Low-Resource Machine Translation for Low-Resource Languages: Leveraging Comparable Data, Code-Switching and Compute Resources

Preprint · March 2021

CITATIONS
0

READS
61

5 authors, including:

Isidora Chara Tourni
Boston University
4 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Low-Resource Machine Translation for Low-Resource Languages: Leveraging Comparable Data, Code-Switching and Compute Resources

Garry Kuwanto*
Institut Teknologi Bandung
gkuwanto@students.itb.ac.id

Afra Feyza Akyürek*
Boston University
akyurek@bu.edu

Isidora Chara Tourni* **Siyang Li*** **Derry Wijaya**
Boston University
{isidora, siyangli, wijaya}@bu.edu

Abstract

We conduct an empirical study of unsupervised neural machine translation (NMT) for truly low resource languages, exploring the case when *both* parallel training data and compute resource are lacking, reflecting the reality of most of the world's languages and the researchers working on these languages. We propose a simple and scalable method to improve unsupervised NMT, showing how adding comparable data mined using a bilingual dictionary along with modest additional compute resource to train the model can significantly improve its performance. We also demonstrate how the use of the dictionary to code-switch monolingual data to create more comparable data can further improve performance. With this weak supervision, our best method achieves BLEU scores that improve over *supervised* results for English→Gujarati (+18.88), English→Kazakh (+5.84), and English→Somali (+1.16), showing the promise of weakly-supervised NMT for many low resource languages with modest compute resource in the world. To the best of our knowledge, our work is the first to quantitatively showcase the impact of different modest compute resource in low resource NMT.

1 Introduction

Approaches in Unsupervised NMT research have seen significant progress recently, facilitating translation with no parallel training data, using techniques such as back-translation and auto-encoding (Lample et al., 2018; Artetxe et al., 2017b; Sen et al., 2019; Artetxe et al., 2019; Sun et al., 2019). However, most works in this area have focused on simulating low-resource scenarios either for high-resource languages (e.g., German), for which unsupervised methods are redundant, or for medium-resource and similar to English ones (e.g., Romanian) rather than truly low-resource languages.

With large language models such as BERT and GPT-3 prevalent in NLP, one can achieve good low resource NMT performance by pretraining these on a large amount of data: hundreds of millions of sentences including monolingual and parallel data from related higher resource languages, on large compute resources: hundreds of large memory GPUs, for a long period of time: weeks (Liu et al., 2020). However, the assumption of abundant monolingual data, available parallel data, and compute resources, remains unsubstantiated when it comes to NMT for truly low-resource languages. The assumption that parallel data from related higher resource languages is available (Zoph et al., 2016; Nguyen and Chiang, 2017; Dabre et al., 2017; Kocmi and Bojar, 2018), such as Russian for Kazakh, only applies to some languages, making the provided approaches inapplicable to hundreds of *left-behind* languages (Joshi et al., 2020).

The lack of parallel data motivates unsupervised NMT. However, its best performance for low resource languages that are distant from English is spectacularly low (BLEU scores of less than 1), giving an impression of its *uselessness* (Kim et al., 2020; Liu et al., 2020). We show there are comparable data that can be easily mined to provide weak supervision to NMT, and that unsupervised NMT, combined with these comparable data is *not useless* for truly low resource languages. Together, they outperform their separate scores as well as *supervised* NMT performance from English to these languages, which is promising given the overwhelming need for automatic translation *from* English *to* other languages, as information in the former is ample compared to the latter (Nekoto et al., 2020).

In this work, among the low resource languages in MT for which there are test corpora, we focus our attention to three: Gujarati (*gu*) and Somali (*so*), the *scraping-by* languages for which *some* monolingual data are available, and Kazakh (*kk*) for which *more* monolingual data are available (Joshi et al.,

*Contributed equally

2020). These are distant languages that are diverse in terms of script, morphological complexity and word order from English (*en*) and for which unsupervised NMT has either never been explored (for so) or has been shown to perform poorly i.e., 0.6 and 0.8 BLEU for *en*→*gu* and *en*→*kk*.

We mine comparable sentences from Gujarati, Somali, and Kazakh Wikipedia, that are linked to their corresponding English articles. Wikipedia has pages in over 300 languages, both high and low resource, allowing our approaches presented here to scale to other languages. We empirically evaluate various lightweight methods to mine comparable sentences from these linked pages. We find that a simple method of using a bilingual lexicon dictionaries to mine sentences with lexical overlap results in comparable data that achieves the best NMT performance in all three languages. We also use the dictionary to code-switch monolingual data harnessing more comparable data.

