CUNI Systems for the Unsupervised News Translation Task in WMT 2019

Ivana Kvapilíková **Dominik Macháček** **Ondřej Bojar**

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
`<surname>@ufal.mff.cuni.cz`

Abstract

In this paper we describe the CUNI translation system used for the unsupervised news shared task of the ACL 2019 Fourth Conference on Machine Translation (WMT19). We follow the strategy of Artetxe et al. (2018b), creating a seed phrase-based system where the phrase table is initialized from cross-lingual embedding mappings trained on monolingual data, followed by a neural machine translation system trained on synthetic parallel data. The synthetic corpus was produced from a monolingual corpus by a tuned PBMT model refined through iterative back-translation. We further focus on the handling of named entities, i.e. the part of vocabulary where the cross-lingual embedding mapping suffers most. Our system reaches a BLEU score of 15.3 on the German-Czech WMT19 shared task.

1 Introduction

Unsupervised machine translation is of particular significance for low-resource language pairs. In contrast to traditional machine translation, it does not rely on large amounts of parallel data. When parallel data is scarce, both neural machine translation (NMT) and phrase-based machine translation (PBMT) systems can be trained using large monolingual corpora (Artetxe et al., 2018b,c; Lample et al., 2018).

Our translation systems submitted to WMT19 were created in several steps. Following the strategy of Artetxe et al. (2018b), we first train monolingual phrase embeddings and map them to the cross-lingual space. Secondly, we use the mapped embeddings to initialize the phrase table of the PBMT system which is first tuned and later refined with back-translation. We then translate the Czech monolingual corpus by the PBMT system to produce several synthetic parallel German-Czech corpora. Finally, we train a supervised NMT system on a filtered synthetic data set, where we exclude sentences tagged as ”not Czech”, shuffle the word order and handle mistranslated name entities. The training pipeline is illustrated in Figure 1.

The structure of this paper is the following. The existing approaches used to build our system are described in Section 2. The data for this shared task is described in Section 3. Section 4 gives details on phrase embeddings. Section 5 describe the phrase-based model and how it was used to create synthetic corpora. Section 6 proceeds to the neural model trained on the synthetic data. Section 7 introduces our benchmarks and Section 8 reports the results of the experiments. Finally, Section 9 summarizes and concludes the paper.

2 Background

Unsupervised machine translation has been recently explored by Artetxe et al. (2018c,b) and Lample et al. (2018). They propose unsupervised training techniques for both the PBMT model and the NMT model as well as a combination of the two in order to extract the necessary translation information from monolingual data. For the PBMT model (Lample et al., 2018; Artetxe et al., 2018b), the phrase table is initialized with an n-gram mapping learned without supervision. For the NMT model (Lample et al., 2018; Artetxe et al., 2018c), the system is designed to have a shared encoder and it is trained iteratively on a synthetic parallel corpus which is created on-the-fly by adding noise to the monolingual text (to learn a language model by de-noising) and by adding a synthetic source side created by back-translation (to learn a translation model by translating from a noised source).

The key ingredient for functioning of the above mentioned systems is the initial transfer from a monolingual space to a cross-lingual space without using any parallel data. Zhang et al. (2017)

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 241–248
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

