“Wikily” Supervised Neural Translation Tailored to Cross-Lingual Tasks

Mohammad Sadegh Rasooli[1]* **Chris Callison-Burch**[2] **Derry Tanti Wijaya**[3]
[1]Microsoft
[2]Department of Computer and Information Science, University of Pennsylvania
[3]Department of Computer Science, Boston University
mrasooli@microsoft.com, ccb@seas.upenn.edu, wijaya@bu.edu

Abstract

We present a simple but effective approach for leveraging Wikipedia for neural machine translation as well as cross-lingual tasks of image captioning and dependency parsing without using any direct supervision from external parallel data or supervised models in the target language. We show that first sentences and titles of linked Wikipedia pages, as well as cross-lingual image captions, are strong signals for a seed parallel data to extract bilingual dictionaries and cross-lingual word embeddings for mining parallel text from Wikipedia. Our final model achieves high BLEU scores that are close to or sometimes higher than strong *supervised* baselines in low-resource languages; e.g. supervised BLEU of 4.0 versus 12.1 from our model in English-to-Kazakh. Moreover, we tailor our *“wikily” supervised* translation models to unsupervised image captioning, and cross-lingual dependency parser transfer. In image captioning, we train a multi-tasking machine translation and image captioning pipeline for Arabic and English from which the Arabic training data is a translated version of the English captioning data, using our wikily-supervised translation models. Our captioning results on Arabic are slightly *better* than that of its supervised model. In dependency parsing, we translate a large amount of monolingual text, and use it as artificial training data in an *annotation projection* framework. We show that our model outperforms recent work on cross-lingual transfer of dependency parsers.

1 Introduction

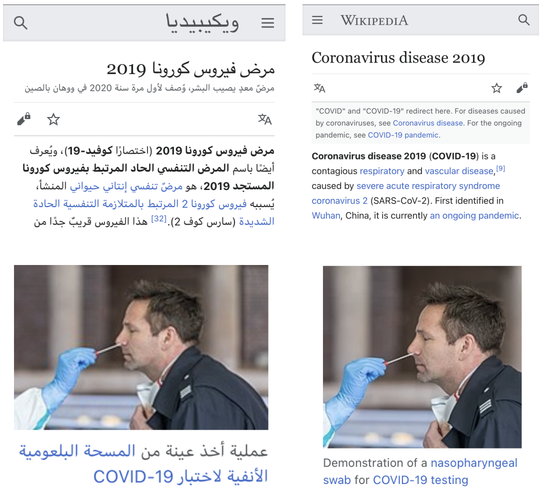
Developing machine translation models without using bilingual parallel text is an intriguing research problem with real applications: obtaining a large volume of parallel text for many languages is hard if not impossible. Moreover, translation models could be used in downstream cross-lingual tasks in which annotated data does not exist for some languages. There has recently been a great deal of interest in unsupervised neural machine translation (e.g. Artetxe et al. (2018a); Lample et al. (2018a,c); Conneau and Lample (2019); Song et al. (2019a); Kim et al. (2020); Tae et al. (2020)). Unsupervised neural machine translation models often perform nearly as well as supervised models when translating between similar languages, but they fail to perform well in low-resource or distant languages (Kim et al., 2020) or out-of-domain monolingual data (Marchisio et al., 2020). In practice, the highest need for unsupervised models is to expand beyond high resource, similar European language pairs.

There are two key goals in this paper: Our first goal is developing accurate translation models for low-resource distant languages *without* any supervision from a supervised model or gold-standard parallel data. Our second goal is to show that our machine translation models can be directly tailored to downstream natural language processing tasks. In this paper, we showcase our claim in cross-lingual image captioning and cross-lingual transfer of dependency parsers, but this idea is applicable to a wide variety of tasks.

We present a fast and accurate approach for learning translation models using Wikipedia. Unlike *unsupervised machine translation* that solely relies on raw monolingual data, we believe that we should not neglect the availability of incidental supervisions from online resources such as Wikipedia. Wikipedia contains articles in nearly 300 languages and more languages might be added in the future, including indigenous languages and dialects of different regions in the world. Different from similar recent work (Schwenk et al., 2019a), we do not rely on any supervision from supervised translation models. Instead, we leverage the fact that many first sentences in linked Wikipedia pages are rough

*Research was conducted at The University of Pennsylvania.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1655–1670
November 7–11, 2021.



ويكيبيديا

مرض فيروس كورونا 2019

مرضٌ معدٍ يصيب البشر، وُصف لأول مرة سنة 2020 في ووهان بالصين

مرض فيروس كورونا 2019 (اختصارًا **كوفيد-19**)، ويُعرف
أيضًا باسم **المرض التنفسي الحاد المرتبط بفيروس كورونا**
المستجد 2019، هو مرضٌ تنفسي إنتاني حيواني المنشأ،
يُسببه فيروس كورونا 2 المرتبط بالمتلازمة التنفسية الحادة
الشديدة (سارس كوف 2).[32] هذا الفيروس قريبٌ جدًا من

عملية أخذ عينة من المسحة البلعومية
الأنفية لاختبار COVID-19

Wikipedia

Coronavirus disease 2019

"COVID" and "COVID-19" redirect here. For diseases caused by coronaviruses, see Coronavirus disease. For the ongoing pandemic, see COVID-19 pandemic.

Coronavirus disease 2019 (**COVID-19**) is a contagious respiratory and vascular disease,[9] caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). First identified in Wuhan, China, it is currently an ongoing pandemic.

Demonstration of a nasopharyngeal swab for COVID-19 testing

Figure 1: A pair of Wikipedia documents in Arabic and English, along with a same image with two captions.

translations, and furthermore, many captions of the same images are similar sentences, sometimes translations. Figure 1 shows a real example of a pair of linked Wikipedia pages in Arabic and English in which the titles, first sentences, and also the image captions are rough translations of each other. Our method learns a seed bilingual dictionary from a small collection of first sentence pairs, titles and captions, and then learns cross-lingual word embeddings. We make use of cross-lingual word embeddings to extract parallel sentences from Wikipedia. Our experiments show that our approach improves over strong unsupervised translation models for low-resource languages: we improve the BLEU score of English→Gujarati from 0.6 to 15.2, and English→Kazakh from 0.8 to 12.1.

In the realm of downstream tasks, we show that we can easily use our translation models to generate high-quality translations of MS-COCO (Chen et al., 2015) and Flickr (Hodosh et al., 2013) datasets, and train a cross-lingual image captioning model in a multi-task pipeline paired with machine translation in which the model is initialized by the parameters from our translation model. Our results on Arabic captioning show a BLEU score of 5.72 that is slightly better than a supervised captioning model with a BLEU score of 5.22. As another task, in *dependency parsing*, we first translate a large amount of monolingual data using our translation models and then apply transfer using the *annotation projection* method (Yarowsky et al., 2001; Hwa et al., 2005). Our results show that our approach performs similarly compared to using gold-standard parallel text in high-resource scenarios, and significantly better in low-resource languages.

