al. (2019) by examining the zero-shot case (i.e. fine-tuning on one language and predicting on others) for a variety of tasks. However, while Pires et al. (2019); Wu and Dredze (2019) each touch on linguistic factors, neither thoroughly investigate their effect on cross-lingual *alignment* (as opposed to *transfer*, which we argue may be correlated but not perfectly so), or make a rigorous effort to control for the other factors at play in cross-lingual LMs, such as monolingual/bilingual training data size and size of same-family training

[15] https://www.sbert.net/examples/applications/parallel-sentence-mining/README.html




(a) Margin scores on English-Kazakh comparable corpora

(b) Margin scores on English-Gujarati comparable corpora

(c) Margin scores on BUCC '17/18 English-French data

(d) Margin scores on Tatoeba English-Maltese data

(e) Margin scores on Tatoeba English-Telugu data

Figure 3: Distributions of margin scores across various datasets, achieved using *intersection* retrieval with no threshold. The left three graphs are mined from comparable corpora, while the right two are mined from gold-standard bitexts and contain 99%+ ground-truth pairs.

data for a given language. While such an investigation lies beyond the scope of this paper, we believe our results make practical use of a deficient cross-lingual alignment, and echo Artetxe et al. (2020)'s call for more thorough probing, linguistic and otherwise, of cross-lingual models.

### 7.2 Energy and Resource Considerations

While catalyzed by such tools as FAISS (Johnson et al., 2017), bitext mining is inherently an incredibly expensive task, especially when performed globally. Though extensive computational optimization has allowed the search space to grow to billions of sentences (Fan et al., 2020), these global-level procedures still require hundreds of GPUs, leaving a sizable environmental footprint (Schwartz et al., 2019; Strubell et al., 2019) and limiting the number of researchers and institutions to whom this method is available. Our approach, while requiring the upfront cost of translating entire corpora of sentence pairs, operates at the document level (which, as Schwenk et al. (2019b) note, is available for Wikipedia but not for corpora like Common Crawl) and provides a heuristic measure that may eliminate the need for laboriously tuning a somewhat arbitrary margin threshold.

### 7.3 Linguistic Diversity

Artetxe et al. (2020) outlines many of the key issues in unsupervised cross-lingual learning and evaluation, among which are the verisimilitude of the training conditions and the lack of a cross-lingual benchmark for many tasks. On the one hand, our method relies on supervised data and thus isn't applicable to the most low-resource language pairs, helping instead a niche of mid-low-resource languages (see Table 4). Also, we report results on the Tatoeba test set from Artetxe and Schwenk (2019b), which contains only English-aligned sentence pairs. However, as Artetxe et al. (2020) note, the fully supervised setting isn't as rare as it's often made out to be, and our relatively lightweight, heuristic-based approach suits a practical research or development setting. Nonetheless, we would like to perform extensive linguistic probing on the bitext mining task using non-English-aligned corpora, and suggest Tiedemann (2020) as a possible resource for this inquiry, as it extends the Tatoeba test set from Artetxe and Schwenk (2019b) to nearly 3000 language pairs. As Schwenk et al. (2019b) note, the factors affecting bitext quality and quantity aren't fully understood. Such massively multilingual, non-English-centric benchmarks will enable richer and more inclusive cross-lingual research (Joshi et al., 2020), building on top of current benchmarks such as XTREME, XGLUE, and XNLI (Hu et al., 2020; Liang et al., 2020; Conneau et al., 2018), and supplementing probing efforts such as Pires et al. (2019) and Wu and Dredze (2019).

## 8 Conclusions

In this paper, we propose a novel method of mining sentence pairs from both comparable and parallel corpora, and demonstrate success on both the Tatoeba gold-standard bitext mining task and on mining pseudo-parallel sentences for NMT. We uncover the problematic nature of setting a similarity score threshold for this task, particularly in the context of margin scoring with document-level mining, showing that thresholding is not a one-size-fits-all approach. On the Tatoeba dataset, we set what we believe to be new benchmarks for 64 languages and reveal an intriguing cross-lingual division across languages by their resource availability with respect to which mining approach performs best, with the voting-based approach involving bidirectional translation providing superior results on languages for which a supervised NMT system was available. We contribute novel insights regarding the cross-lingual alignment of multilingual language models, exploit its deficiencies, and propose further probing efforts to examine the linguistic and technical factors affecting this alignment. In future work, we also hope to investigate how cross-lingual alignment may be improved in cross-lingual LMs, and how our mining methods transfer to the large-scale, global mining scenario.

## 9 Acknowledgements

We would like to thank Mohammad Sadegh Rasooli for offering ideas and insights, as well as Nils Reimers, Iryna Gurevych, and the Facebook AI team for contributing their open-source models and data for research. We would also like to acknowledge Boston University for providing funding and computational resources.

## References

2021. Unsupervised Machine Translation is Useful: It Just Needs Good Company.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively Multilingual Word Embeddings. *arXiv e-prints, arXiv:1602.01925.*

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *International Conference on Learning Representations 2017.*

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised Statistical Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019a. An Effective Approach to Unsupervised Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019b. Bilingual Lexicon Induction through Unsupervised Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised Neural Machine Translation. In *International Conference on Learning Representations 2018.*

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A Call for More Rigor in Unsupervised Cross-lingual Learning. *arXiv e-prints*, page arXiv:2004.14958.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

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.