Unless otherwise indicated, we assume limited access to compute resource and restrict our compute expenses by the average monthly income for the language speaking region. Our assumption of limited access to compute resource is realistic and driven by personal experiences and previous works that have observed how indeterminate access to a large number of GPUs is unrealistic beyond large industrial/academic environments (Ahmed and Wahed, 2020). We observe that without this constraint, adding even a small compute power from 1 GPU (16GB) to 4 GPUs (32GB) *and* training for a longer period of time can already improve NMT performance. We note how pivotal, even misguiding when not accounted for, compute resources can be in achieving compelling results. Furthermore, previous works have shown that increasing compute resources carries with it a high environmental cost (Strubell et al., 2019; Bender et al., 2021). To this end, we echo their suggestions and urge the NLP community to factor in the compute resources when reporting results. In addition, by using limited compute resources, we hope to shed light on the true state of low-resource NMT when both data *and* access to compute resource are lacking.

The contributions of our paper are manifold. We empirically evaluate lightweight methods for mining comparable sentences for training an NMT model. We explore different ways of leveraging monolingual data for NMT, using code switching. We introduce guidelines for training NMT for low-resource languages with no dependency on available parallel data from related languages, and quantify the effects of available computational resources. Finally, we achieve the state-of-the-art results without parallel data from English to these languages, with improvements up to 21 BLEU points from previous unsupervised results and up to 18 BLEU points from supervised results (Kim et al., 2020). We believe that the improvements presented here are remarkable, especially considering the fact that these languages have been treated as "hopeless cases" of unsupervised NMT in recent works. While making no language-specific assumptions, such as availability of a related high-resource language, or parallel data we provide state-of-the-art results, hoping our insights will motivate similar work in other languages.

2 Related Work

Despite the rapid progress in language technologies, research efforts in NLP have only incorporated about 6% of all 7000 world languages (Joshi et al., 2020). In their study, the authors develop a taxonomy of six categories depending on the data available in each language (labeled and unlabeled), and study extensively their resource disparities, and their representation in NLP conferences. Their analysis highlights that NLP methods and venues need to further focus on under-explored and typologically diverse languages. This motivates our choice of languages in this paper since Somali, Gujarati, and Kazakh are low-resource and typologically diverse languages that are also under-explored in unsupervised NMT.

It is important to note that aside from the growing gap in available data, a growing phenomenon in the AI community is the so-called "compute divide" between large firms, elite and non-elite universities, caused by the large computational requirements in GPU usage and the researchers' unequal access to computing resources (Ahmed and Wahed, 2020; Strubell et al., 2019). We believe that a discussion of compute requirement should always be factored in, especially in the study of low resource NMT; since for many of these languages the lack of compute resource, infrastructure, and time constraints can hinder communities in low resourced societies from working and publishing on these languages; and can render the use of techniques developed in high resourced societies inapplicable in this low resource setting (Joshi et al., 2020; Nekoto et al.,

	Wikipedia	WMT[1](2018/2019)	Leipzig Corpora[2](2016)	soWaC16[3](2016)	Total
gu	243K	531K	600K	-	1.36M
kk	1M	7.5M	1M	-	9.51M
so	32K	123K	-	1.83M	1.97M
en	843K(gu) 4.44M(kk) 654K(so)	517K(gu) 5.07M(kk) 1.32M(so)	-	-	1.36M(gu) 9.51M(kk) 1.97M(so)

Table 1: Monolingual data sources (with data size in number of sentences)

2020).

For low resource languages where parallel training data is little to none, unsupervised NMT can play a crucial role. 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 reason that this is because factors that are important for good unsupervised NMT such as linguistic similarity between source and target, and domain proximity along with size and quality of the monolingual corpora are hard to satisfy in the case of low resource languages. In this work we show that we can improve unsupervised performance by leveraging comparable data.

There is a large body of work in mining these *pseudo* parallel sentences (Munteanu et al., 2004; Munteanu and Marcu, 2006; Zweigenbaum et al., 2017; Guo et al., 2018; Grover and Mitra, 2017; Schwenk, 2018; Hangya et al., 2018; Hangya and Fraser, 2019; Wu et al., 2019b); yet most approaches have not been widely applied in the low-resource scenarios we aim to examine. Moreover, some of the recent works rely on supervised systems trained on parallel data(Schwenk et al., 2019a,b; Pourdamghani et al., 2019) and/or require extensive compute resource for iterative mining of comparable sentences (Tran et al., 2020), hence may not be applicable in low resource setting.

The most recent unsupervised bitext extraction method utilizes multilingual contextual embeddings retrieved from mBERT Keung et al. (2020); Devlin et al. (2019). Unlike these approaches that require expensive computation of contextual embedding similarity between sentences, we employ a simpler method of using bilingual dictionaries to mine sentences with lexical overlap. Our method also does not require a multilingual model containing the language but simply an existing dictionary which is available from various sources (Wiktionary, Kamholz et al. (2014); Thompson et al. (2019); Pavlick et al. (2014)) or can be induced from sentence alignment of comparable corpora such as Wikipedia titles (Ramesh and Sankaranarayanan, 2018; Pourdamghani et al., 2018; Wijaya et al., 2017).