A summary of our contribution is as follows: 1) We propose a simple, fast and effective approach towards using the Wikipedia monolingual data for machine translation without any explicit supervision. Our mining algorithm easily scales on large comparable data using limited computational resources. We achieve very high BLEU scores for distant languages, especially those in which current unsupervised methods perform very poorly. 2) We propose novel methods for leveraging our current translation models in image captioning. We show that how a combination of translating caption training data, and multi-task learning with English captioning as well as translation improves the performance. Our results on Arabic shows results slightly superior to that of a supervised captioning model trained on gold-standard datasets. 3) We propose a novel modification to the annotation projection method to be able to leverage our translation models. Our results on dependency parsing performs better than previous work in most cases, and performs similarly to using gold-standard parallel datasets.

Our translation and captioning code and models are publicly available online[1].

2 Background

Supervised neural machine translation Supervised machine translation uses a parallel text $\mathcal{P} = \{(s_i, t_i)\}_{i=1}^{n}$ in which each sentence $s_i \in l_1$ is a translation of $t_i \in l_2$. *Neural machine translation* uses sequence-to-sequence models with attention (Cho et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) for which the likelihood of training data is maximized by maximizing the log-likelihood of predicting each target word given its previous predicted words and source sequence:

$$\mathcal{L}(\mathcal{P}) = \sum_{i=1}^{n} \sum_{j=1}^{|t_i|} \log p(t_{i,j} | t_{i,k<j}, s_i; \theta)$$

where θ is a collection of parameters to be learned.

Unsupervised neural machine translation Unsupervised neural machine translation does not have access to any parallel data. Instead, it tailors monolingual datasets $\mathcal{M}_{l_1}$ and $\mathcal{M}_{l_2}$ for learning multilingual language models. These language

[1] Our code: https://github.com/rasoolims/ImageTranslate. Our modification to Stanza for training on partially projected trees: https://github.com/rasoolims/stanza.

models usually mask parts of every input sentence, and try to uncover the masked words (Devlin et al., 2019). The monolingual language models are used along with iterative back-translation (Hoang et al., 2018) to learn unsupervised translation. An input sentence s is translated to t' using current model θ, then the model assumes that (t', s) is a gold-standard translation, and uses the same training objective as of supervised translation.

Dependency parsing Dependency parsing algorithms capture the best scoring dependency trees for sentences among an exponential number of possible dependency trees. A valid dependency tree for a sentence $s = s_1, \ldots, s_n$ assigns heads h_i for each for word s_i where $1 \leq i \leq n, 0 \leq h_i \leq n$ and $h_i \neq i$. The zeroth word represents a dummy root token as an indicator for the root of the sentence. For more details about efficient parsing algorithms, we encourage the reader to see Kübler et al. (2009).

Annotation projection Annotation projection is an effective method for transferring supervised annotation from a rich-resource language to a low-resource language through translated text (Yarowsky et al., 2001). Having a parallel data $\mathcal{P} = \{(s_i, t_i)\}_{i=1}^{n}$, and supervised source annotations for source sentences s_i, we transfer those annotations through word translation links $0 \leq a_i^{(j)} \leq |t_i|$ for $1 \leq j \leq |s_i|$ where $a_i^{(j)} = 0$ shows a `null` alignment. The alignment links are learned in an unsupervised fashion using unsupervised word alignment algorithms (Och and Ney, 2003a). In dependency parsing, if $h_i = j$ and $a^{(j)} = k$ and $a^{(i)} = m$, we project a dependency $k \rightarrow m$ (i.e. $h_m = k$) to the target side. Previous work (Rasooli and Collins, 2017, 2019) has shown that annotation projection only works when a large amount of translation data exists. In the absence of parallel data, we create artificial parallel data using our translation models. Figure 2 shows an example of annotation projection using translated text.

3 Learning Translation from Wikipedia

The key component of our approach is to leverage the multilingual cues from linked Wikipedia pages across languages. Wikipedia is a great comparable data in which many of its pages explain entities in the world in different languages. In most cases, first sentences define or introduce the mentioned entity in that page (e.g. Figure 1). Therefore, we observe that many first sentence pairs in linked Wikipedia documents are rough translations of each other. Moreover, captions of images in different languages are usually similar but not necessarily direct translations of each other. We leverage this information to extract many parallel sentences from Wikipedia without using any external supervision. In this section, we describe our algorithm which is briefly shown in Figure 3.

3.1 Data Definitions

For languages e and f in which e is English and f is a low-resource target language of interest, there are Wikipedia documents $w_e = \{w_1^{(e)} \ldots w_n^{(e)}\}$ and $w_f = \{w_1^{(f)} \ldots w_m^{(f)}\}$. We refer to $w_{(i,j)}^{(l)}$ as the jth sentence in the ith document for language l. A subset of these documents are aligned (using Wikipedia *languages links*). Thus we have an aligned set of document pairs in which we can easily extract many sentence pairs that are potentially translations of each other. A smaller subset $\mathcal{F}$ is the set of first sentences in Wikipedia $(w_{(i,1)}^{(e)}, w_{(i',1)}^{(f)})$ in which documents i and i' are linked and their first sentence lengths are in a similar range. In addition to text content, Wikipedia has a large set of images. Each image comes along with one or more captions, sometimes in different languages. A small subset of these images have captions both in English and the target language. We refer to this set as $\mathcal{C}$. We use the set of all caption pairs ($\mathcal{C}$), title pairs ($\mathcal{T}$), and first sentences ($\mathcal{F}$) as the seed parallel data: $\mathcal{S} = \mathcal{F} \cup \mathcal{C} \cup \mathcal{T}$.

3.2 Bilingual Dictionary Extraction and Cross-Lingual Word Embeddings

Having the seed parallel data $\mathcal{S}$, we run unsupervised word alignment (Dyer et al., 2013) in both English-to-target, and target-to-English directions. We use the intersected alignments to extract highly confident word-to-word connections. Finally, we pick the most frequently aligned word for each word in English as translation. This set serves as a bilingual dictionary $\mathcal{D}$.

Given two monolingual trained word embeddings $v_e \in \mathbb{R}^{N_e \times d}$ and $v_f \in \mathbb{R}^{N_f \times d}$, and the extracted bilingual dictionary $\mathcal{D}$, we use the method of Faruqui and Dyer (2014) to project these two embedding vectors to a shared cross-lingual space.[2] This method uses a bilingual dictionary along with

[2] There are more recent approaches such as (Lample et al., 2018b). Comparing different embedding methods is not the focus of this paper, thereby we leave further investigation to future work.