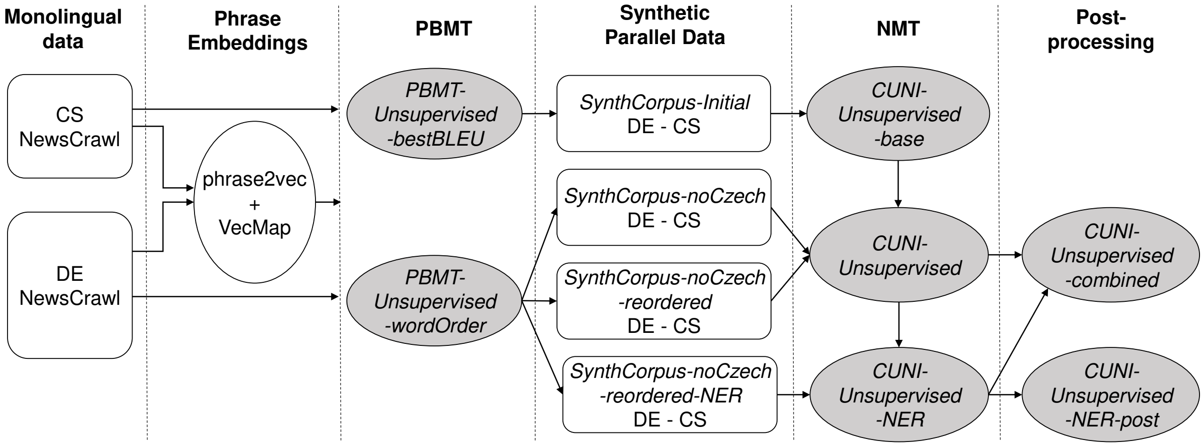
Monolingual data
Phrase Embeddings
PBMT
Synthetic Parallel Data
NMT
Post-processing
CS NewsCrawl
DE NewsCrawl
phrase2vec + VecMap
PBMT-Unsupervised -bestBLEU
PBMT-Unsupervised -wordOrder
SynthCorpus-Initial DE - CS
SynthCorpus-noCzech DE - CS
SynthCorpus-noCzech -reordered DE - CS
SynthCorpus-noCzech -reordered-NER DE - CS
CUNI-Unsupervised -base
CUNI-Unsupervised
CUNI-Unsupervised -NER
CUNI-Unsupervised -combined
CUNI-Unsupervised -NER-post

Figure 1: The training pipeline and an overview of our resulting systems. Corpora are displayed as rounded rectangles, MT systems as grey ovals.

and Conneau et al. (2018) 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.

3 Data

In line with the rules of the WMT19 unsupervised shared task, we trained our models on the NewsCrawl[1] corpus of newspaper articles collected over the period of 2007 to 2018.

We tokenized and truecased the text using standard Moses scripts. Sentences with less than 3 or more than 80 tokens were removed and the resulting monolingual corpora used for training of the unsupervised PBMT system consisted of 70M Czech sentences and 267M German sentences.

We performed further filtering of the Czech corpus before the NMT training stage. Since there are a lot of Slovak sentences in the Czech NewsCrawl corpus, we used a language tagger `langid.py` (Lui and Baldwin, 2012) to tag all sentences and remove the ones which were not tagged as Czech. After cleaning the corpus, the resulting Czech training set comprises 62M sentences.

Since small parallel data was allowed to tune the unsupervised system, we used newstest2013 for development of the PBMT system. Finally, we used newstest2012 to select the best PBMT model and newstest2010 as the validation set for the NMT model.

4 Phrase Embeddings

The first step towards unsupervised machine translation is to train monolingual n-gram embeddings and infer a bilingual dictionary by learning a mapping between the two embedding spaces. The resulting mapped embeddings allow us to derive the initial phrase table for the PBMT model.

4.1 Training

We first train phrase embeddings (up to trigrams) independently in the two languages. Following Artetxe et al. (2018b), we use an extension of the word2vec skip-gram model with negative sampling (Mikolov et al., 2013) to train phrase embeddings. We use a window size of 5, embedding size of 300, 10 negative samples, 5 iterations and no subsampling. We restricted the vocabulary to the most frequent 200,000 unigrams, 400,000 bigrams and 400,000 trigrams.

Having trained the monolingual phrase embeddings, we use *VecMap* (Artetxe et al., 2018a) to learn a linear transformation to map the embeddings to a shared cross-lingual space.

4.2 Output: Unsupervised Phrase Table

The output of this processing stage is the unsupervised phrase table which is filled with source and target n-grams. For the sake of a reasonable phrase table size, only the 100 nearest neighbors are kept as translation candidates for each source phrase. The phrase translation probabilities are de-

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

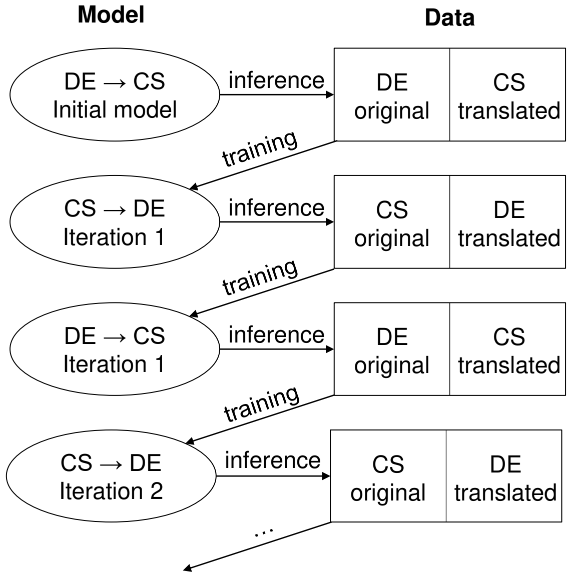

Figure 2: Step-by-step illustration of the iterative back-translation procedure.

rived from a softmax function over the cosine similarities of their respective mapped embeddings (Artetxe et al., 2018a).