3 Model

Since transformer-based architectures have proven successful in MT numerous times in the past (Barrault et al., 2019), for all of our experiments we use XLM (Conneau and Lample, 2019). 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. Somali and English for en-so). For both the LM pretraining and NMT model fine-tuning, unless otherwise noted, we follow the hyper-parameter settings suggested in the XLM repository[4]. For every language pair we extract a shared 60,000 subword vocabulary using Byte-Pair Encoding (BPE) (Sennrich et al., 2015b). Apart from MLM, Conneau and Lample (2019) introduced the Translation Language Model (TLM) for which we use utilize mined comparable data introduced in Section 4.2.

After pretraining the LM, we train a NMT model in an unsupervised or weakly-supervised (using comparable data) manner. We again follow the setup recommended in Conneau and Lample (2019), where both encoder and decoder are initialized using the same pretrained encoder block. For unsupervised NMT, we use back-translation (*BT*) and denoising auto-encoding (*AE*) losses (Lample et al., 2018), and the same monolingual data as in LM pretraining, elucidated in Section 4.1. Lastly, for our experiments that leverage 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.

[1] http://data.statmt.org/news-crawl/
[2] https://wortschatz.uni-leipzig.de/en/download/
[3] http://habit-project.eu/wiki/SetOfEthiopianWebCorpora
[4] http://github.com/facebookresearch/XLM

FIRST	**gu**	કવર્ધા ભારત દેશમાં આવેલા છત્તીસગઢ રાજ્યમાં આવેલું મહત્વનું નગર છે. (Kawardha is an important town in the Indian state of Chhattisgarh.)
	en	Kawardha is a city and a municipality in Kabirdham district in the Indian state of Chhattisgarh.
DATE	**gu**	આ મેચો 5 માર્ચ સોમવારથી શુક્રવાર 9 માર્ચ સુધી રમાઇ હતી. (These matches were played from Monday 5th March to Friday 9th March.)
	en	The matches were played from Monday 5 March until Friday 9 March.
DICT (PROJECTED)	**gu**	ગુજરાતી સાહિત્ય પરિષદના પૂર્વ પ્રમુખ તરીકેનું ગૌરવ પામનાર ડો. (Former President of Gujarati Sahitya Parishad, Dr.)
	en	He also served as the president of Gujarati Sahitya Parishad.
LENGTH	**kk**	Оттрау - Германия Федеративтік Республикасының Гессен жерінде орналасқан муниципалитет. (Otrau is a municipality in Hesse, Germany.)
	en	Ottrau is a community in the Schwalm-Eder-Kreis in Hesse, Germany.
LINK	**kk**	Ең ірі аралдары: Сааремаа, Хийумаа, Муху, Ворсми, т.б. (The largest islands: Saaremaa, Hiiumaa, Muhu, Vorsmi, etc.)
	en	Two of them are large enough to constitute separate counties: Saaremaa and Hiiumaa.
DICT (PROJECTED)	**kk**	Абдуррахман ибн Ауф басқа аты Абдел-Рахман ибн Гауф. (Abdur-Rahman ibn Auf is another name for 'Abd al-Rahman ibn Ghauf.)
	en	His name has also been transliterated as Abdel Rahman Ibn Auf.

Table 2: Sample comparable sentence pairs in Gujarati and Kazakh. Translations in parentheses are obtained using Google Translate which may be imperfect.

4 Dataset

4.1 Monolingual Data

The languages we study are Gujarati, Kazakh and Somali. They are spoken by 55M, 22M and 16M 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 candidates for our low-resource unsupervised NMT research.

Our monolingual data (see Table 1) are carefully chosen from the same domain of news data and from similar time periods (late 2010s) to mitigate domain discrepancy between source and target languages as per previous research (Kim et al., 2020). For English data, we use Wikipedia pages linked to *gu*, *kk* and *so* pages, respectively, combined with the randomly downsampled WMT NewsCrawl corpus so that target and source data are equal in size.

4.2 Comparable Data Mining

Our comparable data which is delineated in Table 3 comes from the linked Wikipedia pages in different languages obtained using the langlinks from Wikimedia dumps and six lightweight methods to extract comparable sentence pairs:

- DATE: Pairs with at least one identical date item (e.g. December).
- FIRST: The first sentence of both articles.
- LENGTH: Pairs in the first paragraph whose lengths are within ± 1 of each other.
- LINK: Pairs with at least two identical URLs.
- NUMBER: Pairs with at least two identical numbers.
- DICT(X): Pairs having at least 20% word overlap after word-by-word translation from source to English. Word overlap is defined as the ratio of matching words divided by the average number of words in the sentences.
- MERGED(X): This category includes all of the above with the corresponding DICT(X).

