Towards Machine Translation from Monolingual Texts

Thesis Proposal

Ivana Kvapilíková
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
kvapilikova@ufal.mff.cuni.cz

Abstract

The current state of the art in machine translation heavily relies on parallel data, i.e. texts that have been pre-translated by humans. This type of resource is expensive and only available for several language pairs in limited domains. A new line of research has emerged to design models which would learn to translate from abundant monolingual texts. This thesis proposal provides an overview of unsupervised techniques for machine translation and shows directions for future research primarily in the area of cross-lingual representation learning.

1 Introduction

Modern machine translation (MT) systems are trained on large parallel corpora, i.e. collections of sentence-aligned text documents translated by humans, ideally professional translators. While there are public sources of parallel data for several widely-spoken languages (e.g. EU legislation, public domain books, movie subtitles), the only parallel corpus available for many other language pairs is the Bible. According to Ethnologue[1], there are 7,111 languages spoken in the world and only a small fraction of them is covered by large parallel data sets, others are considered *low-resource*. This work summarizes and compares different approaches applicable in low-resource settings.

In contrast to the standard MT, unsupervised MT models are trained without any parallel documents, but rather use large monolingual corpora to learn the structure of each language separately. Since monolingual texts are significantly easier to obtain (e.g. crawl from the web) than parallel texts, unsupervised techniques are of particular significance for low-resource language pairs. Alternatively, parallel corpus (bitext) mining can be used to expand existing data resources by finding parallel sentences in comparable corpora (e.g. Wikipedia) and train an MT system in a supervised fashion even for low-resource languages.

There are two main directions this thesis will explore: unsupervised machine translation and bitext mining for low-resource languages. The underlying problem behind both of these tasks is unsupervised cross-lingual representation learning. We encode the input text into a cross-lingual latent space and we aim to either search this space for close sentences (bitext mining) or decode into a different language (unsupervised MT), both without using translation parallel resources for training. We will focus on various techniques to induce such cross-lingual space and enhance the alignment of parallel word and sentence representations. We will explore the effect of multilingual training on the quality of the representations and on the performance of unsupervised MT systems.

Section 2 of this proposal gives an overview of the existing work related to machine translation from monolingual texts. Section 3 expands on the methods used in our experiments. Section 4 summarizes the experiments we have conducted so far and Section 5 introduces our research plan for the future.

2 Related Work

2.1 Unsupervised Machine Translation

Unsupervised machine translation was pioneered by Artetxe et al. (2018c,b) and Lample et al. (2018b). They proposed unsupervised training techniques for both the phrase-based statistical machine translation (SMT) model and the neural machine translation (NMT) model to extract all necessary translation information from monolingual data. For the SMT model (Lample et al.,

[1] https://www.ethnologue.com/guides/how-many-languages

2018b; Artetxe et al., 2018b), the phrase table is initialized with an unsupervised n-gram embedding mapping. For the NMT model (Lample et al., 2018b; Artetxe et al., 2018c), the system is designed with a shared encoder and it is trained on batches of synthetic sentence pairs generated on-the-fly by auto-encoding (He et al., 2016) and by back-translation (Sennrich et al., 2016). Conneau and Lample (2019) obtain state-of-the art results in unsupervised MT by pretraining the encoder and the decoder with a masked language model objective (Devlin et al., 2018). Artetxe et al. (2019) obtain similar results by initializing an unsupervised NMT model by an SMT system and jointly refining both systems by back-translation.

Garcia et al. (2020) explore the multilingual view on unsupervised MT and propose a novel cross-translation loss term utilizing not only monolingual data but also an auxiliary parallel data set for a related language pair. They show that adding one more language to the training framework can lead to improvements in BLEU scores over state-of-the-art unsupervised models.

2.2 Multilingual Machine Translation

There have been successful attempts to jointly train multilingual translation systems on training data from several language pairs, either with full parameter sharing (Ha et al., 2016; Johnson et al., 2017; Aharoni et al., 2019), with language-specific encoders and decoders relying on shared attention (Firat et al., 2016) or an attention bridge (Vázquez et al., 2019). The results show that multilingual models yield comparable or even superior results to the standard bilingual setup and are capable of *zero-shot* translation between unseen language pairs, which demonstrates their ability of abstraction and transfer learning. Although the zero-shot results lag behind the more conventional pivot translation through a third language, Gu et al. (2019) reach competitive results by using encoder pretraining and back-translation. Pham et al. (2019) introduce a regularization term to enforce decoder-level similarity between true and auto-encoded target sentences. The ability of multilingual models to learn language-independent representations is especially relevant to this work.

2.3 Unsupervised Cross-lingual Representation Models

The BERT model by Devlin et al. (2018) introduced a new paradigm into the NLP research, leveraging large amounts of existing text to train universal representations exploitable in various downstream tasks. It initiated an entire family of language models (Liu et al., 2019; Zhang et al., 2019; Dai et al., 2019) which learn contextualized word embeddings through Transformer self-attention (Vaswani et al., 2017).

Aside from the vanilla BERT model of the English language, Devlin et al. (2018) released a multilingual model (M-BERT) trained on non-aligned Wikipedia dumps in 104 languages. Similarly, Conneau and Lample (2019) trained another Transformer-based multilingual model on 100 languages and called it XLM-100, later publishing an even larger model XLM-R (Conneau et al., 2019). The architecture of multilingual models is identical to their monolingual counterparts and relies on the same masked LM training objective, but the training data consists of streams of sentences in different languages. Although there is no cross-lingual training objective and no explicit alignment, the limited capacity of the model forces it to generalize and learn multi-lingual abstractions.