Adoni Azpeitia, Thierry Etchegoyhen, and Eva Martinez 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).

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2017. Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45, Vancouver, Canada. Association for Computational Linguistics.

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.

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).

M.V. Butz and S.W. Wilson. 2002. An Algorithmic Description of XCS. *Soft Computing*, 6:144–153.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual Alignment of Contextual Word Representations. In *International Conference on Learning Representations*.

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2020. word2word: A Collection of Bilingual Lexicons for 3,564 Language Pairs. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3036–3045, Marseille, France. European Language Resources Association.

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.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv e-prints*, page arXiv:1705.02364.

Alexis Conneau and Guillaume Lample. 2019a. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau and Guillaume Lample. 2019b. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. *arXiv e-prints*, page arXiv:1809.05053.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv e-prints*, page arXiv:1412.6568.

Thierry Etchegoyhen and Andoni Azpeitia. 2016. Set-theoretic alignment for comparable corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2009–2018, Berlin, Germany. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. *arXiv e-prints*, page arXiv:2010.11125.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT Sentence Embedding. *arXiv e-prints*, page arXiv:2007.01852.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1051–1057, Geneva, Switzerland. COLING.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised Parallel Sentence Extraction from Comparable Corpora. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 7–13, Bruges, Belgium.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. *arXiv e-prints*, page arXiv:2003.11080.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv e-prints*, page arXiv:1702.08734.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A. Smith. 2020. Unsupervised Bitext Mining and Translation via Self-trained Contextual Embeddings.

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.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-Thought Vectors. *arXiv e-prints*, page arXiv:1506.06726.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv e-prints*, page arXiv:1711.00043.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. *arXiv e-prints*, page arXiv:2004.01401.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{BERT}a: A Robustly Optimized {BERT} Pre-training Approach.

Timothy Campbell Lukins. 2002. *Dynamically Developing Novel and Useful Behaviours: A First Step in Animat Creativity*. Ph.D. thesis.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When Does Unsupervised Machine Translation Work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 265–272, Boston, Massachusetts, USA. Association for Computational Linguistics.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nima Pourdamghani, Nada Aldarrab, Marjan Ghazvininejad, Kevin Knight, and Jonathan May. 2019. Translating translationese: A two-step approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3057–3062, Florence, Italy. Association for Computational Linguistics.

Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. 2018. Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

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

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. *arXiv e-prints*, page arXiv:1907.10597.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv e-prints*, page arXiv:1907.05791.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. *arXiv e-prints*, page arXiv:1911.04944.

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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. *arXiv e-prints*, page arXiv:1905.02450.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In *International Conference on Learning Representations*.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. *arXiv e-prints*, page arXiv:2010.06354.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual Retrieval for Iterative Self-Supervised Training. *arXiv e-prints*, page arXiv:2006.09526.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are All Good Word Vector Spaces Isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework. In *International Conference on Learning Representations*.

Lijun Wu, Jinhua Zhu, Di He, Fei Gao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Machine Translation With Weakly Paired Documents. 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 4375–4384, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv e-prints*, page arXiv:1609.08144.

Hongzhi Xu, Mitchell Marcus, Charles Yang, and Lyle Ungar. 2018. Unsupervised morphology learning with statistical paradigms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 44–54, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3180–3189, Florence, Italy. Association for Computational Linguistics.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, page 745, USA. IEEE Computer Society.

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 *Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan.

