FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

ABSTRACT OF DOCTORAL THESIS

Ivana Kvapilíková

Towards Machine Translation Based on Monolingual Texts

Institute of Formal and Applied Linguistics

Supervisor of the doctoral thesis: doc. RNDr. Ondřej Bojar, Ph.D.
Study programme: Computer Science
Study branch: Mathematical Linguistics

Prague 2024

The results of this thesis were achieved in the period of a doctoral study at the Faculty of Mathematics and Physics, Charles University in the years 2018–2023.

Doctoral Candidate: Mgr. Ivana Kvapilíková

Supervisor: doc. RNDr. Ondřej Bojar, Ph.D.
Institute of Formal and Applied Linguistics

Department: Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské náměstí 25
118 00 Prague 1, Czech Republic

Opponents: Dr. Cristina España-Bonet
Deutsches Forschungszentrum für
Künstliche Intelligenz GmbH (DFKI)
Stuhlsatzenhausweg 3
66123 Saarland, Germany

RNDr. Martin Čmejrek, Ph.D.
The MAMA AI
Sokolovská 130
186 00 Prague 8, Czech Republic

Chairman of Academic Council: doc. Ing. Zdeněk Žabokrtský, Ph.D.
Institute of Formal and Applied Linguistics

This abstract was distributed on January 26, 2024.
The thesis defense will take place on February 9, 2024 at 9:00 a.m. in front of a committee for thesis defenses in Mathematical Linguistics at the Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, Prague 1, room S4.
The thesis can be viewed at the Study Department of Doctoral Studies of the Faculty of Mathematics and Physics, Charles University, Ke Karlovu 3, Prague 2.

MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

AUTOREFERÁT DOKTORSKÉ PRÁCE

Ivana Kvapilíková

Strojový překlad na základě jednojazyčných textů

Ústav aplikované a formální lingvistiky

Školitel:	doc. RNDr. Ondřej Bojar, Ph.D.
Studijní program:	Informatika
Studijní obor:	Matematická lingvistika

Praha 2024

Disertační práce byla vypracována na základě výsledků získaných během doktorského studia na Matematicko-fyzikální fakultě Univerzity Karlovy v letech 2018–2023.

Doktorand:	Mgr. Ivana Kvapilíková
Školitel:	doc. RNDr. Ondřej Bojar, Ph.D. Ústav aplikované a formální lingvistiky
Školící pracoviště:	Ústav aplikované a formální lingvistiky Matematicko-fyzikální fakulta Univerzita Karlova Malostranské náměstí 25 118 00 Praha 1, Česká republika
Oponenti:	Dr. Cristina España-Bonet Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI) Stuhlsatzenhausweg 3 66123 Saarland, Německo
	RNDr. Martin Čmejrek, Ph.D. The MAMA AI Sokolovská 130 186 00 Praha 8, Česká republika
Předseda RDSO:	doc. Ing. Zdeněk Žabokrtský, Ph.D. Ústav aplikované a formální lingvistiky

Autoreferát byl rozeslán dne 26. ledna 2024.

Obhajoba disertační práce se koná dne 9. února 2024 v 9:00 před komisí pro obhajoby disertačních prací v oboru Matematická lingvistika na Matematicko-fyzikální fakultě UK, Malostranské náměstí 25, Praha 1, v místnosti S4.

S disertační prací je možno se seznámit na studijním oddělení Matematicko-fyzikální fakulty UK, Ke Karlovu 3, Praha 2.

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 dominant languages (e.g. EU legislation, public domain books, movie subtitles), the only parallel corpus available for many other language pairs is the Bible. There are more than 7,000 [Eberhard et al., 2023] languages spoken in the world and only a small fraction of them is covered by large data sets, others are considered low-resource for most natural language processing (NLP) tasks, including MT.

The scarcity of parallel data motivated researchers to devise a new training strategy where the MT model can learn from monolingual texts which are significantly easier to obtain (e.g. by web crawling) than parallel texts. Monolingual corpora were then used in combination with existing parallel resources to increase translation quality and it was an open question whether an MT system can be trained purely from monolingual data.

The problem of learning to translate without ever seeing a translation was first tackled as deciphering [Ravi and Knight, 2011] where foreign text was viewed merely as an unknown cipher of the English text. The idea seemed intriguing but quite unrealistic, until the pioneering work of Artetxe et al. [2018] and Lample et al. [2018a]. It was shown that minimal supervision suffices to teach a neural model to align monolingual word representations (embeddings) and find translation equivalents. Unsupervised training of MT systems became a hot topic both for the curiosity of a seemingly unsolvable task as well as for its relevance for low-resource language pairs.

This thesis investigates unsupervised learning strategies to find the most efficient way to exploit monolingual data for cross-lingual signal. There are two main directions this work explores: (1) methods for obtaining parallel data when authentic parallel resources are unavailable, and (2) unsupervised machine translation (UMT) models, their architecture, and training strategies. The two directions are closely intertwined since UMT models are always trained using a form of synthetic parallel data. Moreover, the underlying problem behind the UMT task as well as the unsupervised parallel corpus mining (PCM) task is the building of a cross-lingual space which we can either use to initialize an MT system or to search for similar sentences. In our analysis, we focus on various techniques to induce the cross-lingual space and enhance the alignment of parallel word and sentence representations. We explore the effect of multilingual training on the quality of the representations and on the performance of UMT systems.

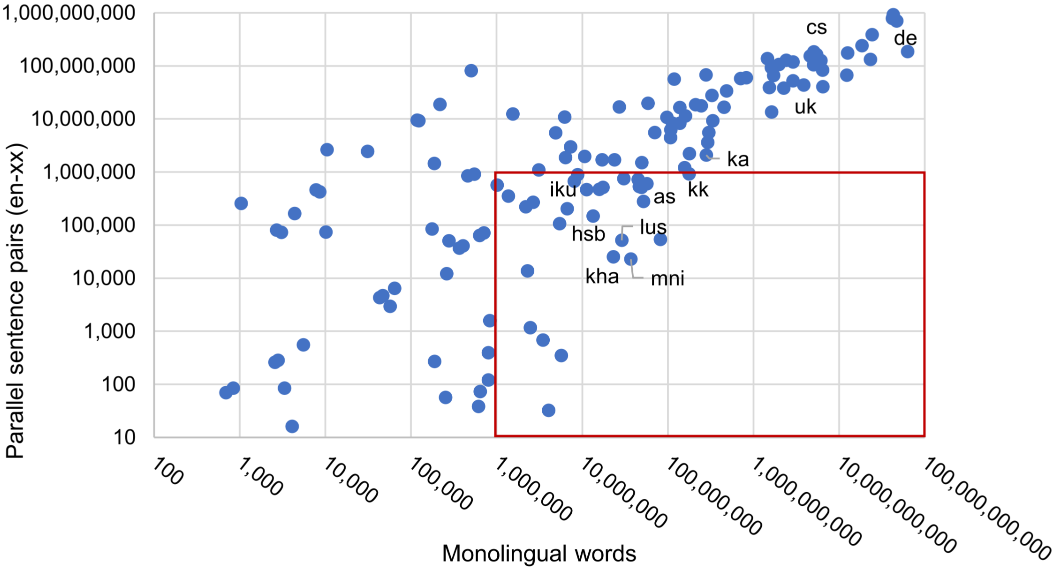

Figure 1.1: World languages plotted in terms of the available textual data – raw monolingual (horizontal axis) and parallel English-aligned (vertical axis). Both axes are in log scale. The rectangle delimits the area of low-resource languages that this thesis focuses on.

1.1 The Extent of This Study

To determine the scope of this work, we need to assess which languages are considered *low-resource* for the task of machine translation and how many such languages there are. We gauge the quantity of parallel data accessible for each language by calculating the number of English-aligned parallel sentences found on the OPUS website, in conjunction with the supplementary corpora provided for the WMT translation shared tasks.[1] As a proxy for the total amount of monolingual data available, we consider the Oscar corpus sizes. It must be noted that both OPUS and Oscar include uncleaned text data with a lot of noise and possible duplicates. We display the languages in terms of their quantities of labeled and unlabeled data in Figure 1.1.