In order to curate dict(X), we first sample 10k most frequent words from Wikipedia dumps of *gu*, *kk* and *so* and obtain translations of these words from Amazon MTurk dictionary (Pavlick et al., 2014), dict(*MTurk*), and using Google Translate API, dict(*Google*). For Kazakh, because MTurk provides very few translations, we add entries from NorthEuraLex bilingual lexicon (Dellert et al., 2020). We also create a higher coverage dictionary based on dict(*Google*) by first training monolingual word embeddings for each language using fastText's skipgram model (Bojanowski et al., 2017) on the monolingual data of the language (Table 1); then learn a linear mapping between the source and target word embeddings, using dict(*Google*) as seed translations with *MUSE* (Conneau et al., 2017). Based on this mapping, we find translations of up to 200k most frequent words from each language monolingual data by projecting the source embedding (*gu*, *kk*, and *so*) to the target language embedding (*en*) and taking as translation the target word that has the highest cosine similarity based Cross-Domain Similarity Local Scaling (CSLS) metric (Conneau et al., 2017), which adjusts cosine similarity values of a word based on the density of the area where its embedding lies. We call this higher coverage dictionary dict(*Projected*).

4.3 Data Augmentation Using Code-Switching

To create more comparable data for training, similar to the idea of back-translation to automatically

Method	# sentences			BLEU					
	gu	kk	so	en-gu	gu-en	en-kk	kk-en	en-so	so-en
DATE	3758	3079	618	0.11	1.47	0.09	0.00	0.00	0.00
DICT(MTURK)	44492	5832	19754	6.17	1.82	0.17	0.30	0.38	0.00
DICT(GOOGLE)	65974	53211	9927	6.39	1.98	0.55	**1.34**	0.73	0.95
DICT(PROJECTED)	71424	58961	9880	5.86	1.79	0.66	1.30	**1.17**	**1.13**
FIRST	8671	103002	4118	0.00	0.00	0.00	0.23	0.52	1.03
LENGTH	9487	52553	1870	1.11	0.21	0.00	0.17	0.25	0.00
LINK	3795	18294	683	1.27	0.00	0.21	0.31	0.00	0.00
NUMBER	926	2604	254	0.85	0.00	0.00	0.00	0.13	0.17
MERGED(MTURK)	64904	156871	25101	6.08	1.90	0.32	0.48	0.00	0.53
MERGED(GOOGLE)	85032	196382	15361	6.73	2.14	0.55	1.28	1.05	0.76
MERGED(PROJECTED)	90229	202147	15377	**7.15**	**2.56**	**0.69**	1.19	0.52	1.03

Table 3: Different methods of mining comparable sentences used, number of comparable sentence pairs extracted with each technique and BLEU scores of models trained using only *MT* objective with these pairs. There is no LM pretraining. Best results are highlighted.

translate the monolingual text to and from the target language (Sennrich et al., 2015a), we use our dict(*Projected*) dictionary to automatically translate monolingual Wikipedia sentences to and from the target language, creating *real* (i.e., monolingual) source to *pseudo* (i.e., dictionary-translated) target sentence pairs and *pseudo* source to *real* target sentence pairs, respectively. Since the coverage of our dictionary is limited, there are words in the monolingual sentence that it cannot translate. In those cases, we leave the words as they are, resulting in an imperfect, *code-switched* translation of the monolingual sentence. Previous works on text classification has shown that fine-tuning multilingual models such as multilingual BERT on code-switched data can improve performance on few-shot and zero-shot classification tasks (Akyürek et al., 2020; Qin et al., 2020) ranging from frame classification (Liu et al., 2019a) to natural language inference (Conneau et al., 2018), sentiment classification (Barnes et al., 2018), document classification (Schwenk and Li, 2018), dialogue state tracking (Mrkšić et al., 2017), and spoken language understanding (Schuster et al., 2019). We include code-switched sentences that have at least 20% of words translated from their original sentences to augment our training. Despite its imperfect nature, we believe that such (code-switched, real) sentence pairs can provide some weak supervision to the unsupervised NMT model to learn to translate. In our experiments that leverage code-switched sentences, we follow the unsupervised NMT training step i.e., *BT+AE* with $BT{+}MT_{cs}$ and then *BT+MT*, where MT_{cs} stands for supervised MT objective, for which we use the (code-switched, real) sentence pairs, and *MT* stands for supervised MT objective for which we use the mined data as before.

5 Experiments and Results

5.1 Experimental Setup

We use WMT 2019 news dataset for evaluation of Gujarati ↔ English and Kazakh ↔ English. We use DARPA's LORELEI (Tracey et al., 2019) validation and test data sets for Somali.