Several authors (Pires et al., 2019; Karthikeyan et al., 2019; Libovický et al., 2019) analyzed how multilingual are the representations learned by these models. Although their training does not require any parallel data, M-BERT and XLM prove surprisingly effective at cross-lingual knowledge transfer in NLP tasks such as cross-lingual natural language inference (XNLI)[2] or named entity recognition (XNER) (Pires et al., 2019). Furthermore, Conneau and Lample (2019) reach state-of-the-art performance on both XNLI and unsupervised MT when using a pretrained XLM model for initialization. While Pires et al. (2019) suspected that the cross-lingual ability of M-BERT is linked to the lexical overlap between related languages, Karthikeyan et al. (2019) show that the transfer exists even for languages with different alphabets and with no lexical overlap at all, suggesting that the cross-lingual ability arises rather due to some structural similarities of the languages.

2.4 Cross-lingual Word Embeddings

Cross-lingual **static word embeddings** can be obtained by post-hoc alignment (Mikolov et al., 2013b) of monolingual word embeddings such as `Word2Vec` (Mikolov et al., 2013c) or

[2] https://www.nyu.edu/projects/bowman/xnli/

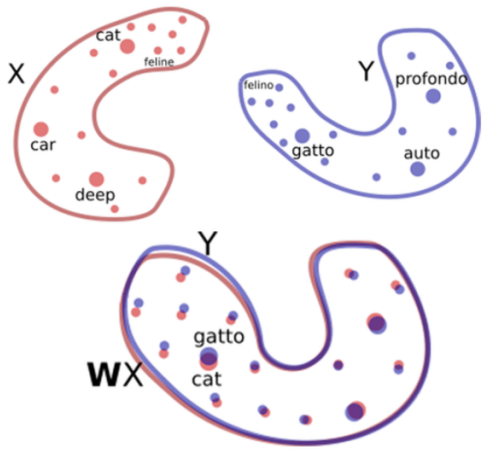

Figure 1: An illustration of mapping monolingual embeddings to a common cross-lingual space.

Source: Conneau et al. (2018a)

`fastText` (Bojanowski et al., 2017), relying on the assumption of isomorphic embedding spaces as illustrated in Figure 1. Aside from a range of supervised methods to learn the mapping matrix, some approaches are completely unsupervised. Zhang et al. (2017) and Conneau et al. (2018a) have inferred a bilingual dictionary in an unsupervised way by aligning monolingual embedding spaces through adversarial training. Artetxe et al. (2018a) propose an alternative method of mapping monolingual embeddings to a shared space by exploiting their structural similarity and iteratively improving the mapping through self-learning. Mohiuddin et al. (2020) propose a semi-supervised method for non-linear mapping in the latent space of two independently trained auto-encoders which even allows them to depart from the questionable assumption that embedding spaces are isomorphic.

Schuster et al. (2019); Wang et al. (2019b) derive **contextualized word embeddings** from masked language models and use the mapping approach to project them into the multilingual space, reaching favorable results on the task of dependency parsing. Other authors improve the alignment of representations in a multilingual LM using a parallel corpus as an anchor (Cao et al., 2020) or using iterative self-learning (Wang et al., 2019a).

2.5 Parallel Corpus Mining

The state-of-the-art approaches to parallel corpus mining are based on similarity retrieval of sentence embedding vectors using a margin based scoring of translation candidates (Artetxe and Schwenk, 2019a). Most models rely on heavy supervision by parallel corpora for the embedding.

Schwenk and Douze (2017); Schwenk (2018); Espana-Bonet et al. (2017) derive sentence embeddings from internal representations of a neural machine translation system with a shared encoder. The top performance in parallel data mining is currently achieved by LASER (Artetxe and Schwenk, 2019b), a multilingual BiLSTM model sharing a single encoder for 93 languages trained on parallel corpora to produce language agnostic sentence representations. LASER has been successfully used to mine billions of sentence pairs from the web (Schwenk et al., 2019).

The universal sentence encoder (USE) (Cer et al., 2018; Yang et al., 2019a) family covers sentence embedding models with a multi-task dual-encoder training framework including the tasks of question-answer prediction or natural language inference. Guo et al. (2018) directly optimize the cosine similarity between the source and target sentences using a bidirectional dual-encoder. Yang et al. (2019b) enhance the model with an *additive margin softmax* loss to separate translations from nearby non-translations.

An entirely different (and possibly unsupervised) approach is to construct sentence representations by aggregating cross-lingual word embeddings either by simple averaging (Arora et al., 2017) or using an IDF weighted average (Litschko et al., 2019). However, since the mapping is applied to static (non-contextualized) embeddings, this strategy gives up on the contextual information which could be exploited in the sentence representation construction.

3 Methodology

This section introduces methodological concepts which will be used in our experiments

3.1 Cross-lingual Pretraining

The goal of unsupervised pretraining is to use abundant unlabeled data to learn a general structure of text. Specifically, language models learn deep bidirectional representations which carry information on each word token and its context.

We will be working with Transformer (Vaswani et al., 2017) encoders trained on a concatenation of monolingual corpora to learn a joint structure of multiple languages. All languages are processed with the same shared vocabulary gener-

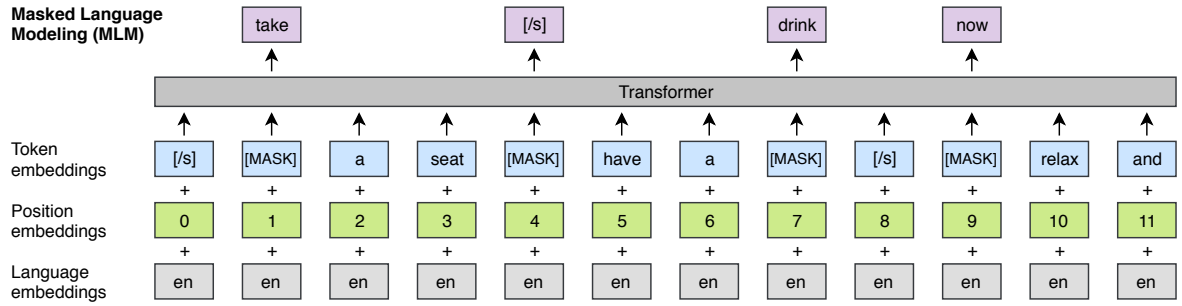

Figure 2: Cross-lingual language model design for training with the masked language modeling (MLM) objective. *Source: Conneau and Lample (2019)*

ated by Byte Pair Encoding (BPE) (Sennrich et al., 2016) from a concatenation of fair samples of the training corpora to not create a bias towards high-resource languages. Possible training objectives include the masked language modeling loss and the translation language modeling loss described below.