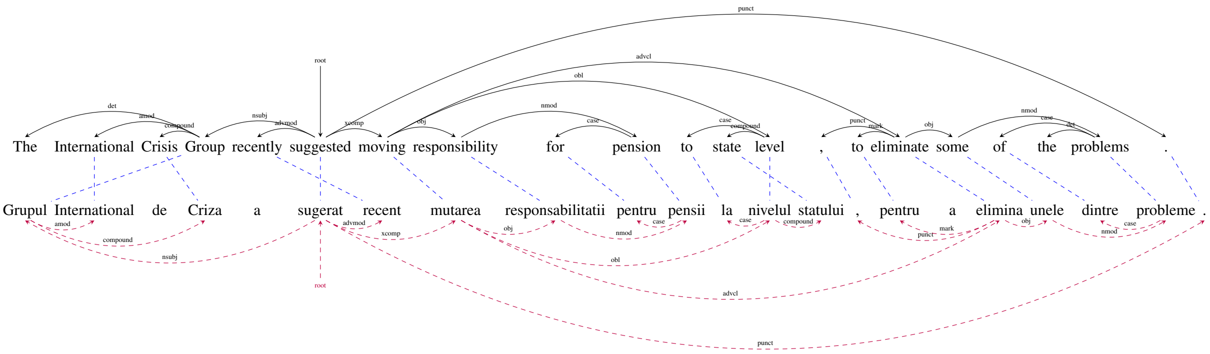

The International Crisis Group recently suggested moving responsibility for pension to state level , to eliminate some of the problems .
Grupul International de Criza a sugerat recent mutarea responsabilitatii pentru pensii la nivelul statului , pentru a elimina unele dintre probleme .
root
det
amod
compound
nsubj
advmod
xcomp
obj
obl
advcl
punct
nmod
case
compound
punct
mark
obj
nmod
case
det
amod
compound
nsubj
advmod
xcomp
obj
nmod
case
case
compound
obl
punct
mark
obj
case
nmod
advcl
punct
root

Figure 2: An example of annotation projection for which the source on top is a translation of the Romanian target via our *wikily* translation model. The supervised source tree is projected using intersected word alignments.

Definitions: 1) e is English, f is the foreign language, and g is a language similar to f, 2) learn_dict (P) extracts a bilingual dictionary from parallel data P, 3) t $(x|m)$ translates input x given model m, , 4) pretrain (x) pretrains on monolingual data x using MASS (Song et al., 2019a), 5) train $(P|m)$ trains on parallel data P initialized by model m, 6) bt_train $(x_1, x_2|m)$ trains iterative back-translation on monolingual data $x_1 \in e$ and $x_2 \in f$ initialized by model m.

Inputs: 1) Wikipedia documents $w^{(e)}$, $w^{(f)}$, and $w^{(g)}$, 2) Monolingual word embedding vectors v_e and v_f, 3) Set of linked pages from Wikipedia COMP, their aligned titles $\mathcal{T}$, and their first sentence pairs $\mathcal{F}$, 4) Set of paired image captions $\mathcal{C}$, and 5) Gold-standard parallel data $\mathcal{P}^{(e,g)}$.

Algorithm:

```
→ Learn bilingual dictionary and embeddings
  $\mathcal{S} = \mathcal{F} \cup \mathcal{C} \cup \mathcal{T}$
  $\mathcal{D}^{(f,e)} =$ learn_dict $(\mathcal{S})$
  $\mathcal{D}^{(g,e)} =$ learn_dict $(\mathcal{P}^{(e,g)})$          ▷ Related language
  Learn $v_e \rightarrow v'_e$ and $v_f \rightarrow v'_f$ using $\mathcal{D}^{(f,e)} \cup \mathcal{D}^{(g,e)}$
→ Mine parallel data
  Extract comparable sentences $\mathcal{Z}$ from COMP
  Extract $\mathcal{P}^{(f,e)}$ from $\mathcal{Z}$.
  $\mathcal{P}^{(f,e)} = \mathcal{P}^{(f,e)} \cup \mathcal{T}$          ▷ Mined Data
→ Train MT with pretraining and back-translation
  $\theta_0 =$ pretrain $(w^{(e)} \cup w^{(f)} \cup w^{(g)})$          ▷ MASS Training
  $\theta_{\rightleftarrows} =$ train $(\mathcal{P}^{(f,e)} \cup \mathcal{P}^{(g,e)}|\theta_0)$          ▷ NMT Training
  $\mathcal{P}^{(e \rightarrow f)} = ($ t $(w^{(f)}|\theta_{\rightleftarrows}), w^{(f)})$
  $\mathcal{P}^{(f \rightarrow e)} = ($ t $(w^{(e)}|\theta_{\rightleftarrows}), w^{(e)})$
  $\mathcal{P}'^{(f,e)} = \mathcal{P}^{(e \rightarrow f)} \cup \mathcal{P}^{(f \rightarrow e)} \cup \mathcal{P}^{(f,e)}$
  $\theta'_{\rightleftarrows} =$ train $(\mathcal{P}'^{(f,e)}|\theta_0)$
  $\theta *_{\rightleftarrows} =$ bt_train $(w^{(e)}, w^{(f)}|\theta'_{\rightleftarrows})$
Output: $\theta^*_{\rightleftarrows}$
```

Figure 3: A brief depiction of the training pipeline.

canonical correlation analysis (CCA) to learn two projection matrices to map each embedding vector to a shared space $v'_e \in \mathbb{R}^{N_e \times d'}$ and $v'_f \in \mathbb{R}^{N_f \times d'}$ where $d' \leq d$.

3.3 Mining Parallel Sentences

We use cross-lingual embedding vectors $v'_e \in \mathbb{R}^{N_e \times d}$ and $v'_f \in \mathbb{R}^{N_f \times d'}$ for calculating the cosine similarity between pairs of words. Moreover, we use the extracted bilingual dictionary to boost the accuracy of the scoring function. For a pair of sentences (s, t) where $s = s_1 \ldots s_n$ and $t = t_1 \ldots t_m$, after filtering sentence pairs with different numerical values (e.g. sentences containing 2019 in the source and 1987 in the target), we use a modified version of cosine similarity between words:

$$sim(s_i, t_j) = \begin{cases} 1.0, & \text{if } (s_i, t_j) \in \mathcal{D} \\ cos(s_i, t_j), & \text{otherwise} \end{cases}$$

Using the above definition of word similarity, we use the average-maximum similarity between pairs of sentences.

$$score(s, t) = \frac{\sum_{i=1}^{n} \max_{j=1}^{m} sim(s_i, t_i)}{n}$$

From a pool of candidates, we pick those pairs that have the highest score in both directions.

3.4 Leveraging Similar Languages

In many low-resource scenarios, the number of paired documents is very small, leading to a small number and often noisy extracted parallel sentences. To alleviate this problem to some extent, we assume to have another language g in which g has a large lexical overlap with the target language f (such as g=Russian and f=Kazakh). We assume that a parallel data exists between language g and English, and we can use it both as an auxiliary parallel data in training, and also for extracting extra lexical entries for the bilingual dictionaries: as shown in Figure 3, we supplement the extracted bilingual dictionary from seed parallel data with the bilingual dictionary extracted from related language parallel data.