Because we are interested in simulating a limited compute resource setting, instead of training our language models for prolonged times, we take into account the average monthly income[1] of each region the language is primarily spoken in (Gujarat for Gujarati, Kazakhstan for Kazakh and Somalia for Somali) and use Amazon AWS EC2 rate[2] as an estimate on how long we should train. Our calculation yields 40, 60 and 72 hours of training time on 1 GPU (16GB) for Gujarati, Somali and Kazakh, respectively. While we train our language models only up to this calculated time, we do not specifically enforce training time limits for NMT training, which stops only if BLEU score on validation does not improve after 10 epochs. We observe that NMT training lasts less than 24 hours in average.

5.2 Which type of comparable data is the best?

In order to empirically evaluate which method provides the best quality comparable data, we train NMTs from a randomly initialized LM using only the MT objective, hence no expensive pre-training on monolingual data. This is an efficient way to see how the comparable sentences fare when used

[1] https://www.numbeo.com/cost-of-living/
[2] https://calculator.aws/#/

in other training scenarios. In Table 3 we see the numbers of sentence pairs mined using a particular method, and the BLEU scores obtained using only these extracted corpora in NMT. We observe that for Kazakh-English and Somali-English, sentence pairs from the dict(*Projected*) yield the best BLEU score when both directions i.e., X-*en* and *en*-X are considered together. Moreover, because all types of sentence pairs in English-Gujarati, except (perhaps surprisingly) FIRST are generally better than other language pairs, merging all of the mined corpora provides a better result.

5.3 Leveraging Comparable Data to Boost Unsupervised MT

Kim et al. (2020) pinpoint the impracticality of unsupervised NMT due to the large monolingual corpora required and the difficulty of satisfying other, domain and linguistic similarity constraints for low-resource languages such as Kazakh and Gujarati. As seen in Table Table 4, they report remarkably low-scores for these languages in both directions, and conclude that only(!) 50k paired sequences will suffice to surpass the performance of unsupervised NMT. We find the call to create 50k parallel sentences for languages which do not even have sizable annotated data for testing to be beyond reach in the near future. Therefore, in this work, we show that by leveraging inexpensive resources such as lexical translations, it is possible to achieve satisfactory results which are on par, if not better, than supervised NMT.

In Table 4, we provide the BLEU scores using the best comparable data (from Table 3) to train weakly-supervised NMT models. We observe significant improvements in all three languages and both directions. In addition, in Table 4 we also provide the BLEU scores for the TLM model, which makes use of comparable data from the beginning. While the TLM model is still much better than using no mined sequences, it is not as good as adding comparable data after training an unsupervised NMT, especially for Kazakh and Somali models. This signals that adding comparable data, which are not of great quality, too early into training may not be the optimal practice. Letting the models first learn from monolingual data for simpler language features and possibly grammar, and then adding comparable data later on is potentially a more efficient way to utilize mined data.

5.4 Accounting for Compute Resources

Previous research demonstrates that LM pretraining significantly improves performance in many downstream natural language processing tasks, including NMT (Lample et al., 2018). Moreover, LMs benefit from large batch sizes during pretraining (Liu et al., 2019b). In this section, we attempt to quantify how increased availability of compute resources (practically enabling large batch sizes) affect the language model perplexity as well as translation performance, even when data is scarce. Note that using gradient accumulation one can mimic larger number of GPUs. However, in this case, replicating a setup of four GPUs with only one can quickly become prohibitive time-wise, especially considering the already prolonged times required to train transformer-based LMs (Devlin et al., 2019).

In Table 5 we provide LM perplexities for different numbers of GPUs for both Kazakh and Somali, and BLEU scores in Table 6. We use NVIDIA V100 32GB GPUs, setting the batch size to 64 for LM pretraining, and the tokens per batch to 3k for MT fine-tuning, per GPU. We keep all other parameters the same for simplicity. The results support the hypothesis that enhanced compute resource bears significant potential to boost LM quality. We observe that the perplexity consistently drops as the number of GPUs increases. In our experiments we found a strong correlation between translation performance and a better LM pretraining.

Notably, without any hyper-parameter tuning other than increasing the number of GPUs, the unsupervised translation scores improve consistently in general, if not dramatically (Table 6). For Somali, BLEU scores almost double in both *en-so* and *so-en* directions, rising from 8.35 to 14.76 for *en-so* and from 8.02 to 14.83 for *so-en*. Kazakh proved to be a more challenging case for low-resource NMT, nonetheless, simply utilizing more resources even for Kazakh results in consistent improvements in Table 6 in any translation direction.