Out of the 151 languages covered by the Oscar corpus, 79 have less than 1M uncleaned parallel sentence pairs, making them suitable candidates for unsupervised training. For the purposes of this work, we call these languages *low-resource*. The threshold of 1M parallel sentences is motivated by Kocmi [2020] who shows that training MT models with fewer sentences leads to fast over-fitting and hindered translation performance. The rectangle in Figure 1.1 delimits the space where unsupervised pre-training techniques are most needed for the lack of parallel data (<1M sentence pairs) and where they are applicable thanks to the availability of monolingual data (>1M words for unsupervised training). The languages to the left of the rectangle can be

[1] https://statmt.org/

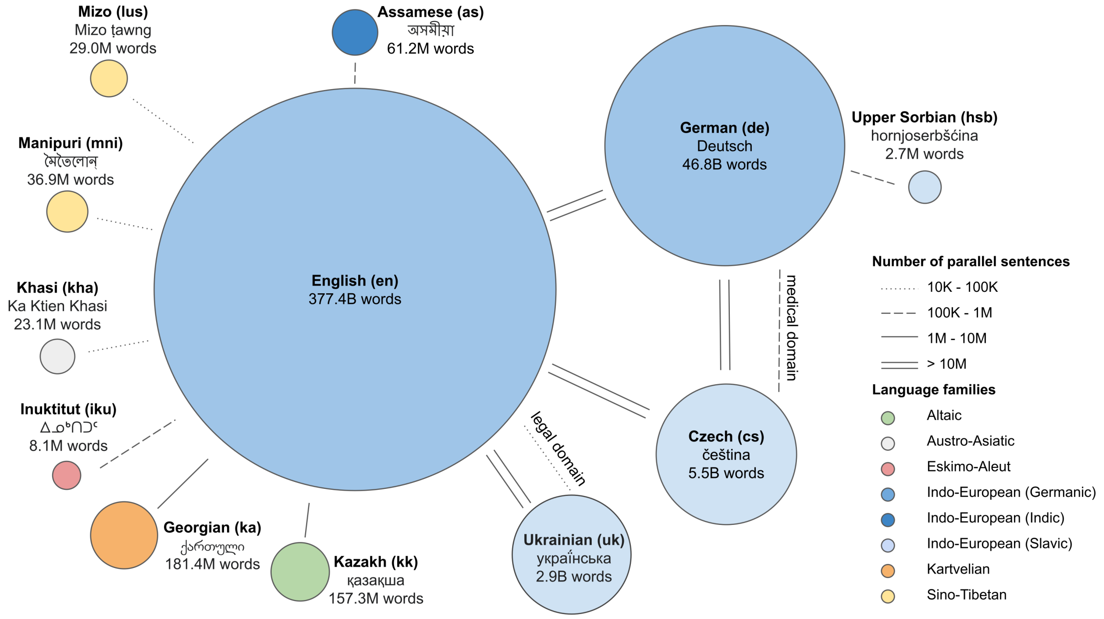

Figure 1.2: Languages used in this thesis in terms of the size of the available monolingual texts. Colors reflect language families and the links between languages represent the amount of parallel data available.

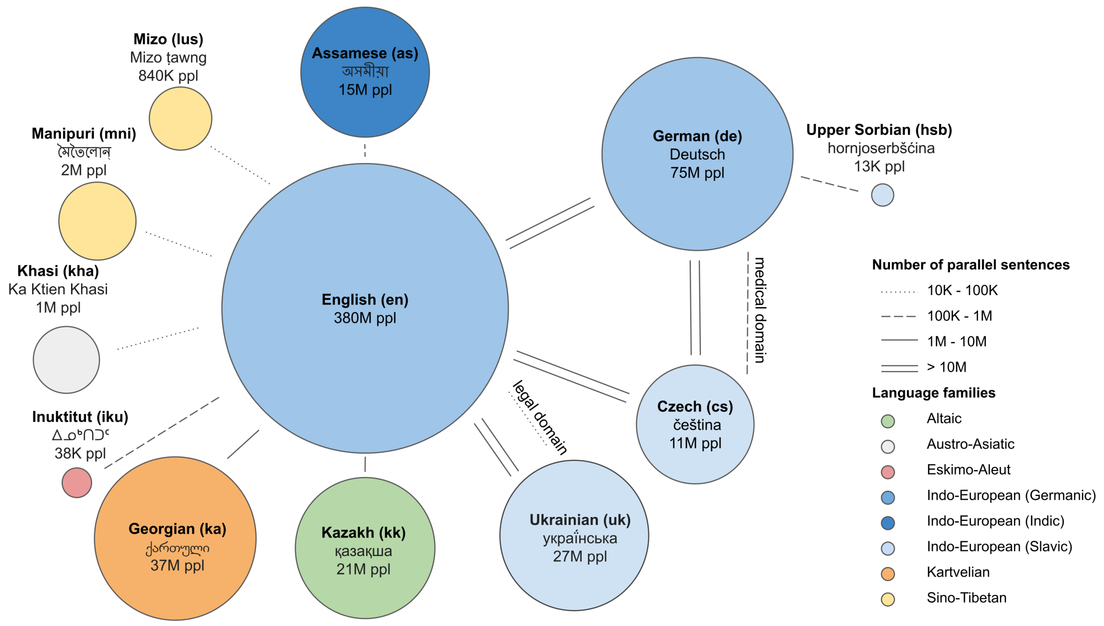

Figure 1.3: Languages used in this thesis in terms of the number of native speakers. Colors reflect language families and the links between languages represent the amount of parallel data available.

called *very low-resource* and they cannot easily benefit from the techniques we propose due to their limited amounts of monolingual data. Many other languages are not even plotted in the chart as they do not have any data available in the OSCAR corpus.

Figure 1.2 illustrates the language pairs relevant for this thesis, their corpus sizes and their linguistic similarity. Figure 1.3 shows the languages in terms of their speaker base rather than their text data amounts. Comparing the two figures allows us to judge how big a language really is (as represented by the number of native speakers) in contrast with how strong its online presence is. The dominance of English or German is less pronounced when measured by the size of their speaker base. On the other hand, Czech is an example of a language that possesses a comparatively abundant volume of data in relation to its number of speakers which reflects a strong NLP community supporting it. Similarly, Inuktitut has only 38k speakers but a relatively big parallel corpus of 1M languages due to the support of the National Research Council of Canada which published the proceedings of the Legislative Assembly of Nunavut in the Hansard corpus.

2. Methodology

2.1 Parallel Corpus Mining

Real data collection from human translators leads to the creation of data sets of the highest quality, but it is also the slowest and the most expensive option. Arguably, if we want to improve the translation quality of a particular low-resource language or domain, collecting new data from native speakers or domain experts is the best thing that we can do. However, when collecting new natural pieces of text is not an option, we can resort to finding parallel sentences in existing comparable corpora. In our work, we explore the possibilities of parallel sentence search and we present a strategy to mine parallel sentences from monolingual corpora. We consider the mined sentence pairs to be *pseudo-parallel* as they should ideally be identical in meaning but in practice only share a certain degree of similarity.

Our approach to parallel corpus mining is the following:

1. embed sentences in a multilingual space using pre-trained Transformer encoders;
2. score all possible candidate sentence pairs;
3. set a threshold score for two sentences to be considered parallel;
4. select sentence pairs that score above the threshold.

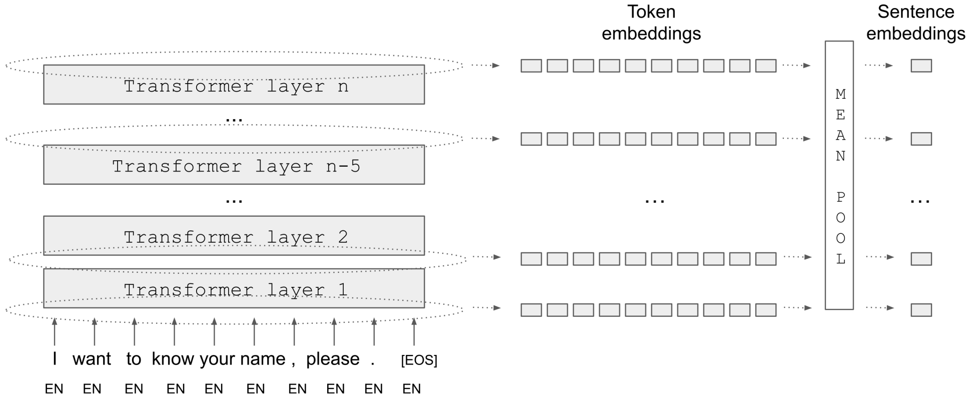

Figure 2.1: Encoding a masked sentence by a Transformer model. Contextualized word embeddings are aggregated by mean-pooling.

2.1.1 Sentence Encoders

Our sentence encoders are Transformer language models trained using the masked language modelling (MLM) training objective in multiple languages [Conneau and Lample, 2019]. To further align their internal representations and make them more language-neutral, we fine-tune the models using a translation language modelling (TLM) objective on synthetic parallel data in one arbitrary language pair.

During TLM fine-tuning, pairs of synthetic parallel sentences are concatenated, random tokens are masked from both sentences 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 thus have the capacity to enrich word representations with information about their monolingual context as well as their translation counterparts. This explicit cross-lingual training objective further enhances the alignment of the embeddings in the cross-lingual space.

We use this objective to fine-tune the pre-trained model on a small synthetic parallel data set obtained via unsupervised MT for one language pair, aiming to improve the overall cross-lingual alignment of the internal representations of the model. In our experiments, we also compare the performance to fine-tuning on a small authentic parallel corpus.

To derive sentence embeddings from the pre-trained encoder, we compute the per-sentence average of token representation vectors generated by the model at each layer as illustrated in Figure 2.1. We experiment with representations at different layers to determine which carry the most cross-lingual information.

2.2 Unsupervised Neural Machine Translation

We describe the methodology of unsupervised neural MT (UNMT) adopted in our experiments. In the full thesis, we also explore unsupervised phrase-based models (UPBMT) which are excluded from this thesis abstract for the sake of conciseness.

2.2.1 Architecture

The design of a neural MT system needs to meet several requirements to be functional for unsupervised translation. Firstly, a significant number of parameters needs to be shared among the languages in order to allow the model to generate a shared latent space where meaning is represented regardless of the language it is expressed in [Lample et al., 2018b]. Secondly, the initialization of the model weights is vital to produce an initial solution and kick-start the training process [Conneau and Lample, 2019].