5 PBMT Model

We followed the Monoses[2] pipeline of Artetxe et al. (2018b) for our unsupervised phrase-based system. The initial translation model is estimated based on the unsupervised phrase table induced from the mapped embeddings and the language model is estimated on the monolingual data. The reordering model is not used in the first step. The initial model is tuned and later iteratively refined by back-translation (Sennrich et al., 2016).

5.1 Training

The models are estimated using Moses (Koehn et al., 2007), with KenLM (Heafield, 2011) for 5-gram language modelling and fast_align (Dyer et al., 2013) for alignments. The feature weights of the log-linear model are tuned using Minimum Error Rate Training.

The back-translation process is illustrated in Figure 2. Both de→cs and cs→de systems are needed at this step. The de→cs system is used to translate a portion of the German monolingual corpus to Czech and create a synthetic parallel data set, which is then used to train the cs→de system and the procedure continues the other way around. We note that we do not make use of the initial model for cs→de. Once the synthetic parallel data set is created, the problem turns into a supervised one and we can use standard PBMT features, including the standard phrase table extraction procedure and the reordering model estimated on the aligned data sets.

Since back-translation is computationally demanding, we experimented with using a synthetic data set of 2 and 4 million sentences for back-translation rather than translating the whole monolingual corpus.

5.2 Output: PBMT Systems (cs→de)

We evaluated various PBMT models to select the best candidate to translate the whole monolingual corpus from Czech to German. The translation quality was measured on newstest2012.

We experimented with tuning the model both on an authentic parallel development set (3K sentence pairs) and a synthetic back-translated development set (10K sentence pairs). In the first scenario, possibly as a result of a smaller development set, the model started diverging after the first round of back-translation. In the second scenario, the best result is achieved after two and three rounds of back-translation for the cs→de and de→cs model, respectively (see the results in Table 1).

PBMT-Unsupervised-bestBLEU system

We selected the cs→de model with the highest BLEU of 14.22 for creating the synthetic corpus for the initial training of the NMT system. This PBMT model was tuned on a synthetic development set with two rounds of back-translation).

PBMT-Unsupervised-wordOrder system

However, after reviewing the translations and despite the BLEU results, we kept also the cs→de model with a BLEU score of 12.06 which was tuned on authentic parallel data. The translations were superior especially in terms of the word order.

5.3 Output: Synthetic Corpora

The training data sets for our NMT models were created by translating the full target monolingual corpus (filtered as described in Section 3) from Czech to German using the best performing cs→de PBMT models. Due to time constraints, we were gradually improving our PBMT models

[2] https://github.com/artetxem/monoses

Iteration No.	Authentic Dev Set		Synthetic Dev Set	
	de→cs	cs→de	de→cs	cs→de
Initial model	9.44	11.46	9.06	11.06
1	11.11	*12.06	4.61	12.92
2	7.26	6.78	11.70	**14.22
3	1.06	2.32	12.06	14.07
4	-	-	5.65	13.67
5	-	-	11.69	14.18
6	-	-	11.56	13.96

Table 1: Results of the PBMT models on newstest2012. The systems in left two columns were tuned on the parallel newstest2013 (3K sentence pairs) and iteratively refined on 2M sentence pairs. The ones in the right two columns were tuned on a synthetic set (10K back-translated sentence pairs) and iteratively refined on 4M sentence pairs. ** indicates the model selected for creating the synthetic training data for the initial training of the NMT model (*PBMT-Unsupervised-bestBLEU*). * indicates the model selected for creating the synthetic training data for further fine-tuning of the NMT model (*PBMT-Unsupervised-wordOrder*).

and already training the NMT model on the synthetic data. As a result, the final NMT model used synthetic data sets of increasing quality in four training stages.