3.5 Translation Model

We use a standard sequence-to-sequence transformer-based translation model (Vaswani et al., 2017) with a six-layer BERT-based (Devlin et al., 2019) encoder-decoder architecture from HuggingFace (Wolf et al., 2019) and Pytorch (Paszke et al., 2019) with a shared SentencePiece (Kudo and Richardson, 2018) vocabulary. All input and output token embeddings are summed up with the language id embedding. First tokens of every input and output sentence are shown by the language ID. Our training pipeline assumes that the encoder and decoder are shared across different languages, except that we use a separate output layer for each language in order to prevent input copying (Artetxe et al., 2018b; Sen et al., 2019). We pretrain the model on a tuple of three Wikipedia datasets for the three languages g, f, and e using the MASS model (Song et al., 2019a). The MASS model masks a contiguous span of input tokens, and recovers that span in the output sequence.

To facilitate multi-task learning with image captioning, our model has an image encoder that is used in cases of image captioning (more details in §4.1). In other words, the decoder is shared between the translation and captioning tasks. We use the pretrained ResNet-152 model (He et al., 2016) from Pytorch to encode every input image. We extract the final layer as a 7×7 grid vector ($g \in \mathbb{R}^{7\times7\times d_g}$), and project it to a new space by a linear transformation ($g' \in \mathbb{R}^{49\times d_t}$), and then add location embeddings ($l \in \mathbb{R}^{49\times d_t}$) by using entry-wise addition. Afterwards, we assume that the 49 vectors are encoded text representations as if a sentence with 49 words occurs. This is similar to but not exactly the same as the Virtex model (Desai and Johnson, 2021).

3.6 Back-Translation: One-shot and Iterative

Finally, we use the back-translation technique to improve the quality of our models. Back-translation is done by translating a large amount of monolingual text to and from the target language. The translated texts serve as noisy input text along with the monolingual data as the silver-standard translations. Previous work (Sennrich et al., 2016b; Edunov et al., 2018) has shown that back-translation is a very simple but effective technique to improve the quality of translation models. Henceforth, we refer to this method as *one-shot back-translation*. Another approach is to use *iterative back-translation* (Hoang et al., 2018), the most popular approach in unsupervised translation (Artetxe et al., 2018b; Conneau and Lample, 2019; Song et al., 2019a). The main difference from one-shot translation is that the model uses an online approach, and updates its parameters in every batch.

We empirically find *one-shot back-translation* faster to train but with much less potential to reach a high translation accuracy. A simple and effective way to have both a reliable and accurate model is to first initialize a model with one-shot back-translation, and then apply iterative back-translation. The model that is initialized with a more accurate model reaches a higher accuracy.

4 Cross-Lingual Tasks

In this section, we describe our approaches for tailoring our translation models to cross-lingual tasks. Note that henceforth we assume that our translations model training is finished, and we have access to trained translation models for cross-lingual tasks.

4.1 Cross-Lingual Image Captioning

Having gold-standard image captioning training data $\mathcal{I} = \{(I_i, c_i)\}_{i=1}^{n}$ where I_i is the image as pixel values, and $c_i = c_i^{(1)}, \ldots, c_i^{k_i}$ as the textual description with k_i words, our goal is to learn a captioning model that is able to describe new (unseen) images. As described in §3.5, we use a transformer decoder from our translation model and a ResNet image encoder (He et al., 2016) for our image captioning pipeline. Unfortunately, annotated image captioning datasets do not exist in many languages. Having our translation model parameter $\theta^*_{\rightleftarrows}$, we can use its translation functionality to translate each caption c_i to $c'_i = translate(c_i|\theta^*_{\rightleftarrows})$. Afterwards, we will have a translated annotated dataset $\mathcal{I}' = \{(I_i, c'_i)\}_{i=1}^{n}$ in which the textual descriptions are not gold-standard but translations from the English captions. Figure 4 shows a real example from MS-Coco (Chen et al., 2015) in which Arabic translations are provided by our translation model. Furthermore, to augment our learning capability, we initialize our decoder with decoding parameters of $\theta^*_{\rightleftarrows}$, and also continue training with both English captioning and translation.

4.2 Cross-Lingual Dependency Parsing

Assuming that we have a large body of monolingual text, we translate that monolingual text to create artificial parallel data. We run unsupervised word alignments on the artificial parallel text. Following previous work (Rasooli and Collins, 2015; Ma and Xia, 2014), we run Giza++ (Och and Ney,

This is an open box containing four cucumbers.

وهذا صندوق مفتوح يحتوي على أربعة خيار.

An open food container box with four unknown food items.

صندوق حاوية طعام مفتوح مع أربعة مواد غذائية مجهولة.

A small box filled with four green vegetables.

مربع صغير مليء بأربعة الخضروات الخضراء.

An opened box of four chocolate bananas.

علبة مفتوحة من أربعة من الموز.

An open box contains an unknown, purple object

مربع مفتوح يحتوي على كائن غير معروف الأرجوان

Figure 4: An image from MS-Coco (Chen et al., 2015) with gold-standard English captions, and Arabic translations from our *wikily* translation model.

2003b) alignments on both source-to-target and target-to-source directions, and extract intersected alignments to keep high-precision one-to-one alignments. We run a supervised dependency parser of English as our rich-resource language. Then, we project dependencies to the target language sentences via word alignment links. Inspired by previous work (Rasooli and Collins, 2015), to remove noisy projections, we keep those sentences that at least 50% of words or 5 consecutive words in the target side have projected dependencies.

5 Experiments

In this section, we provide details about our experimental settings and results for translation, captioning, and dependency parsing. We put more details about our settings as well as thorough analysis of our results in the supplementary material.

5.1 Datasets and Settings

Languages We focus on four language pairs: Arabic-English, Gujarati-English, Kazakh-English, and Romanian-English. We choose these pairs to provide enough evidence that our model works in distant languages, morphologically-rich languages, as well as similar languages. As for similar languages, we use Persian for Arabic (written with very similar scripts and have many words in common), Hindi for Gujarati (similar languages), Russian for Kazakh (written with the same script), and Italian for Romanian (Romance languages).

Monolingual and Translation Datasets We use a shared SentencePiece vocabulary (Kudo and Richardson, 2018) with size 60K. Table 1 shows the sizes of Wikipedia data in different languages. For evaluation, we use the Arabic-English UN data (Ziemski et al., 2016), WMT 2019 data (Barrault et al., 2019) for Gujarati-English and Kazakh-English, and WMT 2016 shared task data (Bojar et al., 2016) for Romanian-English. Following previous work (Sennrich et al., 2016a), diacritics are removed from the Romanian data. More details about other datasets and their sizes, we refer the reader to the supplementary material.