The **masked language model (MLM)** training objective facilitates learning of a bidirectional context of words (Devlin et al., 2018) . Random tokens of a word sequence are masked and the model is trained to fill in the missing tokens given the context, as illustrated in Figure 2. During MLM training, 15% of tokens are randomly sampled to be either replaced by the [MASK] token ($p = 0.8$), replaced by a random token ($p = 0.1$) or not changed at all ($p = 0.1$). In contrast to a causal (left-to-right) language modeling objective, MLM allows the model to see the context from both sides of the predicted word. Conneau and Lample (2019) show that the MLM objective is superior to the causal LM objective in cross-lingual transfer.

When parallel data is available, it can be leveraged in training of the multilingual language model using a **translation language model (TLM)** loss. Pairs of source and target sentences are concatenated, random tokens are masked from both sentences (independently of each other) and the model is trained to fill in the blanks by attending to any of the words of the two sentences. The Transformer self-attention layers are thus free to enrich word representations with the information about their monolingual context as well as their translation counterparts. This explicit cross-lingual training objective further enhances the alignment of the internal representations of the model in the latent cross-lingual space.

3.2 Cross-lingual Embeddings

Encoder hidden states extracted from masked language models are sometimes called *contextualized word embeddings* because they not only carry information about the usual context of each word, but they also change according to the context the word appears in (i.e. are contextualized). These representations can be extracted from any of the model layers and experiments show that different encoder layers represent different linguistic phenomena (Jawahar et al., 2019). Furthermore, when we derive contextualized embeddings from a multilingual model, they already exhibit some cross-lingual properties (Pires et al., 2019). In our experiments, we will attempt to better align these representations to make them fully language agnostic wherever possible.

Static word embeddings are vector representations of words with favorable properties which can be learned using the CBOW (Mikolov et al., 2013c), skipgram (Mikolov et al., 2013a; Kocmi and Bojar, 2016) of GloVe (Pennington et al., 2014) algorithms. These representations are learned from monolingual corpora for each language separately and we can assume that monolingual embedding spaces have similar geometric structures across languages, i.e. are approximately isomorphic, and there exists a linear mapping between them (Mikolov et al., 2013b).[3] Given this assumption, finding the mapping W can be viewed the Procrustes problem (Hurley and Cattell, 1962) which has a closed-form solution given by singular value decomposition (SVD)

$$W^* = \underset{W}{\operatorname{argmin}} ||WX - Y||_F = UV^T \qquad (1)$$

[3]We discuss later in Section 5 that this assumption is not always met.

where $U\Sigma V^T = \text{SVD}(YX^T)$, X and Y are the $(n \times dim)$ matrices of source embeddings $x_1, ..., x_n$ and target embeddings $y_1, ..., y_n$.

In our experiments, we follow Conneau et al. (2018b) (MUSE) and learn an initial proxy of W by adversarial learning in a training framework proposed by Ganin et al. (2017). A discriminator is trained to discriminate between elements randomly sampled from $\{Wx_1, ..., Wx_n\}$ and $\{y_1, ..., y_n\}$ while W is trained to prevent the discriminator from making accurate predictions. Then, we iteratively improve the solution by using the words that match the best as anchor points for Procrustes. More details about the mapping algorithm are given in Conneau et al. (2018b). A similar approach by Artetxe et al. (2018a) (VecMap) initializes the mapping by exploiting structural similarity of embedding spaces and matching similarity matrices of monolingual embeddings.

Our experiments require **searching the cross-lingual embedding space** for translation candidates which can be done using the nearest neighbor search with a cosine similarity metric. However, the margin-based approach of Artetxe and Schwenk (2019a) was proved to yield superior results because it eliminates the hubness problem caused by words which are extra-ordinarily close to many other words. The score relies on cosine similarity to measure the distance between sentences but interprets it in relative terms to the average cosine similarity between the two sentences and their nearest neighbors

$$\text{score}(x, y) = \frac{\cos(x, y)}{\sum_{z \in \text{NN}_k(x)} \frac{\cos(x,z)}{2k} + \sum_{z \in \text{NN}_k(y)} \frac{\cos(z,y)}{2k}} \quad (2)$$

where x and y are the source and target sentences, and $NNk(x)$ denotes the k nearest neighbors of x in the other language.

3.3 Unsupervised Statistical Machine Translation

Unsupervised SMT is a log-linear model (Koehn et al., 2003) consisting of a phrase table, language model, distortion model and word penalties. When only monolingual data is available, we can still estimate the language model without any limitation, as it only depends on monolingual data. We can also calculate the penalties, which are parameterless and we may discard the distortion model. The modification of the MERT (Och, 2003) model do tune the weights of the log-linear model is described in Artetxe et al. (2018b).