Our UNMT systems consist of a Transformer encoder and decoder, both of which are shared between the two languages. The shared encoder is essential for creating the shared space of cross-lingual latent representations, the shared decoder serves for regularization. The encoder and the decoder have the same 6-layer Transformer architecture with 8 attention heads and the hidden size of 1024, language embeddings, GELU [Hendrycks and Gimpel, 2017] activations, and a dropout rate of 0.1.

2.2.2 Pre-Training

We experiment with different methods to initialize the UNMT model:

- The encoder-decoder model is initialized randomly, only the token embedding weights are copied from a pre-trained word embedding model.
- The encoder is pre-trained on monolingual corpora with an MLM objective [Conneau and Lample, 2019]. The weights are also copied into the decoder.
- The encoder-decoder model is initialized with weights of a bilingual or multilingual denoising autoencoder [Liu et al., 2020] pre-trained on the monolingual data in source and target languages, possibly in additional auxiliary languages.

2.2.3 Fine-Tuning for Translation

Our UNMT systems are trained on synthetic data generated by a phrase-based model using standard back-translation [Sennrich et al., 2016] or synthetic data generated on-the-fly by the UNMT model itself using online back-translation [Artetxe et al., 2018]. Training on pseudo-parallel data is performed with the supervised translation objective. The details on the fine-tuning and the objective functions are given in the full thesis.

3. Experiments & Results

We carried out several sets of experiments with different unsupervised MT approaches and different language pairs. In each section, we focus on a specific unsupervised technique: parallel corpus mining (Section 3.1), combining UPBMT and UNMT (Section 3.2), and training on pseudo-parallel data (Section 3.3). Finally, we point out the limitations of unsupervised techniques (Section 3.4). Our experiments with phrase-based MT models, pre-training strategies for neural models, and semi-supervised training are omitted here and can be found in Chapter 7 of the full thesis.

We have the following hypotheses regarding the outcomes of our experiments:

- We hypothesize that UNMT can benefit from different cross-lingual information brought into the training by synthetic corpora produced by phrase-based models (Section 3.2).
- We hypothesize that multilingual sentence encoders trained without any parallel data can be used for parallel corpus mining (Section 3.1).
- We hypothesize that existing UNMT models are not able to fully leverage the cross-lingual signal present in monolingual data and we propose a method to explicitly match similar sentences beforehand to present the model with the matched pseudo-parallel sentence pairs in addition to the unaligned monolingual texts (Section 3.3).

3.1 Parallel Corpus Mining

In this section, we explore unsupervised strategies to mine parallel sentences from monolingual data when there are no parallel data to start with. We experiment with pre-trained multilingual Transformer encoders and propose a method to align their internal representations and use them to derive sentence embeddings. Our approach is completely unsupervised and is applicable also for distant language pairs. The methodology was described in Section 2.1 and more details can be found in Chapter 4 of the full thesis as well as in our published research paper [Kvapilíková et al., 2020].

In our experiments, we empirically evaluate the quality of our cross-lingual sentence embeddings and compare it with state-of-the-art supervised methods and unsupervised baselines. We evaluate the proposed method on the task of parallel corpus mining and parallel sentence matching. We fine-tune the pre-trained *XLM-100*[1]

[1] https://huggingface.co/xlm-mlm-100-1280

	EN-DE	EN-FR	EN-RU	EN-ZH	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	multi (2M sent.)
Azpeitia et al. [2018]	85.52	81.47	81.30	77.45	bitext (2-9M sent.)
Artetxe&Schwenk [2019]	**96.19**	**93.91**	**93.30**	**92.27**	multi (223M sent.)
Word Mapping	32.04	32.94	17.68	20.65	none
Vanilla XLM	62.10	64.77	61.65	44.79	none
Our method (TLM EN↔DE)	**80.06**	**78.77**	**77.16**	**67.04**	none (20k synth sent.)

	EN-DE	EN-FR	EN-RU	EN-ZH	EN-KK	CS-ZH	DE-RU
Artetxe&Schwenk [2019]	**90.30**	**87.38**	**94.34**	**83.92**	12.07	**73.41**	**88.39**
Word Mapping	28.45	30.79	17.81	16.04	2.28	10.86	19.55
Vanilla XLM	72.58	71.92	72.90	59.26	24.00	43.00	58.29
Our method (TLM EN↔DE)	**79.32**	**77.05**	**80.98**	**65.49**	**35.41**	**48.79**	**65.91**

Table 3.1: F1 score on the parallel sentence mining task measured on the BUCC test set (top) and News test set (bottom). The supervised and unsupervised winners are highlighted in bold. Artetxe and Schwenk [2019] values were obtained using the public implementation of the LASER toolkit.

Source: Kvapilíková et al. [2020]

model using either English-German (EN-DE) or Czech-German (CS-DE) synthetic parallel data. For comparison, we fine-tune two alternative models using authentic parallel data in the following two low-resource language pairs: English-Nepali (EN-NE) and English-Kazakh (EN-KK).

3.1.1 Results & Discussion

Evaluation I: Parallel Corpus Mining

We measure the performance of our method on the BUCC shared task of parallel corpus mining where candidate systems are expected to search two comparable non-aligned corpora and identify pairs of parallel sentences. We evaluate on two data sets – the original BUCC 2018 corpus created by inserting parallel sentences into monolingual texts extracted from Wikipedia [Zweigenbaum et al., 2017] and a new BUCC-like data set (News train and test) which we created by shuffling 10k parallel sentence from News Commentary into 400k monolingual sentences from News Crawl. The BUCC and News data sets are comparable in size and contain parallel sentences from the same source, but differ in overall domain.

Table 3.1 shows the results of our proposed model on the BUCC and News test sets. When comparing our method to related work, it must be noted that the underlying *XLM* model was pre-trained on Wikipedia and therefore has seen the monolingual BUCC sentences during training. This could result in an advantage over other systems, as the

	DE-EN	CS-EN	CS-DE	CS-FR	CS-RU	FR-ES	FR-RU
Artetxe&Schwenk [2019]	98.78	99.08	99.23	99.37	98.77	99.42	98.60
Word Mapping	60.60	55.03	75.35	43.33	79.87	71.07	41.25
Vanilla XLM	87.15	79.83	82.87	80.55	85.15	91.07	85.28
Our method (TLM EN↔DE)	93.97	**90.47**	90.48	**90.07**	92.23	**94.68**	**91.80**
Our method (TLM CS↔DE)	**94.43**	90.15	**90.50**	89.48	**92.33**	94.65	91.72

Table 3.2: Accuracy on the deshuffling task (*newstest2012*) averaged over both matching directions. Artetxe and Schwenk [2019] values were obtained using the public implementation of the LASER toolkit.

Source: Kvapilíková et al. [2020]

model could exploit the fact that it has seen the non-parallel part of the comparable corpus during training. However, since both the proposed method and the *vanilla XLM* baseline suffer from this, their results remain comparable. We also report results on the News test set which is free from such potential bias (Table 3.1).

The results reveal that TLM fine-tuning on the synthetic parallel sentences which is the key element of our method brings a substantial improvement over the initial pre-trained model trained only using the MLM objective (*vanilla XLM*). In terms of the F1 score, the gain across four BUCC language pairs major and ranges between 14.0-22.3 points (see Table 3.1). 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-Chinese 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 partially attributed to the superiority of contextualized embeddings for representation of sentences over static ones. Furthermore, word mapping relies on the questionable assumption of isomorphic embedding spaces which weakens its performance, especially for distant languages. In our proposed model, it is possible that joint training of contextualized representations induces an embedding space with more convenient geometric properties which makes it more robust to language diversity.

Although the performance of our model generally 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 or English-Kazakh, see Table 3.1.

Evaluation II: Corpus Deshuffling

To assess the effect of the proposed fine-tuning on other language pairs not covered by BUCC, we evaluate our embeddings on the task of corpus deshuffling. The task entails searching a pool of shuffled parallel sentences to recover correct translation pairs. Cosine similarity is used for the nearest neighbor search.