Direction	ar⇄en	gu⇄en	kk⇄en	ro⇄en
Foreign docs	1.0m	28k	230k	400k
Paired docs	745k	7.3k	80k	270k
First sents.	205k	3.2k	52k	78k
Captions	92k	2.2k	1.9k	35k
Comparable pairs	0.1b	14m	32m	64m
Mined sents.	1.7m	49k	183k	675k
BT	2.1m	1.5m	2.2m	2.1m
Iterative BT	4.0m	3.8m	4.0m	6.1m

Table 1: Data sizes for different pairs. We use a sample of English sentences with similar sizes to each data.

Pretraining We pretrain four models on 3-tuples of languages via a single NVIDIA Geforce RTX 2080 TI with 11GB of memory. We create batches of 4K words, run pretraining for two million iterations where we alternate between language batches, and accumulate gradients for 8 steps. We use the apex library[3] to use FP-16 tensors. This whole process takes four weeks in a single GPU. We use the Adam optimizer (Kingma and Ba, 2015) with inverse square root and learning rate of 10^{-4}, 4000 warm-up steps, and dropout probability of 0.1.

Translation Training Table 1 shows the sizes of different types of datasets in our experiments. We pick comparable candidates for sentence pairs whose lengths are within a range of half to twice of each other. As we see, the final size of mined datasets heavily depends on the number of paired English-target language Wikipedia documents. We train our translation models initialized by pretrained models. More details about our hyperparameters are in the supplementary material. All of our evaluations are conducted using SacreBLEU (Post, 2018) except for en↔ro in which we use BLEU score (Papineni et al., 2002) from Moses decoder scripts (Koehn et al., 2007) for the sake of comparison to previous work.

Image Captioning We use the Flickr (Hodosh et al., 2013) and MS-Coco (Chen et al., 2015) datasets for English[4], and the gold-standard Arabic Flickr dataset (ElJundi. et al., 2020) for evaluation. The Arabic test set has 1000 images with 3 captions

[3] https://github.com/NVIDIA/apex

[4] We have also tried Conceptual Captions (Sharma et al., 2018) in our initial experiments but we have observed drops in performance. Previous work (Singh et al., 2020) have also observed a similar problem with Conceptual Captions as a noisy crawled caption dataset.

per image. We translate all the training datasets to Arabic for having translated caption data. The final training data contains $620K$ captions for about $125K$ unique images. Throughout experiments, we use the pretrained Resnet-152 models (He et al., 2016) from Pytorch (Paszke et al., 2019), and let it fine-tune during our training pipeline. Each training batch contains 20 images. We accumulate gradients for 16 steps, and use a dropout of 0.1 for the projected image output representations. Other training parameters are the same as our translation training. To make our pipeline fully unsupervised, we use translated development sets to pick the best checkpoint during training.

Dependency Parsing We use the Universal Dependencies v2.7 collection (Zeman et al., 2020) for Arabic, Kazakh, and Romanian. We use the Stanza (Qi et al., 2020) pretrained supervised models for getting supervised parse trees for Arabic and Romanian, and use the UDPipe (Straka et al., 2016) pretrained model for Kazakh. We translate about 2 million sentences from each language to English, and also 2 million English sentences to Arabic. We use a simple modification to Stanza to facilitate training on partially projected trees by masking dependency and label assignments for words with missing dependencies. All of our training on projected dependencies is blindly conducted with $100k$ training steps with default parameters of Stanza (Qi et al., 2020). As for gold-standard parallel data, we use our supervised translation training data for Romanian-English and Kazakh-English and use a sample of 2 million sentences from the UN Arabic-English data due to its large size that causes word alignment significant slowdown. For Kazakh *wikily* projections, due to low supervised POS accuracy, we use the projected POS tags for projected words and supervised tags for unprojected words. We observe a two percent increase in performance by using projected tags.

5.2 Translation Results

Table 2 shows the results of different settings in addition to baseline and state-of-the-art results. We see that Arabic as a clear exception needs more rounds of training: we train our Arabic model once again on mined data by initializing it by our back-translation model.[5] We have not seen further improvement by back-translation. To have a fair comparison, we list the best supervised models for all language pairs (to the best of our knowledge). In low-resource settings, we outperform strong supervised models that are boosted by back-translation. In high-resource settings, our Arabic models achieve very high performance but regarding the fact that the parallel data for Arabic has 18M sentences, it is quite impossible to reach that level of accuracy.

Figure 5 shows a randomly chosen example from the Gujarati-English development data. As depicted, we see that the model after back-translation reaches to somewhat the core meaning of the sentence with a bit of divergence from exactly matching the reference. The final iterative back-translation output almost catches a correct translation. We also see that the use of the word “creative” is seen in Google Translate output, a model that is most likely trained on much larger parallel data than what is currently available for public use. In general, unsupervised translation performs very poorly compared to our approach in all directions.

5.3 Captioning Results

Table 4 shows the final results on the Arabic test set using the SacreBLEU measure (Post, 2018). First, we should note that similar to ElJundi. et al. (2020), we see lower scales of BLEU scores due to morphological richness in Arabic. We see that if we initialize our model with the translation model and multitask it with translation and also English captioning, we achieve much higher performance. It is interesting to observe that translating the English output on the test data to Arabic achieves a much lower result. This is a strong indicator of the strength of our approach. We also see that supervised translation fails to perform well. This might due to the UN translation training dataset which has a different domain from the caption dataset. Furthermore, we see that our model outperforms Google Translate which is a strong machine translation system, and that is actually what is being used as seed data for manual revision in the Arabic dataset. Finally, it is interesting to see that our model outperforms supervised captioning. Multi-tasking make translation performance slightly worse.

Figure 6 shows a randomly picked example with

[5] We have seen that during multi-tasking with image captioning, the translation BLEU score for Arabic-English significantly improves. We initially thought that multi-tasking is improving both translation and captioning, but our further investigation shows that it is actually due to lack of training for Arabic. We have tried the same procedure for other languages but have not observed any further gains.

	Model	ar→en	en→ar	gu→en	en→gu	kk→en	en→kk	ro→en	en→ro
UNMT	Conneau and Lample (2019)	–	–	–	–	–	–	31.8	33.3
	Song et al. (2019a) (MASS; 8 GPUs)	–	–	–	–	–	–	33.1	35.2
	Best published results	11.0*	9.4*	0.6^1	0.6^1	2.0^1	0.8^1	37.6^4	36.3^2
Wikily UNMT	First sentences + captions + titles	6.1	3.1	0.7	1.1	2.3	1.0	2.0	1.9
	Mined Corpora	23.1	19.7	4.2	4.9	2.8	1.6	22.1	21.6
	+ Related Language	–	–	9.1	7.8	7.3	2.3	23.2	21.5
	+ One-shot back-translation (bt-beam=4)	23.0	18.8	**13.8**	13.9	7.0	**12.1**	25.2	28.1
	+ Iterative back-translation (bt-beam=1)	24.4	18.9	13.3	**15.2**	**9.0**	10.8	32.5	33.0
	+ Retrain on mined data	**30.6**	**23.4**	–	–	–	–	–	–
	(Semi-)Supervised	48.9*	40.6*	14.2^1	4.0^1	12.5^1	3.1^1	39.9^3	38.5^3

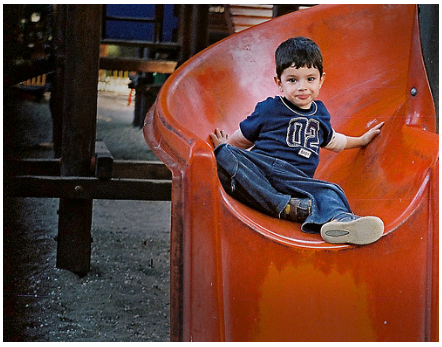
Table 2: BLEU scores for different models. Reference results are from *: Our implementation, 1: Kim et al. (2020), 2: Li et al. (2020), 3: Liu et al. (2020) (supervised), 4: Tran et al. (2020) (unsupervised with mined parallel data).