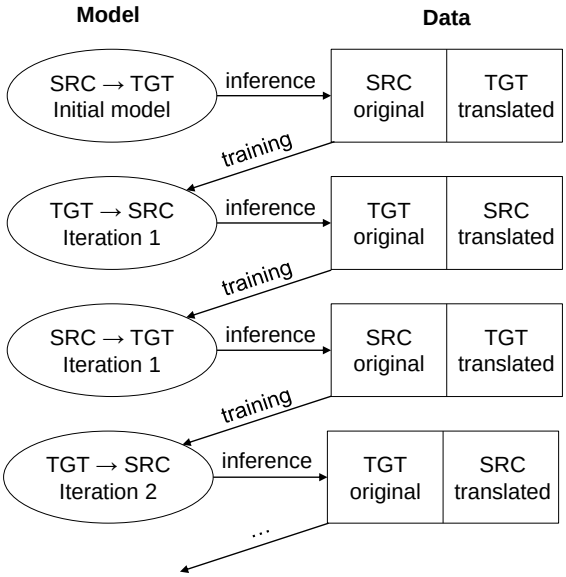

Figure 3: Step-by-step illustration of iterative back-translation.

Source: Kvapilikova et al. (2019)

The key element of an unsupervised SMT is populating the initial **phrase table** with translation pairs derived from cross-lingual word embeddings pretrained as described in Section 3.2. The translation probabilities for each translation candidate pair (e, f) are estimated from cosine distances according to the Equation (3)

$$\phi(f, e) = \frac{\cos(e, f)/\tau}{\sum_{f\prime} \cos(e, f\prime)/\tau} \quad (3)$$

where the temperature parameter τ is tuned as described in Artetxe et al. (2018b).

An essential concept in unsupervised MT is a data augmentation method called **back-translation** (Sennrich et al., 2016). Seed MT models for both translation directions are used to back-translate both monolingual corpora and generate two synthetic parallel corpora. At that point the existing models can be discarded and new ones are estimated from scratch using the synthetic corpora for supervision. This procedure can be repeated several times, creating synthetic corpora of increasing quality (Artetxe et al., 2018b). The procedure is illustrated in Figure 3.

3.4 Unsupervised Neural Machine Translation (UNMT)

Our unsupervised NMT models have a Transformer encoder-decoder architecture. Both the en-

coder and the decoder are shared across languages.

Model initialization is important to introduce the initial cross-lingual signal. Lample et al. (2018b) initialize the embedding layer of their unsupervised NMT model with pretrained cross-lingual word embeddings. We follow Conneau and Lample (2019) who take this idea even further by pretraining the entire encoder and decoder as described in Section 3.1.

We use the following three optimization objectives to fine-tune the initialized model for machine translation. When only monolingual data is available, we train the model iteratively on auto-encoding and back-translation. When parallel data is available, we use the supervised MT objective and back-translation.

When parallel data is available, NMT models are trained on human translation examples using a standard supervised **machine translation (MT)** objective, minimizing the cross-entropy loss function which measures their ability to predict each following target word correctly. The model $(\theta_{enc}, \theta_{dec})$ is penalized every time the predicted token is not the correct one following the loss function

$$L_{MT}(\theta_{\text{enc}}, \theta_{\text{dec}}, l) = E_{(x,y)\sim D, \hat{y}\sim \text{dec}(\text{enc}(x))}(\Delta(\hat{y}, y)) \quad (4)$$

where (x, y) is a sentence pair sampled from the parallel data set D and Δ is the sum of token-level cross-entropy losses.

Denoising auto-encoding (AE) is a monolingual training objective proposed by Lample et al. (2018a) and Artetxe et al. (2018c) to teach the unsupervised model to recover proper sentences from corrupted input. It is especially important in the beginning of the training when there is not enough cross-lingual information for actual inter-language translation. The model for each language l is trained by minimizing

$$L_{AE}(\theta_{\text{enc}}, \theta_{\text{dec}}, l) = E_{x\sim D_l, \hat{x}\sim \text{dec}(\text{enc}(C(x)))}(\Delta(\hat{x}, x)) \quad (5)$$

where x is a sentence sampled from the monolingual data set D_l and $\hat{x}$ is the reconstructed sentence decoded from the noised version of x. The noise process $C(x)$ introduces random noise to a sentence x by dropping words with a probability p_{drop} and shuffling words within a tunable window size.

Back-translation (BT) is a bilingual objective for training an unsupervised model on synthetic translation samples generated by the model itself in previous iterations. This procedure is crucial for unsupervised NMT where we do not have any authentic parallel data available at all. Back-translation is happening *on-the-fly* during training where the model first generates a batch of synthetic parallel data and immediately trains itself on it.

In the back-translation step, the model is first set to the inference mode and used to translate a batch of sentences. The synthetic translations serve as source sentences for a training step where the target side is the original sentence.

$$L_{BT}(\theta_{\text{enc}}, \theta_{\text{dec}}, l) = E_{x\sim D_l, \hat{x}\sim \text{dec}(\text{enc}(T(x)))}(\Delta(\hat{x}, x)) \quad (6)$$

where $T(x)$ is the translation model itself which generates a synthetic translation of a sentence x.

3.5 Evaluation

Machine translation can be evaluated automatically by the BLEU score introduced by Papineni et al. (2002). The metric compares the candidate translation against the reference translation and assigns a score, depending on the number of overlapping n-grams and the overall sentence length. Despite its limitations (mostly due to the large number of ways one can translate a sentence into another language), BLEU has demonstrated a sufficient correlation with human judgment and is widely used to compare results of MT research on standardized WMT[4] test sets. For our final MT models, we will also use manual evaluation where human judges will assess the translations based on their fluency and adequacy. Finally, we will design dedicated measures targeting particular problems of unsupervised MT (e.g. preservation of named entities).

Word translation is an evaluation task applied to measure the quality of cross-lingual word embeddings. We use evaluation dictionaries from the MUSE library (Conneau et al., 2018b) and calculate how many times one of the correct translations of a source word is retrieved, measuring precision@k for k = 1, 5, 10.

Parallel corpus mining can be evaluated and

[4] http://www.statmt.org/wmt19/

compared to other research groups on the BUCC[5] shared task where the system is expected to search two comparable non-aligned corpora and identify pairs of parallel sentences. Precision, recall and the F1 score are calculated based on the gold list of sentence pairs.