# A Appendix

| Procedure | afr | amh | ang | arq | arz | ast | awa | aze | bel | ben | ber | bos | bre |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Raw cosine similarity (*Acc=F1*) | 97.4 | 94 | 64.2 | 46.2 | 78.4 | 90.6 | 73.2 | 96.1 | 96.2 | 91.3 | 10.4 | 96.2 | 17.3 |
| Margin scoring, *intersection*, no threshold (*F1*) | 98.7 | 94.2 | **73.4** | **57.2** | **84.6** | **94.3** | **83.4** | 97.4 | 97.5 | 92.4 | **14.2** | 96.6 | **21.5** |
| *Precision* | *99.9* | *96.9* | *88.4* | *80.0* | *93.6* | *98.3* | *95.5* | *99.3* | *99.1* | *96.6* | *30.9* | *98.0* | *38.5* |
| *Recall* | *97.6* | *91.7* | *62.7* | *44.5* | *77.1* | *90.6* | *74.0* | *95.6* | *95.9* | *88.5* | *9.2* | *95.2* | *14.9* |
| Margin scoring, *intersection*, threshold = 1.06 (*F1*) | 98.2 | 94.5 | 72.9 | 56.0 | 84.0 | 94.2 | 80.5 | 97.2 | 97.3 | 91.8 | 13.4 | 96.4 | 21.3 |
| *Precision* | *100* | *97.5* | *90.1* | *85.0* | *95.7* | *99.1* | *97.0* | *99.3* | *99.1* | *96.9* | *44.4* | *98.0* | *54.1* |
| *Recall* | *96.5* | *91.7* | *61.2* | *41.7* | *74.8* | *89.8* | *68.8* | *95.3* | *95.6* | *87.3* | *7.9* | *94.9* | *13.3* |
| Margin scoring, *intersection*, threshold = 1.20 (*F1*) | 89.5 | 82.5 | 59.1 | 43.6 | 76.9 | 92.4 | 57.2 | 89.6 | 94.8 | 78.6 | 11.8 | 90.5 | 13.4 |
| *Precision* | *100* | *100* | *96.6* | *97.3* | *98.1* | *99.1* | *98.9* | *99.8* | *99.5* | *99.1* | *90.0* | *99.0* | *92.3* |
| *Recall* | *81.0* | *70.2* | *42.5* | *28.1* | *63.3* | *86.6* | *40.3* | *81.4* | *90.5* | *65.1* | *6.3* | *83.3* | *7.2* |
| Margin scoring, *intersection*, en-xx (*F1*) | 98.4 | 93.2 | * | * | * | * | * | 96.7 | 97.6 | 91.8 | * | 96.3 | * |
| *Precision* | *99.6* | *96.8* | * | * | * | * | * | *98.6* | *99.1* | *96.5* | * | *98.2* | * |
| *Recall* | *97.3* | *89.9* | * | * | * | * | * | *94.9* | *96.1* | *87.6* | * | *94.4* | * |
| Margin scoring, *intersection*, xx-en (*F1*) | **99.0** | **95.7** | * | * | * | * | * | **97.6** | 97.6 | 92.0 | * | 97.3 | * |
| *Precision* | *99.8* | *98.1* | * | * | * | * | * | *99.0* | *99.1* | *96.3* | * | *98.8* | * |
| *Recall* | *98.2* | *93.5* | * | * | * | * | * | *96.3* | *96.1* | *88.0* | * | *95.8* | * |
| Margin scoring, *intersection*, strict intersection (*F1*) | 98.1 | 93.7 | * | * | * | * | * | 96.2 | 96.9 | 89.8 | * | 96.0 | * |
| *Precision* | *100* | *100* | * | * | * | * | * | *99.8* | *99.8* | *99.3* | * | *100* | * |
| *Recall* | *96.2* | *88.1* | * | * | * | * | * | *92.8* | *94.2* | *82.0* | * | *92.4* | * |
| Margin scoring, *intersection*, pairwise intersection (*F1*) | 98.9 | 95.4 | * | * | * | * | * | 97.5 | **97.9** | **93.0** | * | 97.1 | * |
| *Precision* | *99.9* | *98.7* | * | * | * | * | * | *99.3* | *99.6* | *97.9* | * | *98.8* | * |
| *Recall* | *97.9* | *92.3* | * | * | * | * | * | *95.9* | *96.2* | *88.6* | * | *95.5* | * |

| Procedure | cbk | ceb | cha | cor | csb | cym | dsb | dtp | epo | eus | fao | fry | gla |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Raw cosine similarity (*Acc=F1*) | 82.5 | 70.9 | 39.8 | 12.8 | 56.1 | 93.6 | 69.3 | 13.3 | 98.4 | 95.8 | 90.6 | 89.9 | 88.8 |
| Margin scoring, *intersection*, no threshold (*F1*) | **89.5** | 79.3 | **49.3** | **18.8** | **69.5** | 96.2 | **80.7** | **18.8** | **99.0** | 96.8 | **94.9** | 93.7 | 91.9 |
| *Precision* | *96.7* | *91.1* | *65.9* | *45.2* | *86.5* | *98.9* | *94.7* | *37.5* | *99.7* | *98.4* | *98.0* | *96.9* | *97.1* |
| *Recall* | *83.2* | *70.2* | *39.4* | *11.9* | *58.1* | *93.6* | *70.4* | *12.5* | *98.4* | *95.2* | *92.0* | *90.8* | *87.3* |
| Margin scoring, *intersection*, threshold = 1.06 (*F1*) | 87.1 | 78.5 | 47.8 | 16.2 | 68.0 | 95.6 | 79.1 | 18.5 | 99.0 | 96.4 | 93.4 | 93.1 | 91.2 |
| *Precision* | *97.8* | *93.3* | *75.0* | *64.1* | *90.2* | *99.1* | *95.6* | *56.1* | *99.9* | *98.5* | *98.7* | *97.5* | *97.3* |
| *Recall* | *78.6* | *67.7* | *35.0* | *9.3* | *54.5* | *92.3* | *67.4* | *11.1* | *98.2* | *94.4* | *88.5* | *89.0* | *85.8* |
| Margin scoring, *intersection*, threshold = 1.20 (*F1*) | 71.5 | 67.4 | 44.3 | 9.0 | 54.2 | 86.0 | 93.4 | 15.2 | 97.9 | 92.6 | 84.5 | 89.5 | 80.3 |
| *Precision* | *99.6* | *98.7* | *85.4* | *100* | *95..0* | *99.3* | *99.6* | *87.4* | *99.9* | *99.2* | *99.0* | *99.3* | *98.9* |
| *Recall* | *55.7* | *51.2* | *29.9* | *4.7* | *37.9* | *75.8* | *46.6* | *8.3* | *96.0* | *86.8* | *73.7* | *81.5* | *67.6* |
| Margin scoring, *intersection*, en-xx (*F1*) | * | 78.6 | * | 15.0 | * | 96.3 | 76.2 | * | 98.5 | 96.4 | * | **96.4** | 92.6 |
| *Precision* | * | *90.6* | * | *36.0* | * | *98.9* | *95.0* | * | *99.5* | *98.6* | * | *98.8* | *97.1* |
| *Recall* | * | *69.3* | * | *9.5* | * | *93.9* | *63.7* | * | *97.6* | *94.3* | * | *94.2* | *88.4* |
| Margin scoring, *intersection*, xx-en (*F1*) | * | **86.1** | * | 17.3 | * | **97.3** | 67.3 | * | 98.9 | **97.6** | * | 95.6 | **93.9** |
| *Precision* | * | *94.2* | * | *41.8* | * | *98.9* | *85.5* | * | *99.6* | *98.8* | * | *97.6* | *97.5* |
| *Recall* | * | *79.2* | * | *10.9* | * | *95.7* | *55.5* | * | *98.3* | *96.4* | * | *93.6* | *90.6* |
| Margin scoring, *intersection*, strict intersection (*F1*) | * | 77.3 | * | 13.0 | * | 95.2 | 63.0 | * | 98.5 | 96.2 | * | 93.9 | 89.9 |
| *Precision* | * | *99.2* | * | *68.6* | * | *100* | *99.1* | * | *100* | *99.5* | * | *98.7* | *99.3* |
| *Recall* | * | *63.3* | * | *7.2* | * | *90.8* | *46.1* | * | *97.1* | *93.1* | * | *89.6* | *82.1* |
| Margin scoring, *intersection*, pairwise intersection (*F1*) | * | 81.8 | * | 18.7 | * | 96.7 | 79.4 | * | 98.8 | 96.8 | * | 95.8 | 93.5 |
| *Precision* | * | *96.0* | * | *47.9* | * | *99.1* | *97.3* | * | *99.6* | *98.6* | * | *98.8* | *98.0* |
| *Recall* | * | *71.3* | * | *11.6* | * | *94.4* | *67.0* | * | *98.1* | *95.2* | * | *93.1* | *89.4* |