For weakly-supervised NMT, *MT* objective uses comparable data mined with dict(*Projected*). Except for one case, we observe the same pattern even after the *MT* step in Table 6 where the scores are already significantly higher compared to the unsupervised case. We think the drop in *en-kk* direction in the 4 GPUs case could be remedied with adjusting the learning rate (Puri et al., 2018), which we avoided for easy interpretation of the results. In these experiments, we simply highlight how a priv-

Method	en-gu	gu-en	en-kk	kk-en	en-so	so-en
Other methods						
Multilingual, Supervised, Google Translate	31.4	26.2	23.1	28.9	22.7	27.7
Multilingual, Supervised, mBART25[2]	0.1	0.3	2.5	7.4	-	-
Bilingual, Unsupervised[1]	0.6*	0.6*	0.8*	2.0*	-	-
Bilingual, Supervised	3.5*[1]	9.9*[1]	2.4*[1]	10.3*[1]	-	25.36[3]
Bilingual, Semi-supervised[1]	4.0*	14.2*	3.1*	12.5*	-	-
Ours						
Bilingual, Unsupervised (*MLM + BT + AE*)	0.44	0.31	1.11	1.60	8.53	7.99
Bilingual, Weakly-supervised (*MLM + TLM*) + (*BT + MT*)	18.05	14.51	4.06	5.65	9.06	10.33
Bilingual, Weakly-supervised *MLM* + (*BT + AE*)	0.44	0.31	1.11	1.6	8.53	7.99
└ (*BT + MT*)	18.13	14.36	5.86	8.02	14.92	14.31
└ 4 GPUs	22.38	21.04	7.26	10.63	16.96	15.83

Table 4: BLEU scores for previous supervised and unsupervised results from [1]Kim et al. (2020) (*report SacreBLEU), [2]Liu et al. (2020), and [3]Liu and Kirchhoff (2018) and our weakly-supervised models which leverage comparable data mined using lexical translations. Test and validation sets are from WMT19 for Gujarati and Kazakh and from Tracey et al. (2019) for Somali. *MLM*, *AE*, *BT* and *MT* stand for Masked Language Model, Auto-Encoding loss, Back Translation loss and Machine Translation loss, respectively. Bolded means best score. *MT* and *TLM* objectives use the best comparable data as paired data according to Table 3. Our models, except for the last row, use a single 32GB GPU. Unlike previous results, we use the standard XLM parameters and do not conduct extensive hyperparameter tuning, which can further improve our performance. For completeness sake, Lakew et al. (2020) reports supervised results for Somali on a pooled test set from sources that doesn't include the test corpus studied here. They achieve 9.16 and 13.38 for en-so and so-en, respectively.

GPUs	en-kk		en-so	
	en	kk	en	so
1	15.17	15.37	16.65	10.41
2	14.28	14.46	14.62	9.07
4	**12.70**	**12.16**	**12.44**	**7.78**

Table 5: Test set perplexities of the bilingual LMs trained using different number of GPUs while everything else is fixed. Best results are highlighted.

GPUs	en-kk	kk-en	en-so	so-en
Unsupervised NMT: MLM + (BT+AE)				
1	0.58	1.32	8.35	8.02
2	1.02	1.73	12.05	11.35
4	**2.91**	**3.86**	**14.76**	**14.83**
Weakly-supervised NMT: MLM + (BT+AE) + (BT+MT)				
1	7.30	10.23	14.79	14.93
2	**8.24**	10.60	16.11	15.72
4	7.26	**10.63**	**16.96**	**15.83**

Table 6: BLEU scores using different number of 32GB GPUs. *MT* objective uses the best comparable data according to Table 3. For both languages best comparable data was yielded using dict(*Projected*). Tokens per batch set at 3k which is different than our basic 1 GPU (16GB) setup. Best results are highlighted.

ileged access to compute resources can drastically affect translation performance without resorting to much engineering. Likewise, as LM pretraining has become a common practice for many language tasks (Yang et al., 2019; Bao et al., 2019), we strongly believe that accounting for the compute resources is essential when creating benchmarks.

5.5 Benchmark Results

Surprisingly, regardless of unsupervised MT being in the spotlight for a few years now (Conneau et al., 2017; Wu et al., 2019a; Song et al., 2019; Pourdamghani et al., 2019), very few works studied the low-resource languages we examine in this paper. In Table 4, we compare the best results we achieved in this paper to the previous works (Liu and Kirchhoff, 2018; Kim et al., 2020), breaking down the specific components that contributed to the results, including the compute power. Except for two cases, our comparable-data-powered model remarkably outperforms previous results, *including* supervised ones which utilize tens of thousands of parallel sequences, if not more. Improvements gained in this paper over the latest reported unsupervised method (Kim et al., 2020) range from around at least 6.5 BLEU in *en-kk* to 21.8 in *en-gu*! Moreover, it is critical to note that we provide the first unsupervised translation benchmark for English translation to and from Somali, as, to our knowledge, all results reported in the literature so far conduct MT in a supervised way.