	AF	AR	AZ	BE	BG	CA	CS	DE	EL	EO
Sup. baseline	89.5	92.0	66.0	66.2	95.0	95.9	96.5	99.0	95.0	97.2
Vanilla XLM	38.1	19.9	25.1	33.7	36.2	51.0	31.5	65.0	27.0	45.8
EN↔DE (synth)	57.3	41.1	46.3	58.4	56.0	66.9	53.5	83.1	51.3	68.0
CS↔DE (synth)	54.2	41.2	44.2	61.8	60.7	68.9	59.9	87.3	53.1	67.4
EN↔KK (auth)	58.4	45.6	51.4	60.2	59.2	72.6	53.9	87.0	54.6	72.1
EN↔NE (auth)	59.9	46.6	54.2	63.1	62.9	71.0	57.6	85.0	51.0	71.2
	ET	**FI**	**FY**	**HI**	**HR**	**IA**	**IS**	**ID**	**JA**	**KA**
Sup. baseline	96.7	96.3	51.7	94.7	97.2	95.2	95.6	94.5	91.8	35.9
Vanilla XLM	19.8	31.4	37.0	26.2	47.2	57.3	25.0	46.4	29.5	22.1
EN↔DE (synth)	39.0	47.5	48.6	53.4	68.2	71.4	43.1	64.9	54.4	41.4
CS↔DE (synth)	41.4	49.5	44.8	51.7	71.8	70.5	43.7	64.1	53.3	39.8
EN↔KK (auth)	43.4	51.3	51.7	60.3	71.3	79.5	45.0	66.4	59.6	44.0
EN↔NE (auth)	44.6	52.7	48.6	59.3	72.1	75.7	47.1	67.8	59.6	47.8
	KK	**KU**	**LT**	**MK**	**ML**	**MN**	**MR**	**MS**	**NE**	**NN**
Sup. baseline	18.6	17.2	96.2	94.7	96.9	8.2	91.5	96.4	20.6	88.3
Vanilla XLM	17.4	10.6	22.0	25.8	17.4	12.6	15.3	52.0	21.3	49.9
EN↔DE (synth)	33.6	16.8	43.9	48.8	51.6	29.0	37.3	67.0	32.8	66.8
CS↔DE (synth)	34.7	16.2	46.2	51.1	44.3	24.5	34.2	65.4	31.4	67.5
EN↔KK (auth)	46.1	20.0	46.2	54.7	54.0	32.7	41.9	69.8	37.3	69.2
EN↔NE (auth)	38.4	20.9	47.7	53.8	56.0	34.9	43.5	72.1	42.8	69.2
	OC	**SL**	**SR**	**SV**	**TA**	**TE**	**TL**	**UK**	**UR**	**YI**
Sup. baseline	61.2	95.9	95.3	96.6	69.4	79.7	50.5	94.5	81.9	5.7
Vanilla XLM	20.0	34.7	35.9	47.2	11.9	14.1	14.6	38.0	19.3	9.9
EN↔DE (synth)	34.3	54.9	58.6	69.7	40.9	44.7	24.0	66.1	43.7	22.1
CS↔DE (synth)	35.9	59.2	64.8	71.8	31.9	37.8	20.4	70.4	43.8	22.8
EN↔KK (auth)	40.3	58.0	64.3	73.3	42.8	44.0	24.4	71.6	48.2	25.8
EN↔NE (auth)	36.9	58.8	65.0	72.0	41.7	53.2	26.8	71.0	49.9	26.7

Table 3.3: Accuracy on the deshuffling task (*Tatoeba*) averaged over both matching directions (to and from English). The supervised baseline was obtained using the public implementation of the *LASER* model [Artetxe and Schwenk, 2019]. Our proposed models were fine-tuned on synthetic parallel data (EN↔DE, CS↔DE) and authentic parallel data (EN↔KK, EN↔NE).

Source: Kvapilíková et al. [2020]

We first evaluate the pairwise matching accuracy on the *newstest* multi-way parallel data set of 3k sentences in 6 languages.[2] We use *newstest2012* for development and *newstest2013* for testing. The results in Table 3.2 show that the fine-tuned model is able to match correct translations in 90-95% of cases, depending on the language pair, which is ∼7 p.p. more than *vanilla XLM*. It is notable that the model which was only fine-tuned on English-German synthetic parallel data has a positive effect on completely unrelated language pairs as well (e.g. Russian-Spanish, Czech-French).

Since the greatest appeal of parallel corpus mining is to enhance the resources for low-resource languages, we also measure the deshuffling accuracy on the Tatoeba [Artetxe and Schwenk, 2019] data set of 0.5–1k sentences in over 100 languages

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

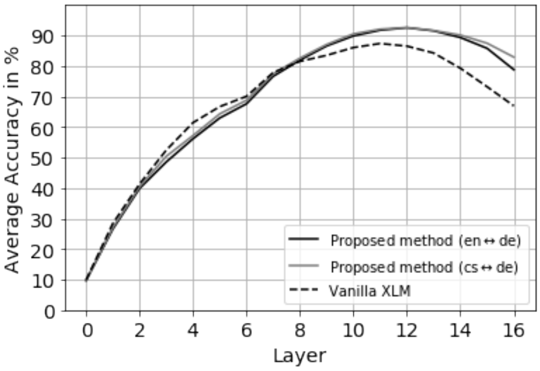

Figure 3.1: Average deshuffling accuracy on *newstest2012* before and after fine-tuning from the input embedding layer (0th) to the deepest layer (16th).

Source: Kvapilíková et al. [2020]

aligned with English. Aside from the two completely unsupervised models, we fine-tune two more models on small authentic parallel data in English-Nepali (5k sentence pairs from the Flores development sets) and English-Kazakh (10k sentence pairs from News Commentary). Table 3.3 confirms that the improvement over *vanilla XLM* is present for every language we evaluated, regardless of the language pair used for fine-tuning. We initially hypothesized that the performance of the English-German model on English-aligned language pairs would exceed the German-Czech model, but their results are equal on average. Fine-tuning on small authentic corpora in low-resource languages exceeds both by a slight margin.

The results are clearly sensitive to the amount of monolingual sentences in the Wikipedia corpus used for XLM pre-training and the matching accuracy of very low-resource languages is significantly lower than we observed for high-resource languages. However, the benefits of fine-tuning are substantial (around 20 percentage points) and for some languages, the results even reach the supervised baseline (e.g. Kazakh, Georgian, Nepali).

It seems that explicitly aligning one language pair during fine-tuning propagates through the shared parameters and improves the overall representation alignment, making the contextualized embeddings more language-agnostic. The propagation effect could also positively influence the ability of cross-lingual transfer within the model in downstream tasks. A verification of this is left to future work.

Analysis: Representations Across Layers

We derive sentence embeddings from each of the layers of the model and show deshuffling results on the development set averaged over all language pairs in Figure 3.1, both before and after fine-tuning. The accuracy differs substantially across the model depth, the best cross-lingual performance is consistently achieved around the 12th (5th-to-

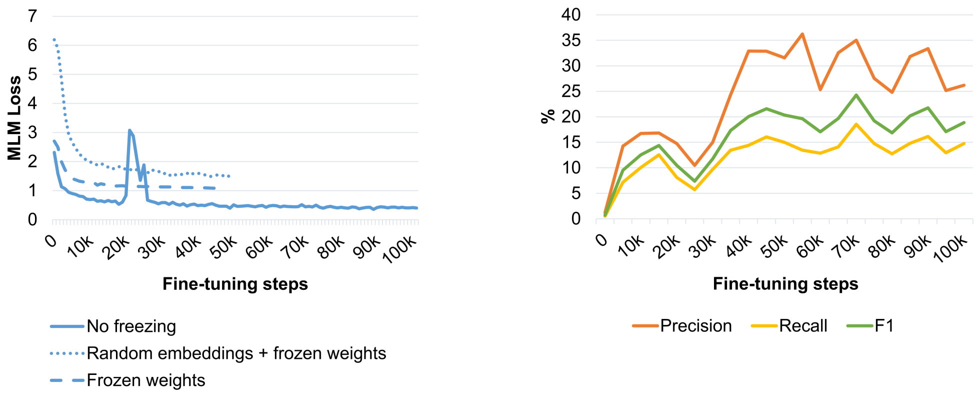

Figure 3.2: Training curves from fine-tuning the proposed model (en↔de) with the MLM objective on English and Inuktitut texts with and without parameter freezing *(left)*. Precision, recall and F1 scores of the model fine-tuned without weight freezing on the task of parallel corpus mining for English and Inuktitut *(right)*.

last) layer of the model. The TLM fine-tuning affects especially the deepest layers.

Parallel Corpus Mining for Unsupported Languages

The *XLM* model only supports the 100 languages covered during pre-training. To use its representations for other languages, the model first has to be fine-tuned.

In the following experiments, we create sentence representations for text in Inuktitut, a language that was not included in the pre-training of the *XLM*, and use them for English-Inuktitut parallel corpus mining.

We create an English-Inuktitut (EN-IKU) encoder by fine-tuning our proposed model (EN↔DE) with the MLM objective on 1M monolingual sentences from the Hansard corpus (IKU) and NewsCrawl (EN). Since the two languages are linguistically distant and Inuktitut has a non-Latin script, this is a particularly difficult scenario.

We experiment with fine-tuning the entire model versus weight-freezing and fine-tuning only the lexical embeddings. Furthermore, we experiment with random initialization of lexical embeddings before the fine-tuning. The training curves are shown in Figure 3.2. Although updating the entire model experiences a sudden drop in performance at the beginning of the training, it recovers and eventually converges to the highest MLM accuracy out of the three approaches. Therefore, in our future experiments, we do not freeze weights and always update the entire model during fine-tuning.

Decreasing MLM loss does not yet guarantee that the model is creating bilingual representations usable for parallel sentence search. We measure the mining performance of the model by trying to recover 5k parallel sentences[3] mixed into 100k monolingual sentences. The precision, recall, and F1 scores are evaluated as the fine-tuning

[3] Parallel sentences are taken from the Hansard dev set.

progresses and plotted in Figure 3.2. We observe an initial performance boost as the model adapts to the new language, followed by fluctuating outcomes, with precision ranging from 25% to 35%. The fact that the model was able to correctly recover up to 18% of the hidden sentences means that it was able to at least partially align its representations of Inuktitut to English.

3.1.2 Takeaways

We proposed a completely unsupervised method for training of multilingual sentence embeddings which can be used for building a parallel corpus with no previous translation knowledge.

We showed that by fine-tuning a pre-trained multilingual encoder with the TLM objective of gap-filling in bilingual sentence pairs, we can significantly enhance the cross-lingual alignment of its representations using as little as 20k synthetic translation pairs. Since the synthetic translations were obtained from an unsupervised MT system, the entire procedure requires no authentic parallel sentences for training.