5.3.1 Frequent Errors in Synthetic Corpora

We read through the translations to detect further error patterns which are not easily detectable by BLEU but have a significant impact on human evaluation. We noticed three such patterns:

- wrong word order (e.g. in contrast to the Czech word order, verbs in subordinate clauses and verbs following a modal verb are at the end of a sentence in German)

- unknown Czech words copied to German sentences during translation

- randomly mistranslated named entities (NEs) (e.g. *king Ludvik* translated as *king Harold* or *Brno* translated as *Kraluv Dvur*);

5.3.2 Heuristics to Improve Synthetic Corpora

In order to reduce the detrimental effects of the above errors, we created several variations of the synthetic corpora. Here we summarize the final versions of the corpora that served in the subsequent NMT training:

SynthCorpus-Initial

The *PBMT-Unsupervised-bestBLEU* model was used for creating the data set for the initial training of the model. All submitted systems were trained on this initial training set.

SynthCorpus-noCzech

This time we translated the Czech corpus by the *PBMT-Unsupervised-wordOrder* model. Despite its lower BLEU, the translations produced by this model seem more fluent. In order to remove Czech words from German sentences in the synthetic corpus, we identified words with Czech diacritics and replaced them on the German side with the *unk* token. As a result, the models trained on this corpus do not learn to simply copy unknown words and therefore, the German translations produced by such models rarely contain copied Czech words.

SynthCorpus-noCzech-reordered

The *SynthCorpus-noCzech* was further treated to improve the word order in the synthetic corpus. We shuffled words in the synthetic German sentences within a 5-word window and mixed the reordered sentences into the original ones. We essentially doubled the size of the training corpus by first reordering odd-indexed sentences while keeping even-indexed sentences intact and then vice versa.

The motivation for this augmentation was to support the NMT system in learning to handle word reordering less strictly, essentially to improve its word order denoising capability. Ideally, the model should learn that German word order need not be strictly followed when translating to Czech. This feature is easy to observe in authentic parallel texts but the synthetic corpora are too monotone. We are aware of the fact that a 5-word window is not sufficient to illustrate the reordering necessary for German verbs but we did not want to introduce too language-specific components to our technique.

SynthCorpus-noCzech-reordered-NER

The *SynthCorpus-noCzech-reordered* was further treated to alleviate the problem of mistranslated NEs present in the data.

NEs were identified in the monolingual Czech corpus by a NE recognition tagger NameTag[3] (Straková et al., 2014). The model was trained on the training portion of the Czech Named Entity Corpus 2.0[4] which uses a detailed two-level named entity hierarchy. We then used automatic word alignments (fast_align) between the Czech side and the synthetic German side of the corpus and checked the German counterparts of automatically-identified Czech NEs. If the German counterpart was close enough (Levenshtein distance of at most 3) to the Czech original, we trusted the translation. In other cases, we either copied the NE from the source or we used *unk* on the German side, preventing the subsequent NMT system from learning a mistranslation. Instead, the *unk* should never match any input and the NMT system should be forced to fall back to its standard handling of unknown words. Ideally, this would be to copy the word, but since there is no copy mechanism in our NMT setups, the more probable solution of the system would be to somehow circumvent or avoid the NE in the target altogether.

Named entity types and their treatment are listed in Table 2. Mistranslated NEs were treated in two stages. First during improving the synthetic corpora and then during post-processing, as described in Section 6.2.

6 NMT Model

6.1 Model and Training

We use the Transformer architecture by Vaswani et al. (2017) implemented in Marian framework (Junczys-Dowmunt et al., 2018) to train an NMT model on the synthetic corpus produced by the PBMT model. The model setup, training and decoding hyperparameters are identical to the CUNI Marian systems in English-to-Czech news translation task in WMT19 (Popel et al., 2019), but in this case, due to smaller and noisier training data, we set the dropout between Transformer layers to 0.3. We use 8 Quadro P5000 GPUs with 16GB memory.

Named Entity Type	Pre-treatment	Post-treatment
Numbers in addresses	copied	copied
Geographical names	removed	copied
Institutions	copied	ignored
Media names	copied	ignored
Number expressions	copied	copied
Artifact names	copied	ignored
Personal names	copied	copied
Time expressions	copied	ignored

Table 2: Named Entity types extracted from Czech Named Entity Corpus 2.0. and their treatment during pre-processing and post-processing. During *pre-treatment* (creation of the synthetic corpus), the NEs were identified in the Czech corpus and their translation on the German synthetic side was either *removed*, *copied* from the source Czech side or completely *ignored*. During *post-treatment* (post-processing of the final NMT outputs), the NEs were identified in the Czech translations and either *copied* from the source German side or *ignored*.