5.6 When life hits you hard, hit back harder!

Among the three languages, Kazakh suffers the most from relatively poor crosslinguality, which is reflected in Table 4. We believe that this may be due to its agglutinative nature, where each of its root words can produce hundreds or thousands of variant word forms (Tolegen et al., 2020). Hence, Kazakh may require many more training examples to generalize compared to other, less morphologically complex, languages. We observe that adding more comparable data in the form of code-switched sentences as discussed in Section 4.3

Somali Sentence	Warbaahinta Ruushka ayaa sheegtay in diyaarad ay leedahay shirkadda diyaaradaha Malaysia oo ka duushay Amsterdam kuna socotay Kuala Lumpur ay ku burburtay bariga Ukraine.
Ours *MLM* + (*BT* + *AE*)	The **Russian** government said **that a** company from Malaysia was on board at **Amsterdam** kuna dhacday **Kuala Lumpur** in a burburtay in bariga Ukraine.
└ (*BT* + *MT*)	**Russian media** said a Malaysia **Airlines** jet that landed in **Amsterdam** at **Kuala Lumpur crashed in eastern Ukraine**.
└ 4 GPUs	**Russian media** said **a Malaysian airline** flight **from Amsterdam to Kuala Lumpur crashed in eastern Ukraine**.
Reference Translation	Russian media reported that a Malaysian Airlines plane flying from Amsterdam to Kuala Lumpur has crashed in Eastern Ukraine.

Table 7: Sample translations from Somali → English by different models.

	en-kk	**kk-en**
Unsupervised	1.11	1.6
Weakly-supervised	5.86	8.02
Code-switched	**7.26**	**9.64**

Table 8: BLEU scores for en-kk and kk-en translation from different NMT models, using one 16GB GPU.

can help, adding a further 1.4 and 1.62 BLEU points for translation to and from Kazakh under the constrained compute resource setting (Table 8). In the future, we believe that for morphologically complex languages such as Kazakh, we need to conduct morphological segmentation preprocessing (Sánchez-Cartagena et al., 2019) to segment words into subword units, which has been shown to outperform BPE for highly inflected languages (Sánchez-Cartagena and Toral, 2016; Huck et al., 2017; Sánchez-Cartagena et al., 2018).

6 Discussion and Future Work

The different scenarios studied have various effects when it comes to the quality of the output translations. Unlike Somali, Gujarati does not benefit much from unsupervised training, potentially due to small monolingual data along with a different script. Nevertheless, mined comparable data for Gujarati is of high quality (see Table 3). Hence, adding these to *BT+AE* model results in BLEU scores of 18.13 and 14.36, which is further boosted with more computational power. Moreover, our Somali model using mined comparable data surpasses the supervised score in *en-so* direction. Table 7 lists sample translations for Somali using different models, showing how translation gets increasingly better with addition of mined data and compute resource. In general, we observe that Kazakh BLEU percentages are lower compared to those of Somali and Gujarati. We believe this is due to the morphologically more complex structure of Kazakh (Briakou and Carpuat, 2019). Yet, the introduction of comparable data substantially increases the BLEU scores for Kazakh, by 4.7 and 6.4 points, even without additional compute power. Further, in Table 8 we introduce code-switched training even before comparable data is used. In addition to back-translation, the code-switched words act as anchors that help align the embedding space of the two languages, yielding a state-of-the-art performance for Kazakh without any parallel data and with mediocre compute resources. Nonetheless, since there is still room for improvement, we deem the need to create more parallel resources for languages such as Kazakh (Joshi et al., 2020) as well as other low-resource and/or highly complex languages with millions of speakers crucial.

Unlike the canonical languages studied in unsupervised NMT i.e. German, French, and Romanian, the core importance of our work lies in that the languages we examine are truly low-resource as well as linguistically, geographically and morphologically diverse from English. Given that in none of our experiments do we assume a related high-resource language to aid in translation, or use any parallel sequences, we set the ground for similar analyses and extension of our approaches to other low-resource languages. We also show that translation scores can almost be doubled by enhancing the compute power used in the experiments. Hence, we urge the community to factor in this significant parameter when reporting results.

It is also worth noting that, for some languages, even lexical translations may be difficult to obtain. Therefore, future work can explore using unsupervised methods to obtain lexical translations in order to yield comparable sequences for MT. Several previous works study harnessing word vectors trained on monolingual corpora in unsupervised manner (Conneau et al., 2017; Irvine and Callison-Burch, 2017), with small bilingual seeds (Artetxe et al., 2017a), or with other modalities or knowledge resources (Hewitt et al., 2018; Wijaya et al., 2017). Others have explored extracting lexical translations from word alignment of compa-

rable corpora such as Bible, Universal Declaration of Human Rights, or Wikipedia Titles (Ramesh and Sankaranarayanan, 2018; Pourdamghani et al., 2018) as Wikipedia is a rich lexical semantic resource (Zesch et al., 2007; Wijaya et al., 2015). We are also interested in exploring other sources for mining comparable sentences outside of Wikipedia such as international news sites (Voice of America, etc.)