| Procedure | gle | gsw | hsb | ido | ile | ina | isl | jav | kab | kaz | khm | kur | kzj |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Raw cosine similarity (*Acc=F1*) | 95.0 | 52.1 | 71.2 | 90.9 | 87.1 | 95.8 | 96.2 | 84.4 | 6.2 | 90.5 | 83.2 | 87.1 | 14.2 |
| Margin scoring, *intersection*, no threshold (*F1*) | 96.6 | 62.0 | 81.6 | **95.1** | **93.0** | **97.4** | **97.9** | **92.2** | **7.7** | 92.6 | 86.8 | 92.1 | **20.8** |
| *Precision* | *98.7* | *85.1* | *94.6* | *98.7* | *98.4* | *99.0* | *99.4* | *98.9* | *19.4* | *96.8* | *93.0* | *98.1* | *41.3* |
| *Recall* | *94.6* | *48.7* | *71.8* | *91.7* | *88.1* | *95.9* | *96.4* | *86.3* | *4.8* | *88.7* | *81.3* | *86.8* | *13.9* |
| Margin scoring, *intersection*, threshold = 1.06 (*F1*) | 95.9 | 60.2 | 79.7 | 94.1 | 91.7 | 96.9 | 97.5 | 91.6 | 7.3 | 92.2 | 86.4 | 91.4 | 20.0 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Precision* | *98.9* | *89.8* | *94.9* | *99.0* | *99.0* | *99.0* | *99.4* | *99.4* | *31.3* | *96.9* | *94.7* | *98.3* | *55.2* |
| *Recall* | *93.1* | *45.3* | *68.7* | *89.7* | *85.4* | *95.0* | *95.7* | *84.9* | *4.1* | *87.8* | *79.5* | *85.4* | *12.2* |
| Margin scoring, *intersection*, threshold = 1.20 (*F1*) | 84.7 | 43.7 | 67.8 | 88.5 | 77.9 | 94.5 | 91.0 | 83.6 | 5.0 | 85.7 | 76.4 | 82.9 | 15.1 |
| *Precision* | *100* | *97.1* | *99.6* | *99.9* | *99.8* | *99.4* | *99.9* | *99.3* | *78.8* | *99.1* | *98.7* | *99.7* | *94.3* |
| *Recall* | *73.5* | *28.2* | *51.3* | *79.5* | *63.8* | *90.0* | *83.6* | *72.2* | *2.6* | *75.5* | *62.3* | *71.0* | *8.2* |
| Margin scoring, *intersection*, en-xx (*F1*) | 96.9 | 58.7 | 76.6 | 80.4 | 76.4 | 96.3 | 91.9 | * | * | 92.6 | 87.3 | 92.0 | * |
| *Precision* | *98.8* | *80.6* | *92.9* | *91.8* | *90.1* | *99.4* | *96.4* | * | * | *97.0* | *93.9* | *97.5* | * |
| *Recall* | *95.2* | *46.2* | *65.2* | *71.6* | *66.3* | *93.5* | *87.8* | * | * | *88.7* | *81.6* | *97.1* | * |
| Margin scoring, *intersection*, xx-en (*F1*) | 97.7 | 59.3 | 80.0 | 82.1 | 78.7 | 95.8 | 80.8 | * | * | **93.5** | 87.5 | **95.6** | * |
| *Precision* | *99.0* | *83.1* | *93.1* | *95.4* | *93.0* | *98.6* | *93.8* | * | * | *96.8* | *93.5* | *99.2* | * |
| *Recall* | *96.4* | *46.2* | *70.2* | *72.0* | *68.2* | *93.2* | *71.0* | * | * | *90.4* | *82.1* | *92.2* | * |
| Margin scoring, *intersection*, strict intersection (*F1*) | 95.6 | 55.1 | 74.7 | 73.2 | 67.0 | 94.9 | 78.2 | * | * | 91.2 | 85.6 | 90.3 | * |
| *Precision* | *99.6* | *91.2* | *96.1* | *100* | *99.8* | *99.7* | *99.8* | * | * | *99.2* | *98.2* | *99.4* | * |
| *Recall* | *92.0* | *39.3* | *61.2* | *57.7* | *50.4* | *90.6* | *64.3* | * | * | *84.3* | *75.9* | *82.7* | * |
| Margin scoring, *intersection*, pairwise intersection (*F1*) | **97.8** | **62.3** | **81.7** | 91.1 | 88.4 | 97.1 | 96.6 | * | * | 93.1 | **87.8** | 94.0 | * |
| *Precision* | *99.3* | *86.4* | *94.8* | *99.5* | *99.3* | *99.1* | *99.3* | * | * | *97.5* | *94.9* | *99.2* | * |
| *Recall* | *73.5* | *28.2* | *51.3* | *79.5* | *63.8* | *90.0* | *83.6* | *72.2* | *2.6* | *75.5* | *62.3* | *71.0* | *8.2* |
| **Procedure** | **lat** | **lfn** | **mal** | **mhr** | **nds** | **nno** | **nov** | **oci** | **orv** | **pam** | **pms** | **swg** | **swh** |
| Raw cosine similarity (*Acc=F1*) | 82.0 | 71.2 | 98.9 | 19.2 | 81.2 | 95.9 | 78.2 | 69.9 | 46.8 | 13.6 | 67.0 | 65.2 | 88.6 |
| Margin scoring, *intersection*, no threshold (*F1*) | **89.0** | **80.7** | **99.3*** | 26.3 | **89.0** | 97.5 | **85.4** | **78.7** | **57.4** | **17.9** | **78.9** | **80.4** | 93.2 |
| *Precision* | *96.8* | *93.4* | *99.7* | *46.0* | *96.9* | *99.4* | *93.5* | *90.6* | *78.6* | *34.6* | *92.8* | *95.1* | *97.7* |
| *Recall* | *82.4* | *71.0* | *98.8* | *18.4* | *82.2* | *95.7* | *78.6* | *69.6* | *45.3* | *12.1* | *68.6* | *69.6* | *89.0* |
| Margin scoring, *intersection*, threshold = 1.06 (*F1*) | 87.2 | 79.4 | **99.3*** | **26.3** | 87.6 | 97.2 | 83.0 | 77.7 | 55.9 | 17.4 | 76.3 | 77.0 | 92.5 |
| *Precision* | *97.6* | *94.7* | *99.7* | *59.3* | *98.3* | *99.5* | *94.5* | *93.1* | *83.6* | *50.2* | *94.4* | *96.0* | *98.8* |
| *Recall* | *78.7* | *68.4* | *98.8* | *16.9* | *79.1* | *95.1* | *73.9* | *66.6* | *42.0* | *10.5* | *64.0* | *64.3* | *86.9* |
| Margin scoring, *intersection*, threshold = 1.20 (*F1*) | 72.6 | 68.8 | 96.4 | 18.0 | 74.8 | 92.1 | 77.3 | 65.8 | 37.0 | 11.7 | 63.0 | 72.3 | 81.8 |