6.2 Post-processing

During post-processing of the translated Czech test set, we always adjusted quotation marks to suit Czech standards. Some systems were subject to further post-processing as indicated in the following section.

6.3 Output: NMT Systems

Our resulting systems share the same architecture and training parameters but they emerged from different stages of the training process as illustrated in Figure 1. The entire training process included training the system on the initial training corpus, fine-tuning on other corpora and final post-processing.

CUNI-Unsupervised-base

This system was trained on the initial synthetic data set *SynthCorpus-Initial* until convergence. We used early stopping after 100 non-improvements on validation cross-entropy, with validation step 1 000. The training finished after 3 days and 11 hours at 249 000 steps. Then we selected the checkpoint with the highest `bleu-detok`, which was at 211 000 steps, in epoch 3.

No further fine-tuning was performed. This system was not submitted to WMT19.

CUNI-Unsupervised

This system was fine-tuned on the *SynthCorpus-noCzech* corpus for 4 hours, when it reached

[3] http://ufal.mff.cuni.cz/nametag

[4] http://ufal.mff.cuni.cz/cnec/cnec2.0

System Name	BLEU uncased	BLEU cased	TER	BEER 2.0	CharacTER
CUNI-Unsupervised-base	13.6	13.3	0.799	0.482	0.688
CUNI-Unsupervised*	15.3	15.0	0.784	0.489	0.672
CUNI-Unsupervised-NER*	14.6	14.3	0.786	0.487	0.675
CUNI-Unsupervised-NER-post**	14.4	14.1	0.788	0.485	0.677
CUNI-Unsupervised-combined*	14.9	14.6	0.785	0.488	0.674
Benchmark-Supervised	19.3	18.8	0.719	0.517	0.636
Benchmark-TransferEN	13.6	13.3	0.793	0.482	0.683

Table 3: Our systems and their performance on newstest2019 (* indicates our WMT submissions and ** indicates our primary system).

a maximum, and for another 4 hours on *SynthCorpus-noCzech-reordered*.

CUNI-Unsupervised-NER

This system is a result of additional 4 hours of fine-tuning of the *CUNI-Unsupervised* system on the *SynthCorpus-noCzech-reordered-NER* corpus. Although the effect of this fine-tuning on the final translation might not be significant in terms of BLEU points, the problem of mistranslated named entities is perceived strongly by human evaluators and warrants an improvement.

CUNI-Unsupervised-NER-post

The translations produced by *CUNI-Unsupervised-NER* were post-processed to tackle the remaining problem with named entities. We first trained GIZA++ (Och and Ney, 2003) alignments on 30K sentences. We used NameTag to tag NEs in Czech sentences and using the alignments, we copied personal names, geographical names and numbers from the German source to the Czech target.

CUNI-Unsupervised-combined

We translated the test set by two models and combined the results. We used NameTag to tag Czech sentences with named entities and translated the tagged sentences by *CUNI-Unsupervised-NER*. The sentences with no NEs were translated by the *CUNI-Unsupervised* system.

7 Benchmarks

For comparison, we created a NMT system using the same model architecture as above but training it in a supervised way on the German-Czech parallel corpus from Europarl (Koehn, 2005) and OpenSubtitles2016 (Tiedemann, 2012), after some cleanup pre-processing and character normalization provided by Macháček (2018). As far as we know, these are the only publicly available parallel data for this language pair. They consist of 8.8M sentence pairs and 89/78M tokens on the German and the Czech side, respectively. The system *Benchmark-Supervised* was trained from scratch for 8 days until convergence.

Our other comparison system, *Benchmark-TransferEN*, was first trained as an English-to-Czech NMT system (see *CUNI Transformer Marian* for the English-to-Czech news translation task in WMT19 by Popel et al. (2019)) and then fine-tuned for 6 days on the *SynthCorpus-noCzech-reordered-NER*. The vocabulary remained unchanged, it was trained on the English-Czech training corpus. This simple and effective transfer learning approach was suggested by Kocmi and Bojar (2018).