Our sentence embeddings yield significantly better results on the tasks of parallel corpus mining and parallel sentence matching than our unsupervised baselines. Interestingly, targeting only one language pair during the fine-tuning phase suffices to propagate the alignment improvement to unrelated languages. It is therefore not necessary to build a working MT system for every language pair we wish to mine.

It is possible to adapt the proposed approach to new languages outside of the original model coverage by MLM fine-tuning. The performance can be further improved by light fine-tuning of the adapted model using synthetic parallel sentences. The source of this improvement deserves further investigation.

3.2 Hybrid Unsupervised MT

In our initial experiments, we worked with UPBMT systems and used them to generate training data for unsupervised neural models. The systems covered in this section are termed “hybrid” due to their neural model architecture which incorporates PBMT-generated synthetic data. We experiment with Czech↔German translation and compare the results to a supervised benchmark to evaluate the gap between unsupervised and supervised models. Furthermore, we compare to a pivoting benchmark where we translate from German to Czech via English.

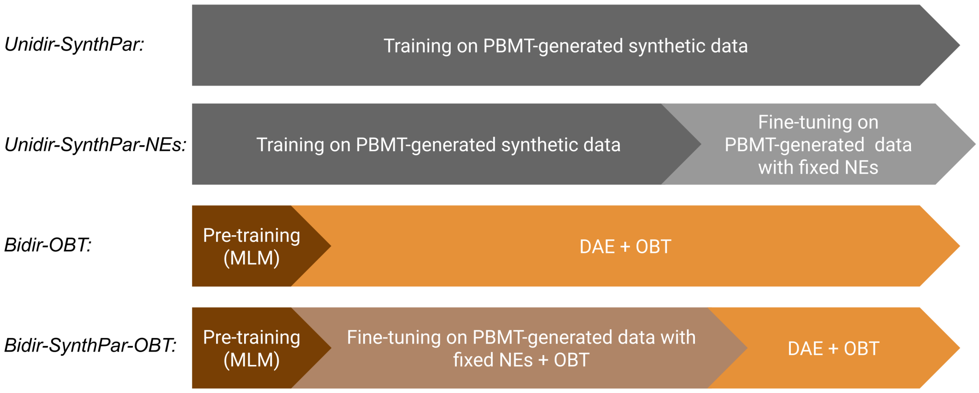
Unidir-SynthPar:
Training on PBMT-generated synthetic data
Unidir-SynthPar-NEs:
Training on PBMT-generated synthetic data
Fine-tuning on PBMT-generated data with fixed NEs
Bidir-OBT:
Pre-training (MLM)
DAE + OBT
Bidir-SynthPar-OBT:
Pre-training (MLM)
Fine-tuning on PBMT-generated data with fixed NEs + OBT
DAE + OBT

Figure 3.3: Schematic illustration of the training pipeline of our models. The size of the blocks is not proportional to training time or raining data.

3.2.1 Model & Training

We use the Transformer architecture described in Chapter 2 to train the DE→CS hybrid models. The training pipelines are illustrated in Figure 3.3. Some models are trained only on UPBMT-generated synthetic data (*SynthPar*), while others employ online back-translation (OBT) to generate synthetic training samples by the UNMT model itself. The systems trained exclusively on the *SynthPar* corpus are unidirectional (DE→CS) whereas systems trained with OBT must be bidirectional (DE↔CS). While the unidirectional models (*Unidir-SynthPar*) are trained from scratch, the bidirectional models *Bidir-OBT* (baseline without synthetic data) and *Bidir-SynthPar-OBT* are pre-trained on the MLM task as described in Section 2.2.

3.2.2 Results & Discussion

The scores of the systems on our test set are reported in Table 3.4. They demonstrate that we can significantly elevate translation quality by training a UNMT system on the UPBMT-generated synthetic data. COMET and chrF++ results are in line with the BLEU score. Our results reveal a narrow difference between the hybrid system and the supervised benchmark, which was trained on 9 million parallel sentences, amounting to only 2.1 BLEU. The pivoting approach yields inferior results compared to the majority of our unsupervised systems. Our findings validate the effectiveness of unsupervised methods in scenarios characterized by large amounts of monolingual data.

Synthetic Training Data Quality

It must be noted that while the UPBMT-generated translations were produced by a finished model, the UNMT-generated synthetic sentence pairs are produced on-the-fly by OBT and are of progressively increasing quality, starting at translations full of

	DE→CS		
	BLEU	**chrF++**	**COMET**
UPBMT	11.6	38.0	0.59
UNMT (Bidir-OBT)	14.6	39.2	0.72
Unidir-SynthPar	15.0	40.8	0.74
*Unidir-SynthPar-NEs**	14.3	40.5	0.74
Bidir-SynthPar-OBT	**16.7**	**42.6**	**0.79**
Benchmark-Supervised	*18.8*	*44.7*	*0.83*
Benchmark-Pivot	*15.1*	*40.1*	*0.75*

Table 3.4: Our unsupervised hybrid systems and their performance on newstest2019. For details on the UPBMT models please refer to Section 7.1 of the full thesis. The *Benchmark-Supervised* model was trained on 9M parallel sentences from OPUS. The *Benchmark-Pivot* model was trained on 26M parallel sentences in DE-EN and EN-CS.

repeating punctuation marks and copied (non-translated) words. We had a closer look at the quality of the back-translated sentences and made the following observations.

- Already after 1k training steps the structure of OBT translations starts corresponding to the source sentence.

- It lasts several more iterations to get rid of most mistranslations and copied German source words. For example, at 1k training steps, the German sentence *"Krähen stehen unter Naturschutz."* ("*Crows are protected by nature conservation laws.*") is translated as *"Krämerovy houby stojí mimo Naturschutz"*, where *"Naturschutz"* is copied and *"Krämerovy houby"* (*"Krämer's mushrooms"*) is a complete mistranslation motivated by a subword overlap of the first word.

- Although the translation is subword-based, it happens only rarely that a part of a word would remain non-translated, e.g. *"Erfolgversprechende"* (*"promising"* translated as a non-existent word *"Erfolgtivní"*. Even long German compound words mostly get copied as a whole (e.g. *"Witterungsbedingungen"*). This is likely the result of MLM pre-training and possibly also the fairly big BPE vocabulary of 60k units.

Named Entity Translation

In our experiments, we tried to mitigate the problem of mistranslated named entities by post-processing the *SynthPar* corpus. We show that it partially alleviates the problem as models trained on *SynthPar* have better results (Table 3.6). However, incorrectly translated names continue to be one of the most serious errors generated by unsupervised translation systems. See Table 3.5 for a sample translation.

Source	**Phrase**
Original	Der Lyriker **Werner Söllner** ist IM **Walter**.
Reference	Básník **Werner Söllner** je tajný agent **Walter**.
PBMT	Český prozaik **~~Miroslav Mišák~~** je agentem StB **~~Josef~~** .
Unidir-SynthPar	Prozaik **~~Filip Bubeníček~~** je agentem StB **~~Josefem~~**.
Unidir-SynthPar-NEs	Prozaik **~~Filip~~ Söllner** je agentem StB **~~Ladislavem Bártou~~**.
Bidir-OBT	Lyrik **~~Jiří~~ Söllner** je IM **Walter~~man~~**.
Bidir-SynthPar-OBT	Prozaik **Werner Söllner** je IM **Walter~~man~~**.

Table 3.5: Sample translations showing that fine-tuning on synthetic corpus with cleaned NEs (*Unidir-SynthPar-NEs* and *Bidir-SynthPar-OBT*) alleviates a part of the NE problem. However, note the imperfect translation of *Lyriker* as *novelist* rather than *poet*. The bidirectional systems seem to be more prone to copying which can help for some NEs but also hurt, e.g. copying the word *IM* rather than recognizing it as a shortcut for *"inoffizieller Mitarbeiter"* and translating it as *secret agent*.

	Sentences with NEs	**Sentences with no NEs**
Unidir-SynthPar	28%	26%
Unidir-SynthPar-NEs	52%	28%
No winner	20%	46%

	Sentences with NEs	**Sentences with no NEs**
Bidir-OBT	22%	18%
Bidir-SynthPar-OBT	38%	40%
No winner	40%	42%

Table 3.6: Results of manual evaluation of three systems on a stratified subset of the validation data set created by randomly selecting 100 sentences with NEs and 100 sentences without NEs.

3.2.3 Takeaways

The UPBMT-generated synthetic corpus serves as a valuable source of cross-lingual signal for UNMT models. Such hybrid models consistently achieve higher quality compared to pure neural models. The synthetic corpus brings the most value at the beginning of the training when the UNMT model is not yet able to generate meaningful translations on its own. Once the UNMT model attains a satisfactory level of quality, it is advisable to phase out the initial synthetic corpus, as it can potentially impede further training. If the UNMT system is initialized well, the training starts successfully, and at 1k training steps we observe that the UNMT starts generating meaningful translations.