Parallel sentence matching is an auxiliary task similar to parallel corpus mining where the model is trying to match correct translations in a pool of shuffled parallel sentences. We use it to evaluate the multilinguality of sentence representations. The Tatoeba (Artetxe and Schwenk, 2019b) collection includes test sets for over 100 languages.

4 Experiments

In this section, we present both our published and unpublished experimental results and we outline the research we plan for the future.

4.1 Combining Statistical and Neural Unsupervised MT

We performed several experiments with training unsupervised machine translation systems described in Section 3 for the German-Czech language pair. The goal was to compare existing approaches in one setting and identify an optimal method for combining models from different families (statistical and neural) together.

Methodology

All neural models in the experiment have an identical Transformer architecture with 6 encoder layers, 1024 hidden units and 8 attention heads. We use a shared subword vocabulary for both the source and the target language. The BPE segmentation is learned from a concatenation of the two corpora with a target vocabulary size of 60,000.

We pretrain one cross-lingual MLM model to initialize the encoder and the decoder of all neural models. We fine-tune the models for machine translation using the BT and AE objectives (when training only on monolingual data – UNMT model) or BT and MT objectives (when synthetic translations by the USMT model are available – hybrid model). For the final experiment (hybrid+UNMT), we train until convergence on the synthetic data set and only then switch to the monolingual data sets and train using the BT+AE objectives. We train all models on 8 GPUs with a batch size of 2400 tokens per GPU.

The statistical USMT model was trained using the Monoses toolkit. The phrase table was populated with 100 nearest neighbors of each source word in the shared embedding space which was generated by the VecMap alignment method. The underlying 300-dimensional skipgram Word2Vec embedding model was applied to 1M most frequent word unigrams, bigrams and trigrams. More details about the experiment are given in Kvapilikova (2020).

Benchmarks

Since German-Czech is not a de-facto low-resource language pair, we can compare the performance of our unsupervised model to a pivoting and a supervised benchmark. For the **pivoting benchmark**, we use English as the pivot and we train two supervised Transformer-based NMT models initialized with a pretrained MLM model with the same hyperparameters that were used for the unsupervised experiments. To generate final translations between German and Czech, we pass each source sentence through both of the models in sequence. For the **supervised benchmark**, we train a supervised NMT model on authentic parallel data, also using the same pretrained model and hyperparameters. The benchmark models are trained using the BT and MT objectives.

Data

Monolingual training data was obtained from NewsCrawl.[6] We used WMT *newstest2013* for validation and *newstest2019* for testing.

For training the supervised benchmark model, we used the following Czech-German parallel corpora available at the OPUS[7] website: OpenSubtitles, MultiParaCrawl, Europarl, EUBookshop, DGT, EMEA and JRC.

For training the pivoting Czech-English-German model, we used the CzEng 1.6 corpus of Czech-English parallel data and the Europarl, EUBookshop and OpenSubtitle corpora of English-German parallel data. The amount of data used for each model is indicated in Table 1.

Results

The results are summarized in Table 1. They reveal that combining features of both statistical and neural modeling has a positive complementary effect on translation quality. To get the most out of

[5] https://comparable.limsi.fr/bucc2018/bucc2018-task.html

[6] http://data.statmt.org/news-crawl/

[7] http://opus.nlpl.eu/

Model Type	BLEU de→cs	BLEU cs→de	Supervision	Training Data
USMT	11.72	12.39	none	26M sent. per lang.
UNMT	15.93	15.79	none	26M sent. per lang.
Hybrid	13.71	-	none	26M synthetic sent. pairs
Hybrid+UNMT	**17.4**	**20.14**	none	26M synthetic sent. pairs
Pivoting	16.50	17.46	2 bitexts	2 x 26M sent. pairs
Supervised	**20.83**	**21.03**	bitext	22M sent. pairs

Table 1: Translation quality measured by BLEU scores on newstest2019. The synthetic training data for the hybrid models was obtained by translating the Czech monolingual corpus into German by the USMT model. The pivoting benchmark was via English.

the hybrid setting, it is optimal to first train the model on the synthetic corpus until convergence and then continue training on the monolingual data by auto-encoding and on-the-fly back-transtaion. Such a model almost reaches the results of a supervised model trained on millions of parallel sentences with the identical architecture.

The benchmark systems do not directly compete with the unsupervised systems since they have higher data requirements (parallel Czech-German data for the supervised benchmark and parallel Czech-English and English-German data for the pivoting benchmark) which are not satisfied for low-resource languages by their definition. However, since Czech and German allow making this comparison, we see that the performance of our unsupervised models surpasses the pivoting benchmark and gets close to the supervised benchmark. It must be noted that both baselines were trained mostly on out-of-domain data (movie subtitles, EU legislation) which might be detrimental to their performance on a test set composed of newspaper articles. Nevertheless, the results suggest that the gap between unsupervised and supervised techniques is narrowing, especially for language pairs with insufficient in-domain parallel training data.

4.2 Contextualized Word Representations from Multilingual Models

In our second experiment, we explore the multilinguality of a large pretrained language model XLM-100[8] by assessing its representations on a task of parallel sentence matching (PSM). Since the model is trained in a completely unsupervised way, any evidence of cross-lingual transfer is surprising. We dissect the model to assess how much cross-lingual information is hidden in its internal representations on different layers and select which layer outputs the most multilingual representations. We use the findings from this experiment when setting hyperparameters in further experiments.

Data

The XLM model was pretrained on the Wikipedia corpus of 100 languages (Conneau et al., 2019). We evaluate the pairwise matching accuracy on a multi-way parallel data set of 3k sentences in 6 languages.[9] We use WMT *newstest2012* for development and *newstest2013* for testing.