| *Precision* | *99.5* | *98.5* | *99.7* | *90.1* | *99.3* | *99.9* | *98.8* | *98.8* | *96.5* | *85.1* | *98.4* | *98.5* | *100* |
| *Recall* | *57.2* | *52.9* | *93.3* | *10.0* | *60.0* | *85.5* | *63.4* | *49.3* | *22.9* | *6.3* | *46.3* | *57.1* | *69.2* |
| Margin scoring, *intersection*, en-xx (*F1*) | 83.5 | * | 98.0 | * | 86.0 | 97.3 | * | * | * | * | * | * | 94.9 |
| *Precision* | *95.1* | * | *99.5* | * | *97.5* | *99.3* | * | * | * | * | * | * | *98.6* |
| *Recall* | *74.4* | * | *96.5* | * | *76.9* | *95.4* | * | * | * | * | * | * | *91.5* |
| Margin scoring, *intersection*, xx-en (*F1*) | 86.1 | * | 98.2 | * | 83.8 | 97.7 | * | * | * | * | * | * | 95.3 |
| *Precision* | *95.6* | * | *99.6* | * | *95.2* | *99.4* | * | * | * | * | * | * | *98.1* |
| *Recall* | *78.3* | * | *96.9* | * | *74.9* | *96.1* | * | * | * | * | * | * | *92.6* |
| Margin scoring, *intersection*, strict intersection (*F1*) | 81.7 | * | 97.1 | * | 80.1 | 96.6 | * | * | * | * | * | * | 92.1 |
| *Precision* | *98.2* | * | *100* | * | *99.3* | *99.8* | * | * | * | * | * | * | *100* |
| *Recall* | *69.9* | * | *94.3* | * | *67.2* | *93.7* | * | * | * | * | * | * | *85.4* |
| Margin scoring, *intersection*, pairwise intersection (*F1*) | 88.8 | * | 99.2 | * | 88.4 | **97.8** | * | * | * | * | * | * | **95.5** |
| *Precision* | *97.1* | * | *99.9* | * | *98.3* | *99.6* | * | * | * | * | * | * | *99.4* |
| *Recall* | *81.7* | * | *98.5* | * | *80.4* | *96.0* | * | * | * | * | * | * | *91.2* |
| **Procedure** | **tam** | **tat** | **tel** | **tgl** | **tuk** | **tzl** | **uig** | **uzb** | **war** | **wuu** | **xho** | **yid** | |
| Raw cosine similarity (*Acc=F1*) | 90.7 | 87.9 | 98.3 | 97.4 | 80.0 | 63.0 | 93.7 | 86.8 | 65.3 | 90.3 | 91.9 | 91.0 | * |
| Margin scoring, *intersection*, no threshold (*F1*) | 93.0 | 92.0 | **99.1*** | 98.6 | 86.8 | **71.0** | 95.4 | 91.1 | **75.8** | **94.8** | 94.2 | 95.2 | * |
| *Precision* | *97.5* | *97.4* | *99.6* | *99.7* | *95.8* | *82.3* | *98.3* | *96.8* | *89.5* | *98.8* | *97.7* | *98.7* | * |
| *Recall* | *88.9* | *87.1* | *98.7* | *97.6* | *79.3* | *62.5* | *92.7* | *86.0* | *65.7* | *91.1* | *90.8* | *92.0* | * |
| Margin scoring, *intersection*, threshold = 1.06 (*F1*) | 92.8 | 91.3 | **99.1*** | 98.4 | 87.3 | 70.9 | 95.1 | 90.7 | 73.8 | 94.0 | 94.2 | 94.3 | * |
| *Precision* | *97.8* | *97.9* | *99.6* | *99.8* | *99.4* | *87.3* | *98.3* | *97.1* | *93.5* | *99.0* | *97.7* | *99.1* | * |
| *Recall* | *88.3* | *85.5* | *98.7* | *97.1* | *77.8* | *59.6* | *92.2* | *85.0* | *60.9* | *89.4* | *90.8* | *90.0* | * |
| Margin scoring, *intersection*, threshold = 1.20 (*F1*) | 88.9 | 83.9 | 97.1 | 93.3 | 58.8 | 56.0 | 91.5 | 85.8 | 57.6 | 86.6 | 87.6 | 87.6 | * |
| *Precision* | *98.8* | *98.9* | *100* | *100* | *98.8* | *91.3* | *99.6* | *99.4* | *99.8* | *99.5* | *97.4* | *99.5* | * |
| *Recall* | *80.8* | *72.8* | *94.4* | *87.5* | *41.9* | *40.4* | *84.6* | *75.5* | *40.5* | *76.7* | *79.6* | *78.2* | * |
| Margin scoring, *intersection*, en-xx (*F1*) | 93.0 | 89.8 | 98.5 | 97.5 | 85.9 | * | 94.8 | 93.5 | * | * | 92.9 | 93.6 | * |
| *Precision* | *98.2* | *95.4* | *99.1* | *99.2* | *95.8* | * | *98.2* | *98.7* | * | * | *98.4* | *98.2* | * |
| *Recall* | *88.3* | *84.8* | *97.9* | *95.8* | *77.8* | * | *91.6* | *88.8* | * | * | *88.0* | *89.5* | * |
| Margin scoring, *intersection*, | | | | | | | | | | | | | |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| xx-en (*F1*) | **93.7** | **93.9** | 97.6 | **99.4** | **97.0** | * | **95.5** | **95.2** | * | * | **97.2** | **97.2** | * |
| *Precision* | *97.5* | *97.7* | *99.1* | *99.9* | *99.5* | * | *98.6* | *97.8* | * | * | *97.9* | *98.8* | * |
| *Recall* | *90.2* | *90.4* | *96.2* | *98.9* | *94.6* | * | *92.5* | *92.8* | * | * | *96.5* | *95.8* | * |
| Margin scoring, *intersection*, strict intersection (*F1*) | 92.0 | 89.9 | 97.4 | 97.6 | 79.9 | * | 93.7 | 91.2 | * | * | 91.3 | 92.7 | * |
| *Precision* | *99.2* | *99.5* | *99.6* | *100* | *100* | * | *99.7* | *100* | * | * | *98.4* | *99.6* | * |
| *Recall* | *85.7* | *81.9* | *95.3* | *95.3* | *66.5* | * | *88.5* | *83.9* | * | * | *85.2* | *86.7* | * |
| Margin scoring, *intersection*, pairwise intersection (*F1*) | 93.7 | 92.5 | **99.1*** | 98.8 | 94.0 | * | 95.4 | 93.6 | * | * | 95.7 | 95.9 | * |
| *Precision* | *98.6* | *97.9* | *99.6* | *100* | *100* | * | *98.7* | *99.2* | * | * | *98.5* | *99.1* | * |
| *Recall* | *89.3* | *87.6* | *98.7* | *97.6* | *88.7* | * | *92.3* | *88.6* | * | * | *93.0* | *92.8* | * |