In our view, one of the most significant types of translation errors in unsupervised systems involves a high frequency of randomly mistranslated named entities. This problem is not adequately addressed by the BLEU score but it has a considerable impact on the perceived translation quality. We have invested our efforts in mitigating this issue during the fine-tuning of the UNMT system by rectifying NEs in the synthetic training corpus (for details see Section 7.1 of the full thesis). Some names where

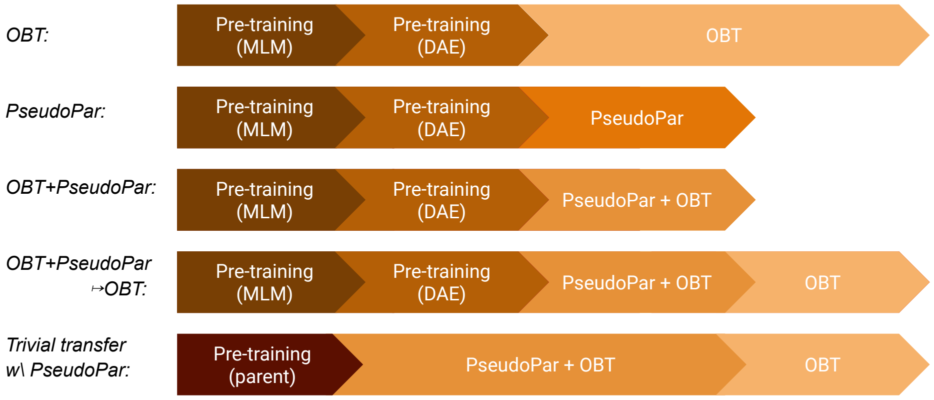

Figure 3.4: Schematic illustration of the training pipeline of our models. The size of the blocks is not proportional to training time or training data.

deleted, others were replaced by their copy. While our approach may not be flawless, we believe that an omitted named entity or a non-translated named entity is less detrimental than a randomly substituted one. Unfortunately, this approach to amending NEs can only be applied to languages with a name tagger available, which is not the case for many truly low-resource languages.

3.3 Boosting Unsupervised MT with Pseudo-Parallel Data

We measure the effect of incorporating pseudo-parallel sentences obtained by parallel corpus mining into unsupervised MT. We hypothesize that they can serve as a new source of cross-lingual information that the model can benefit from. Although pseudo-parallel sentences are not perfect translation equivalents, we believe that they can improve the translation quality nonetheless, especially when used at the beginning of the training. We evaluate on four language pairs: DE-HSB, EN-KA, EN-KK, EN-UK.

3.3.1 Model & Training

We first create a pseudo-parallel corpus (*PseudoPar*) as described in Section 2.1. We experiment with different fine-tuning strategies for unsupervised machine translation as illustrated in Figure 3.4. For each language pair, all translation models are initialized with the same weights obtained in the pre-training stage during MLM and DAE training.

OBT (baseline) models are fine-tuned solely with the online back-translation loss. *PseudoPar* models are fine-tuned with the standard supervised MT loss on our pseudo-parallel corpora. *OBT+PseudoPar* models are fine-tuned simultaneously with the

	DE-HSB	**HSB-DE**	**EN-KA**	**KA-EN**	**EN-KK**	**KK-EN**	**EN-UK**	**UK-EN**
WMT22 best	17.9	18.0	-	-	-	-	-	-
ChatGPT	6.6	-	3.9	-	5.2	-	**25.8**	-
OBT (baseline)	29.6	36.3	3.6	5.2	0.8	1.0	8.4	12.9
PseudoPar	11.3	12.0	1.9	4.8	1.0	3.1	4.6	8.6
OBT+PseudoPar	32.9	36.3	6.8	12.7	5.9	11.3	12.2	20.8
↦OBT	**35.0**	**39.6**	**7.7**	**14.0**	**7.2**	**12.1**	**15.7**	**23.7**

	DE-HSB	**HSB-DE**	**EN-KA**	**KA-EN**	**EN-KK**	**KK-EN**	**EN-UK**	**UK-EN**
de Gibert Bonet (2022)	-	-	12.0	-	6.4	-	20.8	-
OBT (baseline)	-	-	9.0	12.7	0.3	0.3	14.9	12.6
PseudoPar	-	-	2.1	6.8	8.0	11.6	14.6	13.1
OBT+PseudoPar	-	-	11.5	22.0	**16.3**	**18.6**	**29.3**	**21.7**
↦OBT	-	-	**15.0**	**23.5**	9.3	12.7	27.5	**21.8**

Table 3.7: MT performance of our systems measured by BLEU scores on the general test set (top) and the legal test set (bottom). Compared to the WMT22 winner [Shapiro et al., 2022], ChatGPT, and the hybrid system trained by de Gibert Bonet et al. [2022].

back-translation loss on the monolingual sentences and with the standard MT loss on the pseudo-parallel sentence pairs. *OBT+PseudoPar↦OBT* models are a continuation from different checkpoints of the *OBT+PseudoPar* models where the supervised MT objective using pseudo-parallel data is dropped and the training continues with online back-translation only. We experiment with different checkpoints to find the optimal point to switch the training.

3.3.2 Results & Discussion

We observed a significant improvement in translation quality over the baseline for all translation pairs. Table 3.7 shows that the baseline *OBT* system falls short of our proposed method by between 4.7 BLEU points (EN→KK) and 10.7 BLEU points (UK→EN) on the general test set. The differences on the legal test set are even more pronounced: we observe an increase of up to 14.5 BLEU over the baseline (EN→UK). Our DE→HSB system outperforms the WMT22 winner by 17 BLEU points. When translating from English to Kazakh, our approach reaches a BLEU score of 16.3 while the baseline which solely relies on iterative back-translation does not receive enough cross-lingual signal to start learning at all. The hybrid system by de Gibert Bonet et al. [2022] which uses additional translation information from an unsupervised phrase-based system falls behind with a BLEU score of 6.4.

The results of translation by ChatGPT from English or German into truly low-resource languages (HSB, KA, KK) are significantly worse than our results. After manually evaluating several translations with a zero BLEU score, we suspected that the automatic metric puts ChatGPT's fluent but less literal translations at a disadvantage. We calculated the COMET score which is better able to capture the meaning similarity between texts but this hypothesis was not confirmed. The COMET score ranks

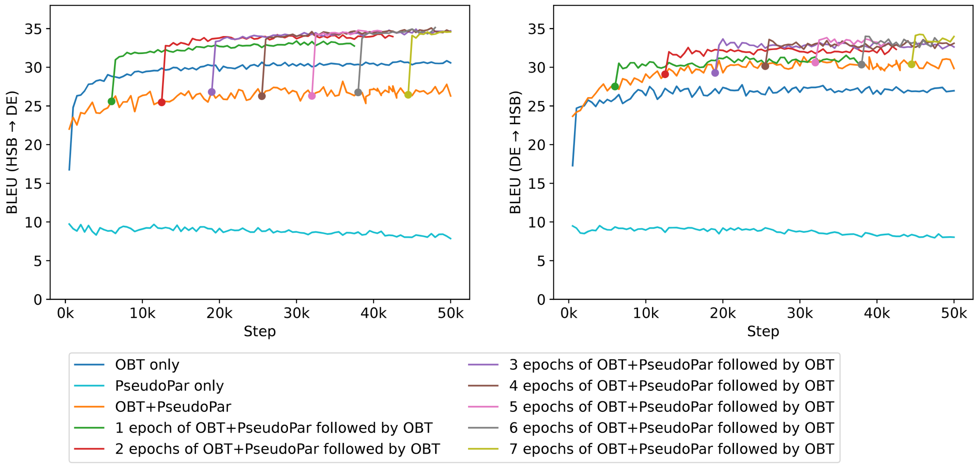

Figure 3.5: The development of validation BLEU scores during the training of HSB→DE (left) and DE→HSB (right) models. Any parallel resources were prohibited.

ChatGPT outputs similarly as the BLEU score.

Nonetheless, the EN→UK translation by ChatGPT is better than all unsupervised MT systems according to all used metrics. It must be noted that the systems cannot be directly compared to ChatGPT since its training corpus is larger and might include parallel texts (possibly even the test set).

Training Schedules

Figure 3.5 shows training curves with validation BLEU scores of all our DE↔HSB systems. We see that the *OBT+PseudoPar* system trained simultaneously on back-translated and pseudo-parallel data without any special schedule outperforms the baseline for DE→HSB but not in the opposite direction. For HSB→DE, the baseline performance is surpassed as soon as we remove the pseudo-parallel corpus from the training.

We trained several DE-HSB models starting from *OBT+PseudoPar* after each completed epoch of 770k pseudo-parallel sentences. Upon examination of the training curves in Figure 3.5, we see an immediate increase in validation BLEU score of ∼0.9–4.9 BLEU points which occurred within the first 500 training steps after removing the pseudo-parallel corpus from the training. This observation confirms our hypothesis that pseudo-parallel sentence pairs aid the training in the beginning but the quality of the corpus itself poses an upper bound on the performance of the system. However, removing the corpus too early (after one or two epochs) leads to a lower final BLEU score. Therefore, we recommend to keep training the *OBT+PseudoPar* model until convergence and only then switch to iterative back-translation alone in our *OBT+PseudoPar*↦*OBT* setup.