	Method	Version	Token and POS	Arabic			Kazakh			Romanian		
				UAS	LAS	BLEX	UAS	LAS	BLEX	UAS	LAS	BLEX
Previous	Rasooli and Collins (2019)	2.0	gold/supervised	61.2	48.8	–	–	–	–	76.3	64.3	–
	Ahmad et al. (2019)	2.2	gold	38.1	28.0	–	–	–	–	65.1	54.1	–
	Kurniawan et al. (2021)	2.2	gold	48.3	29.9	–	–	–	–	–	–	–
Projection	*Wikily* translation	2.7	gold	62.5	50.7	46.3	46.8	28.5	25.0	74.1	57.7	52.6
			supervised	60.2	48.7	42.1	46.2	27.8	14.1	73.6	57.4	50.9
	Gold-standard Parallel data		gold	61.5	47.3	42.4	22.2	9.3	7.9	75.9	62.4	57.3
			supervised	59.1	45.3	38.5	21.8	9.2	3.8	75.6	62.0	55.6
	Supervised		supervised	84.2	79.8	72.7	48.0	29.8	13.7	90.8	86.0	80.0

Table 3: Dependency parsing results on the Universal Dependencies dataset (Zeman et al., 2020). Previous work has used different sub-versions of the Universal Dependencies data in which slight differences are expected.

	Input	અર્થાત આપણે પહેલા તુલનાએ વધુ રચનાત્મક બનવું પડશે.
Outputs	Unsupervised	Ut numerousત્રીit the mother, onwards, in theover અધિકાંશexualit theotherit theIN રોડ 19
	First sentences + captions + titles	A view of the universe from the present to the present day.
	Mined Corpora	For example, if the ghazal is more popular than ghazal.
	+ Related Language	We need to become more creative than before.
	+ One-shot back-translation	For example, we must become more creative than before.
	+ Iterative back-translation	Meanwhile, we 'll have to become more constructive than before.
	Google Translate	That means we have to be more creative than before.
	Reference	That means we have to be more constructive than before.

Figure 5: An example of a Gujarati sentence and its outputs from different models, as well as Google Translate.



English gold	A child on a red slide. A little boy sits on a slide on the playground. A little boy slides down a bright red corkscrew slide. A little boy slides down a red slide. a young boy wearing a blue outfit sliding down a red slide.
English supervised	A boy is sitting on a red slide.
En– supervised translate En– unsupervised translate En– Google translate	– صبي صبي يجلس على شاحنة خفيفة . الطفل يجلس على شريحة حمراء. صبي يجلس على شريحة حمراء.
Supervised MT	صبي صبي على شظية
Unsupervised (mt + ar + en) Unsupervised (mt + ar)	يجلس صبي صغير على شريحة برتقالية. صبي صغير يجلس على شريحة حمراء.
Supervised	صبي في قميص أزرق يقفز في الهواء
Arabic Gold	طفل على منزلقة حمراء صبي صغير يجلس على زلاجة في الملعب ينزلق صبي صغير أسفل منزلقة حمراء

Figure 6: An example of different outputs in our captioning experiments both for English and Arabic, as well as Arabic translations of English outputs on the Arabic Flickr dataset (ElJundi. et al., 2020).

different model outputs. We see that the two outputs from our approach with multi-tasking are roughly the same but one of them as more syntactic order overlap with the reference while both orders are correct in Arabic as a free-word order language. The word برتقالية means "orange" which is close to حمراء that means "red". The word شريحة means "slide" which is correct but other meanings of this word exist in the reference. In general, we observe

	Supervision	Pretrained	Multi-task		BLEU	
			EN	MT	@1	@4
Translate train data	*wikily*	✗	✗	✗	33.1	4.57
	wikily	✓	✗	✗	32.9	5.28
	wikily	✓	✓	✗	32.8	4.37
	wikily	✓	✗	✓	33.3	🟩 5.72
	wikily	✓	✓	✓	🟩 36.8	5.60
	supervised	✓	✗	✗	17.7	1.26
Translate test		English test performance→			68.7	20.42
	wikily	✓	✗	✗	30.6	4.20
	supervised	✓	✗	✗	15.8	0.92
	Google	✓	✗	✗	31.8	5.56
	Gold	✓	✗	✗	33.7	3.76
		✓	✓	✗	🟦 37.9	5.22

Table 4: Image captioning results evaluated on the Arabic Flickr dataset (ElJundi. et al., 2020) using SacreBLEU (Post, 2018). "pretrained" indicates initializing our captioning model with our translation parameters.

that although superficially the BLEU scores for Arabic is low, it is mostly due to its lexical diversity, free-word order, and morphological complexity.

5.4 Dependency Parsing Results

Table 3 shows the results for dependency parsing experiments. We see that our model performs very high in Romanian with a UAS of 74 which is much higher than that of Ahmad et al. (2019) and slightly lower than that of Rasooli and Collins (2019) which uses a combination of multi-source annotation projection and direct model transfer. Our work on Arabic outperforms all previous work and performs even better than using gold-standard parallel data. One clear highlight is our result in Kazakh. As mentioned before, by projecting the part-of-speech tags, we achieve roughly 2 percent absolute improvement. Our final results on Kazakh are significantly higher than that of using gold-standard parallel text (7K sentences).

6 Related Work

Kim et al. (2020) has shown that unsupervised translation models often fail to provide good translation systems for distant languages. Our work solves this problem by leveraging the Wikipedia data. Using pivot languages has been used in previous work (Al-Shedivat and Parikh, 2019), as well as using related languages (Zoph et al., 2016; Nguyen and Chiang, 2017). Our work only explores a simple idea of adding one similar language pair. Most likely, adding more language pairs and using ideas from recent work might improve the performance.