The scores of the systems on newstest2019 are reported in Table 3.

8 Final Evaluation

The systems submitted to WMT19 are listed in Table 3 along with our benchmarks. In addition to BLEU, we also report BEER (Stanojević and Sima'an, 2014) and CharacTER (Wang et al., ”2016”) scores.

Table 5 summarizes the improvement we gained by introducing a special named entity treatment. We manualy evaluated three systems, *CUNI-Unsupervised, CUNI-Unsupervised-NER* and *CUNI-Unsupervised-NER-post* on a stratified subset of the validation data set created by randomly selecting 100 sentences with NEs and 100 sentences without NEs. The results are presented in two steps, the first table shows that fine-tuning the system *CUNI-Unsupervised-NER* on a synthetic corpus with amended NEs proved beneficial in 52% of tested sentences which included NEs and it did not harm in 20% of sentences. When comparing the two systems on sentences

Source	Phrase
Original	Der Lyriker **Werner Söllner** ist IM **Walter**.
Reference	Básník **Werner Söllner** je tajný agent **Walter**.
CUNI-Unsupervised	Prozaik **Filip Bubeníček** je agentem StB **Josefem**.
CUNI-Unsupervised-NER	Prozaik **Filip Söllner** je agentem StB **Ladislavem Bártou**.
CUNI-Unsupervised-NER-post	Prozaik **Werner Söllner** je agentem StB **Walter**.

Table 4: Sample translations showing that fine-tuning on synthetic corpus with cleaned NEs (*CUNI-Unsupervised-NER*) alleviates a part of the NE problem while post-processing can handle the rest. However, note the imperfect translation of *Lyriker* as *novelist* rather than *poet* and the extra word *StB* which was not tagged as a NE and therefore not treated during post-processing.

Winning Systems	Sentences with NEs	Sentences with no NEs
CUNI-Unsup	28%	26%
CUNI-Unsup-NER	52%	28%
No winner	20%	46%

Winning Systems	Sentences with NEs	Sentences with no NEs
CUNI-Unsup-NER	14%	0%
CUNI-Unsup-NER-post	18%	0%
No winner	68%	100%

Table 5: 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.

with no NEs, their performance is very similar.

Furthermore, adjusting NEs during post-processing proved useful in 18% of sentences with NEs and it did not harm in 68% of sentences. Post-processing introduced two types of errors: copying German geographical names into Czech sentences (e.g. translating *Norway* as *Norwegen* instead of *Norsko*) and replacing a Czech named entity with a word which does not correspond to it due to wrong alignments (e.g. translating *Miss Japan* as *Miss Miss*). On the other hand, when alignments were correct, the post-processing was able to fix remaining mismatches in named entities. See Table 4 for a sample translation.

9 Conclusion

This paper contributes to recent research attempts at unsupervised machine translation. We tested the approach of Artetxe et al. (2018b) on a different language pair and faced new challenges for this type of translation caused by the non-similar nature of the two languages (e.g. different word order, unrelated grammar rules).

We identified several patterns where the initial translation models systematically failed and we focused on alleviating such issues during fine-tuning of the system and final post-processing. The most severe type of a translation error, in our opinion, was a large number of randomly mistranslated named entities which left a significant impact on the perceived translation quality. We focused on alleviating this problem both during fine-tuning of the NMT system and during the post-processing stage. While our treatment is far from perfect, we believe that an omitted named entity or a non-translated named entity causes less harm than a random name used instead.

While the performance of our systems still lags behind the supervised benchmark, it is impressive that the translations reach their quality without ever seeing an authentic parallel corpus.

Acknowledgments

This study was supported in parts by the grants SVV 260 453, 1050119 of the Charles University Grant Agency, 18-24210S of the Czech Science Foundation and the EU grant H2020-ICT-2018-2-825460 (ELITR).

This work has been using language resources and tools stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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

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

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

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Dominik Macháček. 2018. Enriching Neural MT through Multi-Task Training. Master's thesis, Institute of Formal and Applied Linguistics, Charles University.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc.

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

Martin Popel, Dominik Macháček, Michal Aueršperger, Ondřej Bojar, and Pavel Pecina. 2019. English-czech systems in wmt19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation: Volume 2, Shared Task Papers*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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*.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. "2016". Character: Translation edit rate on character level. In *"Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers"*.

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: Volume 1, Long Papers*.