Methodology

Aggregating subword embeddings to fixed-length sentence representations necessarily leads to an information loss. We derive sentence embeddings from subword representations by simple element-wise averaging. Even though mean-pooling is a naive approach to subword aggregation, it is often used for its simplicity (Reimers and Gurevych, 2019; Ruiter et al., 2019; Ma et al., 2019) and in our scenario it yields better results than max-pooling. Cosine similarity is used for the nearest neighbor search in the multilingual sentence embedding space.

Results

We derive sentence embeddings from all layers of the model and show PSM results on the development set averaged over all language pairs in Figure 4. The accuracy differs substantially across the model depth, the best cross-lingual performance is consistently achieved around the 12th (5th-to-last) layer of the model. The results show that the model is able to match correct translations in 85% of cases on average. It must be noted that the measurement was performed for high-resource languages which are all very well represented in

[8] https://github.com/facebookresearch/XLM.

[9] Czech, English, French, German, Russian, Spanish

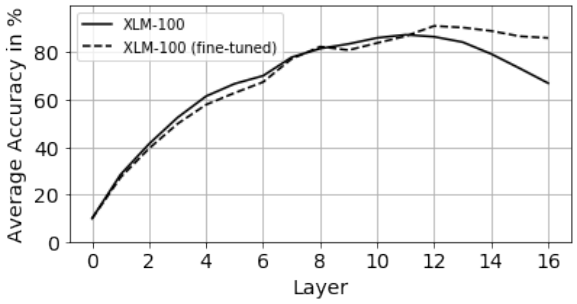

Figure 4: PSM accuracy of sentence embeddings on *newstest2012* from the input embedding layer (0th) to the deepest layer (16th). XLM-100 is the original pretrained model and XLM-100 (fine-tuned) was fine-tuned as described in Section 4.3.

the XLM pretraining corpus. The matching accuracy for low resource languages is significantly lower and will be targeted in future experiments.

4.3 Improving Multilingual Representations with a Translation Objective

We propose a method to enhance the cross-lingual ability of a pretrained multilingual model by fine-tuning it on a small synthetic parallel corpus (Kvapilikova et al., 2020). The parallel corpus is obtained via unsupervised machine translation (MT) so the method remains unsupervised.

Data

The monolingual English and German data was randomly selected from NewsCrawl 2018. Evaluation was performed on the BUCC data sets. The Chinese-Czech and English-Kazakh evaluation data sets were compiled by shuffling parallel sentences from NewsCommentary or WMT newstests into monolingual sentences from NewsCrawl in a 1:40 ratio (similarly to BUCC).

Methodology

We train an unsupervised English-German model using MLM pretraining and AE+BT fine-tuning and use it to create a synthetic parallel English-German corpus. Then we fine-tune the pretrained XLM-100 model on the synthetic data set and measure the effect on its internal representations.

We apply the fine-tuned model on the task of parallel sentence mining. We encode each sentence of the mining dataset and mean-pool the contextualized encoder representations from the 12th[10] layer to fixed-length sentence embeddings. We use the margin-based approach by Artetxe and Schwenk (2019a) when searching the multilingual space for translation candidates. We measure the performance of our method on the task of parallel corpus mining using the BUCC datasets. The threshold was tuned on the training part of the BUCC data sets.

Benchmarks

We compare our proposed model to a **vanilla XLM** baseline where contextualized token representations are extracted from the 12th layer of the original XLM-100 model and mean-pooled into sentence embeddings.

For a **word mapping** baseline we use Word2Vec embeddings with 300 dimensions pretrained on NewsCrawl and map them into the cross-lingual space using the unsupervised version of VecMap (Artetxe et al., 2018a). As above, word embeddings are aggregated by simple mean-pooling to represent sentences.

Results

The results in Table 2 reveal that TLM fine-tuning on only 20k synthetic sentence pairs brings a substantial improvement over the initial pretrained model (*vanilla XLM*). In terms of the F1 score, the gain across four BUCC language pairs is 15.0 - 21.6 points. Even though the fine-tuning focused on a single language pair (English-German), the improvement is notable for all evaluated language pairs. The largest margin of 21.6 points is observed for the English-Russian mining task. We observe that using a small parallel data set of authentic translation pairs instead of synthetic ones does not have a significant effect.

The weak results of the *word mapping* baseline can be attributed to the superiority of contextualized embeddings for representation of sentences over static ones. Even if the contextualized embeddings were effectively limited to a very near context, they still have the capacity of correctly representing multi-word expressions which can greatly help in sentence matching.

Although the performance of our model lags far behind the supervised LASER benchmark, it is valuable because of its fully unsupervised nature and it works even for distant languages such as Chinese-Czech.

In comparison with the *word mapping* baseline, the results of our method are less sensitive to the relatedness of the languages. They are,

[10]The layer hyperparameter as well as the pooling method were tuned on the PSM task described in Section 4.2

	en-de	en-fr	en-ru	en-zh	zh-cs	en-kk	Supervision	
Leong et al. (2018)	-	-	-	56.00	-	-	bitext	0.5M sent.
Bouamor and Sajjad (2018)	-	76.00	-	-	-	-	bitext	2M sent.
Schwenk (2018)	76.90	75.80	73.80	71.60	-	-	9-way parallel	2M sent.
Azpeitia et al. (2018)	85.52	81.47	81.30	77.45	-	-	bitext	2-9M sent.
Artetxe and Schwenk (2019b)	**96.16**	**93.91**	**93.30**	**92.27**	-	-	2- or 3-way parallel	223M sent.
Baseline (Word Mapping)	32.04	32.94	17.68	20.65	-	-	none	n/a
Baseline (Vanilla XLM)*	62.10	64.77	61.65	44.79	43.00	24.00	none	n/a
Proposed method*	**80.06**	**78.77**	**77.16**	**67.04**	**48.69**	**35.41**	none	20k sent.**

Table 2: F1 score on the parallel sentence mining task. The supervised (upper part) and unsupervised (lower part) winners are highlighted in bold. * The model was pretrained on Wikipedia. ** Synthetic translations produced by unsupervised MT.

however, sensitive to the amount of monolingual sentences in the Wikipedia corpus used for XLM pretraining. Representations of languages with small Wikipedia collections (∼100k sentences, e.g. Nepali, Khmer) are not aligned well enough to perform filtering using *vanilla XLM* or our method.