The flat *PseudoPar* training curves indicate that the quality of the pseudo-

#	DE	HSB	Score
1	Thomas de Maizière	Thomas de Maizière	1.286
2	*Knut* ist tot.	*Bayer* ist tot.	1.245
3	Es ist ein harter Kampf, die Konkurrenz ist groß.	To bě napjata hra, a konkurenca bě wulka.	1.185
4	Der Roman hat *1200* Seiten.	*Kniha* ma *300* stronow.	1.178
5	Er passt zu diesem Team wie der Deckel auf den Topf.	Wón so k mustwu hodźi kaž wěko na hornc.	1.161
6	Die größte misst über *fünf Meter, die kleinste wenige Milli*meter.	Najkrótša měri *10 cm*, najdlěša *1 meter*.	1.101
7	Wer Wohlstand will, braucht Wissenschaft.	Štóž chce *něšto změnić*, *trjeba sylnu wolu*.	1.063
8	*Morgen ist doch auch noch ein Tag*!	*Ale to njeje hišće wšo*!	1.053
7	*Auch für Apple ist das iPhone wichtig.*	*Tež aleje su jara važne.*	1.037

Table 3.8: A sample from the DE-HSB mined parallel corpus. Non-matching words in italics.

parallel corpus alone is inadequate for training a functional MT system without back-translation.

Domain-Specific MT

Interestingly, removing the pseudo-parallel corpus from the training harms the translation quality measured on the legal test sets. There the best performance for EN→KK, KK→EN and EN→UK is achieved by *OBT+PseudoPar*. We suspect that this is the result of the repeating terminology in the domain-specific test sets which is better handled by the *OBT+PseudoPar* for some language pairs. This is consistent with the fact that the *PseudoPar* system trained exclusively on pseudo-parallel data performs quite well on the EN-KK and EN-UK legal test set (8.0 on EN→KK, 11.6 on KK→EN and 14.6 on EN→UK) while having poor results on the general test set (1.0 on EN→KK, 3.1 on KK→EN and 4.6 on EN→UK) as demonstrated in Table 3.7. Based on our findings, we believe that utilizing pseudo-parallel sentences extracted from domain-specific monolingual corpora has the potential to enhance the training of domain-specific MT in general. However, further experiments are out of the scope of this thesis.

Data Quality

The sentence pairs in the pseudo-parallel corpus are far from equivalent in meaning. As illustrated in Table 3.8, many of the sentences are paired because they share a named entity, a numeral or a unit name (not necessarily identical), a punctuation mark, or one distinctive word. Others have a similar sentence structure, they contain a similar segment or they contain words that are somehow related, e.g. Apple/alleys (*"aleje"*), although the word Apple is not the fruit in this context. On the other hand, synthetic

SRC	Ich musste mich laufend weiterbilden, und so legte ich im April 1952 die erste und ein Jahr darauf die zweite Lehramtsprüfung ab.
REF	Dyrbjach so běžnje dale kwalifikować, a tak złožich w aprylu 1952 prěnje a lěto po tym druhe wučerske pruwowanje.
PseudoPar	*Hańža Winarjec-Orsesowa* wotpołoži prěnje wučerske pruwowanje *w lěće 1949 a druhe w lěće* 1952.
OBT @ 500	Dyrbjach so *laufend* dale *kubłać*, a tak *legte w měrcu* 1952 *prěnje* a lěto na to druhe *Lejnjanske* pruwowanje *ab.*
OBT @ 3000	Dyrbjach so běžnje dale *kubłać*, a tak w *měrcu* 1952 prěnju a lěto na to druhu *lektoratu serbšćiny wotpołožichmy.*
OBT @ 10000	Dyrbjach so běžnje dale *kubłać*, a tak wotpołožich w *měrcu* 1952 prěnju a lěto na to druhu *lektoratu.*

Table 3.9: A sample sentence translated by the *OBT* model after 500, 3,000 and 10,000 training steps compared to the closest neighbor of such sentence from the bilingual sentence space (*PseudoPar*). The mistranslated words are indicated in italics.

sentences in the first training iterations are also extremely noisy, and even later they contain artifacts such as non-translated words or mistranslated named entities.

Table 3.9 shows what the back-translated and pseudo-parallel data can look like. We observed how the back-translated (*OBT*) version of one sentence changes as the training progresses and witnessed several types of error, e.g. the German word *"laufend"* is not translated at all in the initial iterations; the word "April" remains mistranslated as "March" ("*měrc*") throughout the entire training. On the other hand, the pseudo-parallel sentence (*PseudoPar*) matched based on its distance from the source sentence has a similar meaning but is factually inaccurate.

It is not clear what are the attributes of the pseudo-parallel corpus that the UNMT training benefits from the most. We believe that the benefits of training on such noisy data are twofold: 1) the perfect matches are a valuable source of correct supervision, and 2) the abundant less-than-perfect matches still introduce a new translation signal which can help the model leave a suboptimal situation which we often observe during back-translation when the model learns to mistranslate a word and never forgets it. An example of error pattern induced by back-translation can be seen in Table 3.9 where the model in different stages of the training consistently mistranslates the word *"weiterbilden"* as *"kubłać"* ("to pour") when the meaning is "to further educate oneself". On the other hand, the word *"laufend"* was first mistranslated but later fixed and at 3k training steps it was correctly translated as *"běžnje"*.

3.3.3 Takeaways

We have demonstrated the benefits of MT training on pseudo-parallel data in situations when true parallel data is not available. While the pseudo-parallel corpus alone does not reach sufficient quality for standard supervised MT training, it works well in combination with online back-translation. We found it optimal to train the model until convergence on both pseudo-parallel and synthetic sentence pairs, then remove the pseudo-parallel corpus and continue training with iterative back-translation only.

We confirm our hypothesis that UNMT models are not able to fully exploit the cross-lingual knowledge present in monolingual data. If we match similar sentences prior to the training using an external tool and present the model with the matched pairs, translation quality significantly improves.

3.4 Limitations of Unsupervised MT

In the previous sections, we established that if parallel texts are not available, MT models can learn using unsupervised techniques from monolingual data only. We tested on four language pairs exhibiting rich linguistic variety, out of which DE-HSB, EN-KA and EN-KK can be considered low-resource.

While the results are promising, the absolute BLEU scores for the more remote language pairs are still fairly low. It has been argued [Marchisio et al., 2020], that unsupervised techniques fail when

- languages are linguistically dissimilar;
- or there is a domain mismatch between the training corpora;
- or there is not enough monolingual sentences (less than 1M) for training.

In the previous section, we showed that we can train functional UMT systems even in the scenarios above. In particular, Georgian and Kazakh are linguistically far from English, and the Upper Sorbian training corpus is only 0.9M sentences. Here we perform several experiments in even more adverse conditions and train MT models for translation between English and four low-resource Indic languages: Assamese (AS), Khasi (KHA), Mizo (MZ), and Manipuri (MNI). All of these languages are linguistically dissimilar from English, the amount of monolingual data is limited (only 183k sentences in Khasi), and the corpora exhibit a domain mismatch. We employ our approach of training on pseudo-parallel corpora to determine whether it can help in situations where other unsupervised techniques fail.

3.4.1 Model & Training

We employ the same methodology as in Section 3.3.

System	Sentence Encoder	EN-AS	AS-EN	EN-MNI	MNI-EN
OBT (baseline)	-	0.2	0.3	0.1	0.1
OBT+PseudoPar	XLM-100 (Indic)	1.0	**1.4**	0.2	0.3
OBT+PseudoPar	XLM-100 (Indic+EN-DE synth)	**1.4**	**1.5**	**2.8**	0.7

System	Sentence Encoder	EN-MZ	MZ-EN	EN-KHA	KHA-EN
OBT (baseline)	-	2.0	0.8	7.7	2.3
OBT+PseudoPar	XLM-100 (Indic)	4.1	**2.3**	7.4	2.0
OBT+PseudoPar	XLM-100 (Indic+EN-DE synth)	**4.8**	**2.5**	**12.6**	**4.6**

Table 3.10: BLEU score of Indic unsupervised MT systems on the WMT23 test set. COMET and chrF++ results are reported in the Appendix.

3.4.2 Results & Discussion

The unsupervised results are reported in Table 3.10. We observe that the BLEU scores for EN-AS and EN-MNI are less than 1 BLEU using the baseline unsupervised approach, meaning that the models learn almost zero translation knowledge. The performance can be significantly improved by adding noisy pseudo-parallel sentences, but BLEU still remains below 3 points. Upon closer analysis of the best translation candidates, we see that such low scores correspond to an average of two word matches per reference-candidate sentence pair. We review the translations and observe that the models generate fluent sentences within the same topic as the source sentence but their meaning is completely off. This finding points in the direction that unsupervised techniques could be useful for domain adaptation or style transfer even in high-resource languages.

There are several possible explanations for such subpar results. Both AS and MNI share a non-Latin script. We experienced problems with the Moses tokenization where words containing compound Unicode characters were often incorrectly split or even segmented at the character level. The amount of monolingual data ($\sim$2M) is lower than we had in our previous experiments. Both languages are linguistically distant from English (which, however, also applies to KA and KK where the unsupervised methods work). And finally, Indic texts contain segments from religious texts whereas English training data is from the news domain.

The results for EN-KHA and EN-MZ are slightly more promising. The effect of training on pseudo-parallel sentences is significant for both language pairs and amounts to $\sim$5 BLEU points. However, we see that the models quickly converge to these values, marking a distinct training trajectory compared to what we witnessed in our experiments from Section 3.3. Moreover, we see very low results in the translation direction from the Indic languages into English which contrasts with our prior experiments where translating into English was less problematic than the reverse direction.