Wikipedia is an interesting dataset for solving NLP problems including machine translation (Li et al., 2012; Patry and Langlais, 2011; Lin et al., 2011; Tufiş et al., 2013; Barrón-Cedeño et al., 2015; Wijaya et al., 2017; Ruiter et al., 2019; Srinivasan et al., 2021). The WikiMatrix data (Schwenk et al., 2019a) is the most similar effort to ours in terms of using Wikipedia, but with using supervised translation models. Bitext mining has a longer history of research (Resnik, 1998; Resnik and Smith, 2003) in which most efforts are spent on using a seed supervised translation model (Guo et al., 2018; Schwenk et al., 2019b; Artetxe and Schwenk, 2019; Schwenk et al., 2019a; Jones and Wijaya, 2021). Recently, a number of papers have focused on unsupervised extraction of parallel data (Ruiter et al., 2019; Hangya and Fraser, 2019; Keung et al., 2020; Tran et al., 2020; Kuwanto et al., 2021). Ruiter et al. (2019) focus on using vector similarity of sentences to extract parallel text from Wikipedia. Their work does not leverage structural signals from Wikipedia.

Cross-lingual and unsupervised image captioning has been studied in previous work (Gu et al., 2018; Feng et al., 2019; Song et al., 2019b; Gu et al., 2019; Gao et al., 2020; Burns et al., 2020). Unlike previous work, we do not have a supervised translation model. Cross-lingual transfer of dependency parser have a long history. We encourage the reader to read a recent survey on this topic (Das and Sarkar, 2020). Our work does not use gold-standard parallel data or even supervised translation models to apply annotation projection.

7 Conclusion

We have described a fast and effective algorithm for learning translation systems using Wikipedia. We show that by wisely choosing what to use as seed data, we can have very good seed parallel data to mine more parallel text from Wikipedia. We have also shown that our translation models can be used in downstream cross-lingual natural language processing tasks. In the future, we plan to extend our approach beyond Wikipedia to other comparable datasets like the BBC World Service. A clear extension of this work is to try our approach on other cross-lingual tasks. Moreover, as many captions of the same images in Wikipedia are similar sentences and sometimes translations, multimodal machine translation (Specia et al., 2016; Caglayan et al., 2019; Hewitt et al., 2018; Yao and Wan, 2020) based on this data or the analysis of the data, such as whether more similar languages may share more similar captions (Khani et al., 2021) are other interesting avenues.

Acknowledgments

We would like to thank reviewers and the editor for their useful comments. We also would like to thank Alireza Zareian, Daniel (Joongwon) Kim, Qing Sun, and Afra Feyza Akyurek for their help and useful comments througout this project. This work is supported in part by the DARPA HR001118S0044 (the LwLL program), and the Department of the Air Force FA8750-19- 2-3334 (Semi-supervised Learning of Multimodal Representations). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA, the Air Force, and the U.S. Government.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. 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 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. 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, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Mikel Artetxe and Holger Schwenk. 2019. 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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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.

Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, and Lluís Màrquez. 2015. A factory of comparable corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13, Beijing, China. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondrej Bojar, Vojtech Diatka, Pavel Rychlỳ, Pavel Stranák, Vít Suchomel, Ales Tamchyna, and Daniel Zeman. 2014. Hindencorp-hindi-english and hindi-only corpus for machine translation. In *LREC*, pages 3550–3555.

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A Plummer. 2020. Learning to scale multilingual representations for vision-language tasks. In *European Conference on Computer Vision*, pages 197–213. Springer.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Ayan Das and Sudeshna Sarkar. 2020. A survey of the model transfer approaches to cross-lingual dependency parsing. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5):1–60.

Karan Desai and Justin Johnson. 2021. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*.

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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Obeida ElJundi., Mohamad Dhaybi., Kotaiba Mokadam., Hazem Hajj., and Daniel Asmar. 2020. Resources and end-to-end neural network models for arabic image captioning. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 233–241. INSTICC, SciTePress.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134.

Jiahui Gao, Yi Zhou, Philip LH Yu, and Jiuxiang Gu. 2020. Unsupervised cross-lingual image captioning. *arXiv preprint arXiv:2010.01288*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 2018. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 503–519.

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10323–10332.

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

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.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

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.

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.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(03):311–325.

Alex Jones and Derry Tanti Wijaya. 2021. Majority voting with bidirectional pre-translation for bitext retrieval.

Omid Kashefi. 2018. Mizan: a large Persian-English parallel corpus. *arXiv preprint arXiv:1801.02107*.

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*.

Nikzad Khani, Isidora Tourni, Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya. 2021. Cultural and geographical influences on image translatability of words across languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 198–209.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency parsing. *Synthesis lectures on human language technologies*, 1(1):1–127.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kemal Kurniawan, Lea Frermann, Philip Schulz, and Trevor Cohn. 2021. Ppt: Parsimonious parser transfer for unsupervised cross-lingual adaptation. *arXiv preprint arXiv:2101.11216*.

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, and Derry Wijaya. 2021. Low-resource machine translation for low-resource languages: Leveraging comparable data, code-switching and compute resources.

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.

Shen Li, Joao V Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1389–1398. Association for Computational Linguistics.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020. Data-dependent gaussian prior objective for language generation. In *International Conference on Learning Representations*.

Wen-Pin Lin, Matthew Snover, and Heng Ji. 2011. Unsupervised language-independent name translation mining from wikipedia infoboxes. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 43–52.

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. *arXiv cs.CL 2001.08210*.

Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1337–1348, Baltimore, Maryland. Association for Computational Linguistics.

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.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Franz Josef Och and Hermann Ney. 2003a. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney. 2003b. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.

Alexandre Patry and Philippe Langlais. 2011. Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in Wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95, Portland, Oregon. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338, Lisbon, Portugal. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *Transactions of the Association for Computational Linguistics*, 5:279–293.

Mohammad Sadegh Rasooli and Michael Collins. 2019. Low-resource syntactic transfer with unsupervised source reordering. 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 3845–3856, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Resnik. 1998. Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Conference of the Association for Machine Translation in the Americas*, pages 72–82. Springer.

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

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith. 2019. Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. 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 cs.CL 1907.05791*.

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, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019a. MASS: Masked sequence to sequence pre-training for language generation. *arXiv cs.CL 1905.02450*.

Yuqing Song, Shizhe Chen, Yida Zhao, and Qin Jin. 2019b. Unpaired cross-lingual image caption generation with self-supervised rewards. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 784–792.

Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Yunwon Tae, Cheonbok Park, Taehee Kim, Soyoung Yang, Mohammad Azam Khan, Eunjeong Park, Tao Qin, and Jaegul Choo. 2020. Meta-learning for low-resource unsupervised neural machine translation. *arXiv preprint arXiv:2010.09046*.

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.

Dan Tufiş, Radu Ion, Stefan Daniel Dumitrescu, and Dan Stefanescu. 2013. Wikipedia as an SMT training corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 702–709.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, et al. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Cross-Lingual Embedding

We use the off-the-shelf 300-dimensional FastText embeddings (Grave et al., 2018) as monolingual embedding vectors. We run FastAlign (Dyer et al., 2013) on the seed parallel text from both source-to-target and target-to-source directions, run alignment intersection to get intersected alignments, and extract the highest occurring alignment for every word as the dictionary entry. We make use of the cross-lingual CCA tool (Faruqui and Dyer, 2014) to extract 150-dimensional vectors. This tool can be run on a single CPU within a few hours.