5 Research Plan

In our future research, we would like to explore two experimental paths. The first relates to multilingual vector representations of text and how to induce them without supervision. It is a conceptual problem which lies behind all attempts to solve cross-lingual tasks using only monolingual data. We will continue with our research in sentence embeddings and their usability in unsupervised parallel sentence mining. The second path concerns directly the task of machine translation for low-resource language pairs. We would like to attempt to improve the state of the art by using auxiliary corpora in other languages via transfer learning. Garcia et al. (2020) already showed a positive effect of including an extra language pair into the unsupervised MT training pipeline.

A possible extension of the parallel corpus mining task is to score back-translation candidates generated throughout the unsupervised NMT training process and use only high quality translations for further training. Ruiter et al. (2019) use a self-supervised approach where the translation model itself is used for generating the sentence embeddings.

5.1 Aligning Representations of Multilingual Models

Motivation

Post-hoc linear mapping of monolingual embeddings to a cross-lingual space were described in 3.2. Since the mapping approaches rely on similar geometric properties (e.g. isomoprhism) of embedding spaces across languages, the alignment can only by as good as the properties allow. Several recent studies (Patra et al., 2019; Ormazabal et al., 2019a) have criticized this simplified approach of linear mapping, showing that even the embedding spaces of closely related languages do not exhibit similar geometric properties. Nakashole and Flauger (2018) argue that the transformation of the embedding space from one language to another can be linear only at small local regions, but not globally. Furthermore, the assumption of isomorphism weakens with decreasing language relatedness and the mapping can even fail completely for very distant languages. In practice, the linear mapping approaches lead to a word translation accuracy (P@1) of up to 84 % for high resource languages (English-Spanish) to as low as 19 % for low-resource languages (English-Tamil) (Artetxe et al., 2018a). Such simplified method can serve to bootstrap an USMT system but it imposes an initial upper bound on the translation performance.

We hypothesize that isomorphism is a property of an embedding model rather than a linguistic limitation and we want to test whether joint training of contextualized representations induces a word embedding space with more convenient geometric properties than skipgram and CBOW embedding models. This view is supported by the study of Ormazabal et al. (2019b) which suggest that divergences across languages can be effectively mitigated by jointly learning their representations.

Methodology

We would like to experiment with pretrained multilingual language models (M-BERT, XLM-100, XLM-R), distil their internal representations into

static cross-lingual embeddings and test whether the induced word embedding spaces suffer from the non-isomorphism as severely as the embedding spaces learned separately. Even though the embeddings derived from multilingual Transformer models already exhibit some level of cross-linguality, we will try to improve their quality either by further fine-tuning the pretrained model or by post-hoc alignment of the extracted embeddings. The former approach offers a possibility of utilizing the entire fine-tuned model for other downstream tasks, e.g. XNLI.

Contextualized word embeddings can be averaged over monolingual corpora to obtain static word embedding spaces in different languages which can be further aligned using one of the existing mapping methods. Alternatively we could weight different contexts differently (e.g. upweight common contexts and downweight rare ones) or explicitly model polysemy by clustering over different contexts.

Evaluation

The performance will be measured directly on a word translation task using precision metrics P@1, P@5 or P@10. Secondly, we will assess the quality of sentence translations (measured by BLEU score) produced by an unsupervised statistical MT model with a phrase table induced from the cross-lingual embeddings. The second evaluation task is more difficult as it also uses word pair distances to estimate their translation probabilities.

The unsupervised baselines for this experiment are the VecMap and MUSE alignment methods for post-hoc mapping of static Word2Vec embeddings.

5.2 Fine-tuning Pretrained Models with the Focus on Low-resource Languages

Motivation

Our experiment described in Section 4.3 revealed some interesting properties of pretrained multilingual models which should be investigated further. We showed that fine-tuning the pretrained model on one language pair improved the quality of cross-lingual representations of completely unrelated languages. However, the cross-lingual transfer did not generalize well to all languages the model was pretrained on. In particular, very low resource languages like Nepali or Sinhala account for only around 100k training sentences each and their representation are not aligned well enough to allow for cross-lingual transfer.

Methodology

We will experiment with optimizing the performance of the pretrained model for low-resource languages by fine-tuning it with more monolingual data (e.g. Common Crawl) using the MLM objective or with small bilingual (possibly synthetic) data (e.g. Flores dataset) using the TLM objective.

Evaluation

The performance of a fine-tuned model will be measured on several downstream cross-lingual tasks. The quality of its internal representation will be evaluated on the task of **parallel corpus mining** as described in our previous experiment in 4.3.

5.3 Machine Translation with Synthetic Data Scoring

Motivation

Throughout the training unsupervised MT models are constantly learning from synthetic data and the quality of the data improves as the training progresses. Based on our experiments as well as related work by Artetxe et al. (2020), the bottleneck of unsupervised MT is the iterative back-translation which generates synthetic training examples of differing quality on-the-fly. Scoring these translations and only selecting the *good* ones for training could enhance the training.

Methodology

We will encode each synthetic sentence pair we want to score using an unsupervised multilingual model. We will derive a fixed-length representation of every individual sentence and score each pair based on cosine similarity, setting a threshold for *good* and *bad* translations. The threshold could be learned on a data set created from translations of different quality, e.g. obtained in the first (bad) and last (good) iterations of an unsupervised MT system. Alternatively, the scoring could be done by fine-tuning the multilingual model for classification and feeding in both sentences together, also using a data set of bad and good synthetic translations. Since translation quality evolves during training, this approach calls for self-training, similarly to Ruiter et al. (2019). The pitfall is that using such scoring is computationally demanding,

especially since it is supposed to happen on-the-fly during training.