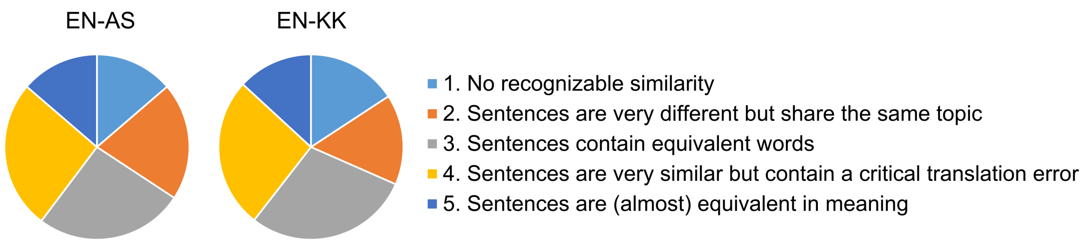

Figure 3.6: Manual evaluation of 100 sentence from English-Kazakh and English-Assamese pseudo-parallel corpora. The evaluation was carried out in English based on the translations from Google Translate.

Data Quality

To have a clearer view of what the pseudo-parallel data look like, we carried out a manual evaluation of EN-KK and EN-KA pseudo-parallel corpora (see Figure 3.6) and found that the structure of the two corpora is relatively similar. However, given the smaller AS monolingual corpus, the AS-EN pseudo-parallel corpus has only 33k sentence pairs. Moreover, the AS-EN data suffers a domain mismatch since the AS corpus contains a significant amount of religious texts. These challenges, together with the linguistic dissimilarity and the problematic Assamese script, might be the reasons why the model fails to start learning.

Surprisingly, despite the low amount of KHA training data (183k sentences), the KHA-EN MT system was able to reach a reasonable level of translation quality without seeing any authentic KHA-EN translations.

3.4.3 Takeaways

We confirm that in the situation of training data domain mismatch, linguistic dissimilarity, different scripts (AS, MNI) and limited amounts of monolingual data, unsupervised MT models struggle. Without *PseudoPar* data in the training mix, the majority of unsupervised models we experimented with did not even start learning. Upon the introduction of *PseudoPar* texts, the BLEU score increases but remains low.

In situations where unsupervised techniques fail, adding a thousand authentic translations into the training can significantly improve the results. With 50k parallel sentences and online back-translation, the models reach a solid translation quality. Data augmentation with noisy pseudo-parallel data is beneficial when we do not have more than 10k authentic sentence pairs.

4. Conclusion

Our research aim was to determine the most effective way of exploiting cross-lingual signal from monolingual data. We conclude that the optimal approach is not to select the single best strategy and use it but rather to rely on a combination of methods. Our contribution lies in a systematic exploration of the approaches and in extending the pool of available methods with a modified pre-training strategy and a novel fully unsupervised way of training data creation.

Unsupervised MT models relying only on model pre-training and back-translation often fail in truly low-resource conditions. We showed that they are not able to fully exploit the translation signal present in monolingual data and they benefit from explicit supervision when extracted from the same data using an external model – pseudo-parallel sentence pairs selected by a multilingual sentence encoder or synthetic sentence pairs generated by a phrase-based MT model.

We proposed a training strategy where we included pseudo-parallel data mined from monolingual corpora in unsupervised MT training and reached a significant improvement across all evaluated language pairs. Although pseudo-parallel texts obtained in a completely unsupervised way are noisy with a majority of sentence pairs being similar rather than equivalent, they offer the model a source of external translation knowledge that can effectively complement the training on synthetic back-translated examples.

For the practical applications of low-resource MT translation, we see the highest potential in large-scale parallel corpus mining and subsequent MT training on the mined corpora. If we relax the strict requirement of no parallel data, it is possible to employ multilingual sentence encoders trained on large parallel corpora in high-resource languages. Using very small amounts of parallel texts coupled with English then suffices for knowledge distillation to new languages.

We see two possible directions of future research in continuation to this work. First of all, exploring the representations hidden in pre-trained multilingual models and improving their cross-lingual alignment is a very relevant topic, especially in the era of large language models. We showed a simple fine-tuning strategy which makes the representations more language-agnostic but the source of that improvement deserves more investigation. Secondly, we believe that the techniques from unsupervised MT are applicable in high-resource scenarios where they can serve for domain adaptation or style transfer. Exploring how to effectively use them for that purpose constitutes an interesting research avenue.

Bibliography

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, Mar 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00288. URL `http://dx.doi.org/10.1162/tacl_a_00288`.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, April 2018. URL `https://arxiv.org/pdf/1710.11041`.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. Extracting parallel sentences from comparable corpora with stacc variants. In Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-07-8. URL `http://lrec-conf.org/workshops/lrec2018/W8/pdf/6_W8.pdf`.

Houda Bouamor and Hassan Sajjad. H2@bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-07-8. URL `http://alt.qcri.org/~hsajjad/publications/papers/2018_BUCC_Bouamor_Comparable_Corpora.pdf`.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf`.

Ona de Gibert Bonet, Iakes Goenaga, Jordi Armengol-Estapé, Olatz Perez-de Viñaspre, Carla Parra Escartín, Marina Sanchez, Mārcis Pinnis, Gorka Labaka, and Maite Melero. Unsupervised machine translation in real-world scenarios. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3038–3047, Marseille, France, June 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.325`.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 26th edition, 2023. URL http://www.ethnologue.com.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2017. URL https://arxiv.org/abs/1606.08415.

Tom Kocmi. Exploring benefits of transfer learning in neural machine translation [dissertation thesis], 2020.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-srw.34. URL https://aclanthology.org/2020.acl-srw.34.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*, 2018a. URL http://arxiv.org/abs/1711.00043.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on EMNLP*, pages 5039–5049, 2018b. URL https://aclanthology.info/papers/D18-1549/d18-1549.

Chongman Leong, Derek F. Wong, and Lidia S. Chao. Um-paligner: Neural network-based parallel sentence identification model. In Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-07-8. URL http://lrec-conf.org/workshops/lrec2018/W8/pdf/7_W8.pdf.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online, November 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.wmt-1.68`.

Sujith Ravi and Kevin Knight. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1002`.

Holger Schwenk. 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, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2037. URL `https://www.aclweb.org/anthology/P18-2037`.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL `https://aclanthology.org/P16-1009`.

Ahmad Shapiro, Mahmoud Salama, Omar Abdelhakim, Mohamed Fayed, Ayman Khalafallah, and Noha Adly. The AIC system for the WMT 2022 unsupervised MT and very low resource supervised MT task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1117–1121, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.wmt-1.110`.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2512. URL `https://www.aclweb.org/anthology/W17-2512`.

List of publications

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, Ondřej Bojar (2020): Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 255-262, Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-952148-03-3

- Citations (excluding self-citations): 25

Md Mahfuz ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Philipp Koehn, Vassilina Nikoulina, Kweon Woo Jung (2021): Findings of the WMT Shared Task on Machine Translation Using Terminologies. In: *Proceedings of the Sixth Conference on Machine Translation*, pp. 652-663, Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-954085-94-7

- Citations (excluding self-citations): 15

Ivana Kvapilíková, Tom Kocmi, Ondřej Bojar (2020): CUNI Systems for the Unsupervised and Very Low Resource Translation Task in WMT20. In: *Fifth Conference on Machine Translation - Proceedings of the Conference*, pp. 1123-1128, Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-948087-81-0

- Citations (excluding self-citations): 5

Borek Požár, Klára Tauchmanová, Kristýna Neumannová, Ivana Kvapilíková, Ondřej Bojar (2022): CUNI Submission to the BUCC 2022 Shared Task on Bilingual Term Alignment. In: *Proceedings of the LREC 2022 15th Workshop on Building and Using Comparable Corpora*, pp. 43-49, European Language Resources Association, Paris, France, ISBN 979-10-95546-94-8

- Citations (excluding self-citations): 3

Ivana Kvapilíková, Dominik Macháček, Ondřej Bojar (2019): CUNI Systems for the Unsupervised News Translation Task in WMT 2019. In: *Fourth Conference on Machine Translation - Proceedings of the Conference*, pp. 241-248, Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-950737-27-7

- Citations (excluding self-citations): 3

Ivana Kvapilíková, Ondřej Bojar (2021): Machine Translation of Covid-19 Information Resources via Multilingual Transfer. In: *ITAT 2021 2nd Workshop on*

Automata, Formal and Natural Languages – WAFNL 2021, pp. 176-181, Faculty of Mathematics and Physics, Praha, Czechia

- Citations (excluding self-citations): 1

IVANA KVAPILÍKOVÁ, ONDŘEJ BOJAR (2022): CUNI Submission to MT4All Shared Task. In: *Proceedings of the LREC 2022 Workshop of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022)*, pp. 78-82, European Language Resources Association (ELRA), Paris, France, ISBN 979-10-95546-91-7

- Citations (excluding self-citations): 1

IVANA KVAPILÍKOVÁ, ONDŘEJ BOJAR (2023): Low-Resource Machine Translation Systems for Indic Languages. In: *Proceedings of the Eighth Conference on Machine Translation*, pp. 954–958, Association for Computational Linguistics: Stroudsburg, PA, USA

- Citations (excluding self-citations): 1

IVANA KVAPILÍKOVÁ, ONDŘEJ BOJAR (2023): Boosting Unsupervised Machine Translation with Pseudo-Parallel Data. In:*Proceedings of Machine Translation Summit XIX vol. 1: Research Track*, pp. 135-147, Asia-Pacific Association for Machine Translation (AAMT), Kyoto, Japan

- Citations (excluding self-citations): 0

Only publications relevant to this thesis are included. The number of citations was computed using Google Scholar and modified manually. Total number of citations (without self-citations): 81