Table 3: Tatoeba test set results for a subset of low-resource language pairs, broken down by the mining method used. These language pairs are ones *without* parallel data for the multilingual distillation process described in Reimers and Gurevych (2020) (cf. Table 10 in that paper). Note that LaBSE has training data for most of these languages. Descriptions of the various mining methods are found in Section 4.

| **Procedure** | **Average gain over baseline (best results only)** | **Average gain over baseline (all results)** | **Average gain over baseline (langs with transl. support)** | **Best results by resource capacity*** | **Average gain over baseline (by resource capacity)** |
|---|---|---|---|---|---|
| Margin scoring, *intersection*, no threshold | **+6.9** | **+5.2** | +3.6 | Level 0: 6 lang.<br>Level 1: 18 lang.<br>Level 2: 2 lang.<br>Level 3: 2 lang.<br>2†, 6‡ | Level 0: +7.2<br>Level 1: +**5.2**<br>Level 2: +1.8<br>Level 3: +3.4<br>Level 4: +1.0 |
| Margin scoring, *intersection*, threshold = 1.06 | +2.8 | * | * | Level 0: 1 lang.<br>Level 1: 1 lang.<br>Level 2: 1 lang. | Level 0: +6.1<br>Level 1: +4.3<br>Level 2: +1.6<br>Level 3: +2.9<br>Level 4: +0.6 |
| Margin scoring, *intersection*, threshold = 1.20 | * | * | * | * | Level 0: −3.8<br>Level 1: −4.3<br>Level 2: −6.8<br>Level 3: −5.2<br>Level 4: −3.2 |
| Margin scoring, *intersection*, en-xx | +6.5 | +2.4 | +2.4 | Level 0: 1 lang. | Level 0: +4.7<br>Level 1: +1.1<br>Level 2: +0.5<br>Level 3: +2.4<br>Level 4: +0.6 |
| Margin scoring, *intersection*, xx-en | +5.2 | +3.3 | +3.3 | Level 0: 1 lang.<br>Level 1: 7 lang.<br>Level 2: 2 lang.<br>Level 3: 7 lang.<br>Level 4: 1 lang. | Level 0: +3.9<br>Level 1: +2.8<br>Level 2: +0.1<br>Level 3: +**4.3**<br>Level 4: +**1.8** |
| Margin scoring, *intersection*, strict intersection | * | * | * | * | Level 0: +0.0<br>Level 1: −1.3<br>Level 2: −2.8<br>Level 3: −1.3<br>Level 4: +0.4 |
| Margin scoring, *intersection*, pairwise intersection | +4.6 | +4.0 | **+4.0** | Level 0: 2 lang.<br>Level 1: 3 lang.<br>Level 2: 2 lang.<br>Level 3: 1 lang. | Level 0: +**7.3**<br>Level 1: +3.9<br>Level 2: +**2.6**<br>Level 3: +4.0<br>Level 4: +1.0 |