The applicability of this method is conditional on the success of the previous experiments which aim at aligning cross-lingual representations of unsupervised masked language models.

Evaluation

The systems will be evaluated by the BLEU score of translation quality and compared against unsupervised MT models which do not employ synthetic data scoring.

5.4 A Multilingual Approach to Unsupervised MT

A possible path to enhance unsupervised MT leads via exploring the options offered by multilingual training. Since multilingual MT is out of scope of this thesis, we will focus on how multilingual pretraining could help unsupervised MT rather than how unsupervised MT findings can help in multilingual or zero-shot translation.

Methodology

Kocmi and Bojar (2018); Zoph et al. (2016) showed that pretraining a parent model (high-resource language pair) can substantially boost translation performance of a child model (low-resource language pair). They start training the parent and after convergence they switch the parent training corpus for the child corpus, leaving all training parameters and optimizer states intact. The only requirement is a shared subword vocabulary. The results are the best when the target language is the same in the parent and child language pair. However, some transfer happens even for completely unrelated languages.

In the first experiment, we will depart from the completely unsupervised setting and use a parallel corpus for the parent language pair to test whether trivial transfer applies also in the unsupervised training pipeline. The assumption of existence of such auxiliary corpus is realistic even in the low-resource setting. We will use the MLM-initialized neural MT model, pretrain it with a supervised MT objective on a high resource pair containing one of the languages we are interested in and fine-tune on the low-resource language pair using only the BT and AE objectives (no parallel data for the child model).

In the second experiment, we will explore the effect of using a multilingual encoder. Normally we initialize the encoder and decoder of the UNMT model with a bilingual MLM model pretrained on the two languages in question. We want to explore the effect of multilingual pretraining of the MLM model using monolingual corpora in several languages on the initialization of the MT model. Tiedemann (2018) suggests that massively multilingual pretraining leads to more language agnostic representations. We want to test whether a multilingual encoder will have a positive effect on translation quality over a more specialised bilingual encoder.

Evaluation

The systems will be evaluated by the BLEU score of translation quality and compared against a baseline NMT model initialized with a bilingual MLM and trained only on the child language pair using the BT and AE objectives.

6 Conclusion

In this proposal we described the current state of the art for unsupervised MT, outlined the outstanding problems and our plan to tackle some of them. We also gave an overview of our previous experiments where we proposed an unsupervised method for parallel corpus mining. We also trained several unsupervised MT models, compared their performance and found a positive effect of pretraining a neural model on a synthetic parallel corpus generated by an unsupervised statistical MT model.

In our future research we would like to explore two research paths: unsupervised multilingual representations for parallel corpus mining and enhancing unsupervised MT models with transfer learning techniques.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual 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 3874–3884.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 789–798.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on EMNLP*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Mikel Artetxe, Gorka Labaka, Noe Casas, and Eneko Agirre. 2020. Do all roads lead to rome? understanding the role of initialization in iterative back-translation.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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

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.

Houda Bouamor and Hassan Sajjad. 2018. H2@bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

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

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv [e-Print archive]*, abs/1810.04805.

Cristina Espana-Bonet, Adam Csaba Varga, Alberto Barron-Cedeno, and Josef van Genabith. 2017. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2017. Domain-adversarial training of neural networks. *Advances in Computer Vision and Pattern Recognition*, page 189–209.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur P. Parikh. 2020. A multilingual view of unsupervised machine translation.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268.

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.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 820–828. Curran Associates, Inc.

John R. Hurley and Raymond B. Cattell. 1962. The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 7(2):258–262.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistic*, 5:339–351.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual BERT: An empirical study. *arXiv [e-Print archive]*, abs/1912.07840.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252.

Tom Kocmi and Ondřej Bojar. 2016. Subgram: Extending skip-gram word representation with substrings. *Lecture Notes in Computer Science*, page 182–189.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the ACL on Human Language Technology*, NAACL '03, pages 48–54.

Ivana Kvapilikova. 2020. *Unsupervised Machine Translation between Czech and German Language*. Ph.D. thesis, Czech Technical University in Prague, Faculty of Information Technology.

Ivana Kvapilikova, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondrej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. In print.

Ivana Kvapilikova, Dominik Machacek, and Ondrej Bojar. 2019. CUNI systems for the unsupervised news translation task in WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation*, pages 241–248.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on EMNLP*, pages 5039–5049.

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

Jindřich Libovický, Rudolf Rosa, and Alexander M. Fraser. 2019. How language-neutral is multilingual BERT? *CoRR*, abs/1911.03310.

Robert Litschko, Goran Glavaš, Ivan Vulic, and Laura Dietz. 2019. Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1109–1112.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal text representation from bert: An empirical study. *CoRR*, abs/1910.07973.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. 2020. LNMap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space.

Ndapa Nakashole and Raphael Flauger. 2018. Characterizing departures from linearity in word translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 221–227.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019a. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019b. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995.

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

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on EMNLP*, pages 1532–1543.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation*, pages 13–23.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

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.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *Proceedings of the 2019 Conference of the North American Chapter of the ACL*.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 228–234.

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

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*, volume 2084 of *CEUR Workshop Proceedings*, pages 188–197.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39.

Shuai Wang, Lei Hou, Juanzi Li, Meihan Tong, and Jiabo Jiang. 2019a. Learning multilingual sentence embeddings from monolingual corpus. In *China National Conference on Chinese Computational Linguistics*, pages 346–357.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019b. Cross-lingual BERT transformation for zero-shot dependency parsing. 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 5720–5726.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Multilingual universal sentence encoder for semantic retrieval.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019b. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1959–1970.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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 EMNLP*, pages 1568–1575.