7 Conclusion

In this work, we explore ways to mine, create, and utilize comparable data to improve performance of unsupervised NMT, which can scale to many low resource languages; while shedding light at another relevant but often overlooked factor that is compute resource. Without using any parallel data aside from bilingual lexicons and without assuming any availability of similar, higher resource languages, we lift up NMT results for truly low-resource languages from disconcerting scores (i.e. fractional BLEU points) to two-digit BLEU scores. We hope this work will set the ground for application of our methods to other languages, as well as further exploration of similar approaches.

References

Nur Ahmed and Muntasir Wahed. 2020. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017a. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017b. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. *arXiv preprint arXiv:1805.09016*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Eleftheria Briakou and Marine Carpuat. 2019. The university of maryland's kazakh-english neural machine translation system at wmt19. In *Proceedings of the Fourth Conference on Machine Translation*

(Volume 2: Shared Task Papers, Day 1), pages 134–140.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, et al. 2020. Northeuralex: A wide-coverage lexical database of northern eurasia. *Language resources and evaluation*, 54(1):273–301.

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.

Jeenu Grover and Pabitra Mitra. 2017. Bilingual word embeddings with bucketed cnn for parallel sentence extraction. In *Proceedings of ACL 2017, Student Research Workshop*, pages 11–16.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, et al. 2018. Effective parallel corpus mining using bilingual sentence embeddings. *arXiv preprint arXiv:1807.11906*.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised parallel sentence extraction from comparable corpora. In *Proc. IWSLT*.

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.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67.

Ann Irvine and Chris Callison-Burch. 2017. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310.

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.

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *LREC*, pages 3145–3150.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A Smith. 2020. Unsupervised bitext mining and translation via self-trained contextual embeddings. *arXiv preprint arXiv:2010.07761*.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Surafel M Lakew, Matteo Negri, and Marco Turchi. 2020. Low resource neural machine translation: A benchmark for five african languages. *arXiv preprint arXiv:2003.14402*.

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

Angli Liu and Katrin Kirchhoff. 2018. Context models for oov word translation in low-resource languages. *arXiv preprint arXiv:1801.08660*.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019a. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? *arXiv preprint arXiv:2004.05516*.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324.

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.

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.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Nima Pourdamghani, Nada Aldarrab, Marjan Ghazvininejad, Kevin Knight, and Jonathan May. 2019. Translating translationese: A two-step approach to unsupervised machine translation. *arXiv preprint arXiv:1906.05683*.

Nima Pourdamghani, Marjan Ghazvininejad, and Kevin Knight. 2018. Using word vectors to improve word alignments for low resource machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 524–528.

Raul Puri, Robert Kirby, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Large scale language modeling: Converging on 40gb of text in four hours. In *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, pages 290–297. IEEE.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402*.

Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. 2018. Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. *arXiv preprint arXiv:1806.09652*.

Víctor M Sánchez-Cartagena, Marta Bañón, Sergio Ortiz Rojas, and Gema Ramírez-Sánchez. 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.

Víctor M Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2019. The universitat d'alacant submissions to the english-to-kazakh news translation task at wmt 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 356–363.

Víctor M Sánchez-Cartagena and Antonio Toral. 2016. Abu-matran at wmt 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 362–370.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492*.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. *arXiv preprint arXiv:1805.09822*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. *arXiv preprint arXiv:1805.09821*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised nmt using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint 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.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Unsupervised bilingual word embedding agreement for unsupervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1235–1245.

Brian Thompson, Rebecca Knowles, Xuan Zhang, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Hablex: Human annotated bilingual lexicons for experiments in machine translation. 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 1382–1387.

Gulmira Tolegen, Alymzhan Toleu, Orken Mamyrbayev, and Rustam Mussabayev. 2020. Neural named entity recognition for kazakh. *arXiv preprint arXiv:2007.13626*.

Jennifer Tracey, Stephanie Strassel, Ann Bies, Zhiyi Song, Michael Arrigo, Kira Griffitt, Dana Delgado, Dave Graff, Seth Kulick, Justin Mott, et al. 2019. Corpus building for low resource languages in the darpa lorelei program. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 48–55.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. *Advances in Neural Information Processing Systems*, 33.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463.

Derry Tanti Wijaya, Ndapandula Nakashole, and Tom Mitchell. 2015. "a spousal relation begins with a deletion of engage and ends with an addition of divorce": Learning state changing verbs from wikipedia revision history. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 518–523.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019a. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. *arXiv preprint arXiv:1904.02331*.

Lijun Wu, Jinhua Zhu, Di He, Fei Gao, QIN Tao, Jianhuang Lai, and Tie-Yan Liu. 2019b. 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 4366–4375.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007. Analyzing and accessing wikipedia as a lexical semantic resource. *Data Structures for Linguistic Resources and Applications*, 197205.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

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.