B Monolingual and Translation Datasets

We use an off-the-shelf Indic-transliteration library[6] to convert the Devanagari script to Hindi script to make the Hindi documents look like Gujarati by removing the graphical vertical bars from Hindi letters, thus increasing the chance of capturing more words in common. We boost the Romanian, Gujarati, and Kazakh monolingual data with newstext dataset from WMT. For parallel data in similar languages, we use the Mizan parallel data for Persian (Kashefi, 2018) with one million sentences, the IITB data (Kunchukuttan et al., 2018) and HindiEnCorp 0.5 (Bojar et al., 2014) for Hindi with a total of 367K sentences, ParaCrawl for Russian (Esplà et al., 2019) with 12M sentences, and Europarl for Italian (Koehn, 2005) with 2M sentences. We use the Arabic-English UN data (Ziemski et al., 2016), WMT 2019 data (Barrault et al., 2019) for Gujarati-English and Kazakh-English, and WMT 2016 shared task data (Bojar et al., 2016) for Romanian-English. Following previous work (Sennrich et al., 2016a), diacritics are removed from the Romanian data.

C Translation Training Parameters

We pick comparable candidates for sentence pairs whose lengths are within a range of half to twice of each other. As we see, the final size of mined datasets heavily depends on the number of paired English-target language Wikipedia documents. We train our translation models initialized by pretrained models. Each batch has roughly 4K tokens. Except for Arabic, for which the size of mined data significantly outnumbers the size of Persian-English parallel data, we use the related language data before using iterative back-translation in which we only use the source and target monolingual datasets. We use similar learning hyper-parameters to pretraining except for iterative back-translation in which we accumulate gradients for 100 steps, and use a dropout probability of 0.2 and 10000 warmup steps since we find smaller dropout and warmup make the model diverge. Our one-shot back-translation experiments use a beam size of 4, but we use a beam size of one for iterative back-translation since we have not seen much gains in using beam-based iterative back-translation except for purely unsupervised settings. All of our translations are performed with a beam size of 4 and $max\_len\_a = 1.3$ and $max\_len\_b = 5$. We alternate between supervised parallel data of a similar language paired with English and the mined data.

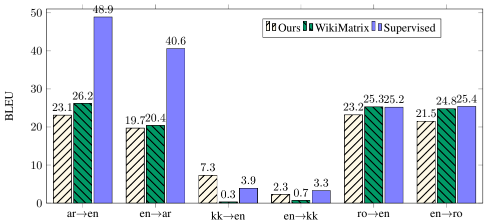


Figure 7: Results using our mined data versus WikiMatrix (Schwenk et al., 2019a) and gold-standard data.

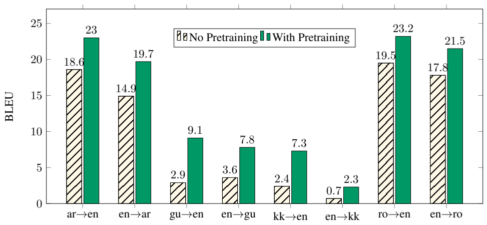


Figure 8: Results using mined data (no back-translation) with and without pretraining.

We train translation models for roughly 400K batches except for Gujarati that has smaller mined data for which we train for 200K iterations. We have seen a quick divergence in Kazakh iterative back-translation, thereby we stopped it early after running it for one epoch of all monolingual data. Most likely, the mined data for Kazakh-English has lower quality (see the supplementary material for more details), and that leads to very noisy translations in back-translation outputs. All of our evaluations are conducted using SacreBLEU (Post, 2018) except for en↔ro in which we use BLEU score (Papineni et al., 2002) from Moses decoder scripts (Koehn et al., 2007) for the sake of comparison to previous work.

D Quality of Mined Data

The quality of parallel data matters a lot for getting high-accuracy. For example, we manually observe that the quality of mined data for all languages are very good except for Kazakh. Our hypothesis is that the Kazakh Wikipedia data is less aligned with the English content. We compare our mined data to that of the supervised mined data from Wiki-

[6] https://pypi.org/project/indic-transliteration

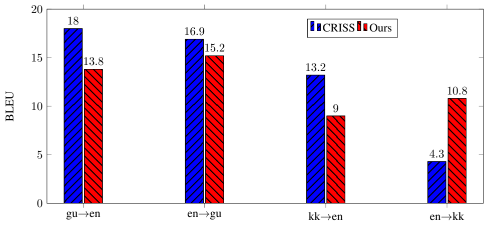


Figure 9: Our best results versus the *supervised* model of Tran et al. (2020).

Matrix (Schwenk et al., 2019a) as well as gold-standard data. Figure 7 shows the difference between the three datasets of three language pairs (WikiMatrix does not contain Gujarati). As we see, our data has BLEU scores near to WikiMatrix in all languages, and in the case of Kazakh, the model trained on our data performs higher than WikiMatrix. In other words, in the case of having very noisy comparable data, as is the case for Kazakh-English, our model even outperforms a contextualized supervised model. It is also interesting to see that our model outperforms the supervised model for Kazakh that has only 7.7K gold-standard training data. These are all strong evidences of the strength of our approach in truly low-resource settings.

E Pretraining Matters

It is a truth universally acknowledged, that a single model in possession of a small training data and high learning capacity, must be in want of a pre-trained model. To prove this, we run our translation experiments with and without pretraining. In this case, all models with the same training data and parameters are equal, but some models are more equal. Figure 8 shows the results on the mined data. Clearly, there is a significant gain by using pre-trained models. For Gujarati, which is our the lowest-resource language in our experiments, the distance is more notable: from BLEU score of 2.9 to 9.0. If we had access to a cluster of high-memory GPUs, we could potentially obtain even higher results throughout all of our experiments. Therefore, we believe that part of the blame for our results in English-Romanian is on pretraining. As we see in Figure 7, our supervised results without back-translation are also low for English-Romanian.

F Comparing to CRISS

The recent work of Tran et al. (2020) shows impressive gains using high-quality pretrained models and iterative parallel data mining from a larger comparable data than that of Wikipedia. Their pretrained model is trained using 256 Nvidia V100 GPUs in approximately 2.5 weeks (Liu et al., 2020). Figure 9 shows that by considering all these facts, our model still outperforms their *supervised* model in English-to-Kazakh with a big margin (4.3 cs 10.8) and gets close to their performance in other directions. We should emphasize on the fact that Tran et al. (2020) explores a much bigger comparable data than ours. One clear addition to our work is exploring parallel data from other available comparable datasets. Due to limited computational resources, we skip this part but we do believe that using our current unsupervised models can help extract even more high-quality parallel data from comparable datasets, and this might lead to further gains for low-resource languages.