* Using resource categorizations found here: rb.gy/psmfnz    † Extinct languages    ‡ Constructed languages

Table 4: Average gain (F1) over the baseline for each mining method on the low-resource subset of the Tatoeba test data given in Table 3, broken down by several categories. The baseline is the F1 achieved using raw cosine similarity with LaBSE. The "best results" for a given method are those results on which that method achieved superior results compared to all other methods. "All results" refers to all languages in the Tatoeba test set. The "resource capacity" refers to the amount of resources a language has available. The languages with "transl. support" are those which we translated before mining (applicable for the last four methods).

| **Procedure** | **Languages on which best result was achieved (ISO 639-2 code)**<br>**Gain over baseline**<br>**Resource capacity*** | | |
|---|---|---|---|
| Margin scoring, *intersection*, no threshold | ang +9.2 (1†)<br>ast +3.7 (1)<br>bre +4.2 (1)<br>cor +6.0 (1)<br>epo +0.6 (1‡)<br>ile +5.9 (1‡)<br>jav +7.8 (1)<br>lat +7.0 (3)<br>nds +7.8 (0)<br>orv +10.6 (?†)<br>swg +15.2 (?)<br>war +10.5 (0) | arq +11.0 (?)<br>awa +10.2 (0)<br>cbk +7.0 (1)<br>dsb +11.4 (0)<br>fao +4.3 (1)<br>ila +1.6 (1‡)<br>kdz +6.6 (0)<br>lfn +9.5 (?‡)<br>nov +7.2 (1‡)<br>pam +4.3 (?)<br>tel +0.8 (1)<br>wuu +4.5 (1) | arz +6.2 (3)<br>ber +3.8 (0)<br>cha +9.5 (1)<br>dtp +5.5 (?)<br>ido +4.2 (1‡)<br>isl +1.7 (2)<br>ksb +13.4 (1)<br>mal +0.4 (2)<br>occ +8.8 (1)<br>pms +11.9 (1)<br>tzl +8.0 (?) |
| Margin scoring, *intersection*, threshold = 1.06 | mal +0.4 (2) | mhr +7.1 (0) | tel +0.8 (1) |
| Margin scoring, *intersection*, threshold = 1.20 | * | | |
| Margin scoring, *intersection*, en-xx | fry +6.5 (0) | | |
| Margin scoring, *intersection*, xx-en | afr +1.6 (3)<br>bos +1 (3)<br>eus +1.8 (4)<br>kur +8.5 (0)<br>tgl +2.0 (3)<br>uzb+8.4 (3) | amh +1.7 (2)<br>ceb +15.2 (3)<br>gla +5.1 (1)<br>tam +3.0 (3)<br>tuk +17.0 (1)<br>xho +5.3 (2) | aze +1.5 (1)<br>cym +3.7 (1)<br>kaz +3.0 (3)<br>tat +6.0 (1)<br>uig+1.8 (1)<br>yid+6.2 (1) |
| Margin scoring, *intersection*, strict intersection | * | | |
| Margin scoring, *intersection*, pairwise intersection | bel +1.7 (0)<br>gsw +10.2 (?)<br>nno +1.9 (1) | ben +1.7 (3)<br>hsb +10.5 (0)<br>swa +6.9 (2) | gle +2.8 (2)<br>khm +4.6 (1)<br>tel +0.8 (1) |

* Using resource categorizations found here: rb.gy/psmfnz † Extinct languages ‡ Constructed languages

Table 5: LaBSE performances by mining method for each language in the Tatoeba test data. As in Tables 3 and 4, the baseline here is F1 (accuracy) obtained using raw cosine similarity with LaBSE. The resource capacity scores are on a 0-5 scale, with 5 indicating highest availability of resources.

# B Mining on the BUCC '17/18 Training Data

## B.1 Secondary Rule-based Retrieval Methods

We also experimented with many rule-based mining procedures on top of margin-based mining with LaBSE on the BUCC '17/18 English-French training data. That is, we performed the initial mining pass described in Algorithm 1, and then used rule based metrics to filter these sentence pairs. The measures we tried included:

- Length ratio
- Lexical overlap: Translate the source or target sentence, and then measure the BOW overlap
- Non-stopword lexical overlap
- Named entity overlap: Multiset named entity overlap using StanfordNER[16] (Finkel et al., 2005).
- Continuous constituent overlap: Using Kitaev et al. (2019)'s constituency parser[17] to compute longest continuous overlap (Butz and Wilson, 2002; Lukins, 2002)[18] between constituents of French sentence and word-by-word translated English sentence (Choe et al., 2020)[19].
- BLEU score: Similar to Bouamor and Sajjad (2018), computed BLEU score between English/French and translated French/English sentence. We experimented with NMT systems from Wu et al. (2016) and Tiedemann and Thottingal (2020)[20] for translation, as well as word-by-word translation from Choe et al. (2020) and Lample et al. (2018b)[21].
- METEOR score: Similar to BLEU, as speculated on by Bouamor and Sajjad (2018)
- Hybrid methods: Combinations of the rule-based metrics above, in addition to threshold-

[16] http://www.nltk.org/api/nltk.tag.html#module-nltk.tag.stanford

[17] https://github.com/nikitakit/self-attentive-parser

[18] https://github.com/timlukins/pylcs

[19] https://github.com/kakaobrain/word2word

[20] https://github.com/Helsinki-NLP/Opus-MT

[21] https://github.com/facebookresearch/MUSE

ing (including ensemble thresholding using LaBSE and LASER).

Unfortunately, none of these rule-based metrics were able to improve margin-based scoring in isolation in terms of F1, suggesting state-of-the-art similarity-based metrics have reached the level where they may not even be supplemented by rule-based metrics, including rule-ensembles, at least on high-resource language pairs.

### B.2 Confirming Flaws in Dataset

We confirm some of the problems with the BUCC dataset that others have pointed out. In particular, we corroborate Reimers and Gurevych (2019)'s observation—which they make on the EN-DE data, and us on EN-FR—that the BUCC data contains many ”false false positives”—that is, sentence pairs that are translations of each other but are not labeled as such. For instance, the following sentence pairs from the EN-FR train data are flagged as false positives:

**En** According to ecological economist Malte Faber, ecological economics is defined by its focus on nature, justice, and time.
**Fr** *Selon Malte Faber, l'économie écologique se définit par son intérêt pour la nature, la justice, et l'évolution au cours du temps.*
**En** Almost all parties have highly active student wings, and students have been elected to the Parliament.
**Fr** *Presque tous les partis ont des branches universitaires très actives, et des étudiants ont été élus au Parlement.*
**En** Many researchers at the time strongly supported the natural selection theory.
**Fr** *De nombreux chercheurs ont fortement soutenu la théorie de la sélection naturelle.*

Out of the first 100 sentence pairs flagged as false positives, we counted 72 that we would consider valid translations under rather strict criteria[22]. Extrapolating this to the rest of the false positives, we estimated the actual precision attainable using LaBSE with F1-based margin threshold optimization is around 97.5, in contrast with the 90.8 we originally recorded. While we don't repeat this procedure for false negatives, we notice that many of these so-called gold-standard translations suffer from coverage issues, which is why LaBSE and other similarity-based measures fail to catch them. Overall, we conclude that the actual F1 obtainable on the BUCC data with current methods is much closer to 100 than has been previously recorded (Reimers and Gurevych, 2020; Artetxe and Schwenk, 2019b), and we caution others against future leadboard-chasing on this benchmark, as we believe it may be ”conquered.”

[22] https://github.com/AlexJonesNLP/alt-bitexts/tree/main/BUCC_EN-FR_fp_fn