The LMU Munich Unsupervised Machine Translation System for WMT19

Dario Stojanovski, Viktor Hangya, Matthias Huck and **Alexander Fraser**
Center for Information and Language Processing
LMU Munich
{stojanovski,hangyav,mhuck,fraser}@cis.lmu.de

Abstract

We describe LMU Munich's machine translation system for German→Czech translation which was used to participate in the WMT19 shared task on unsupervised news translation. We train our model using monolingual data only from both languages. The final model is an unsupervised neural model using established techniques for unsupervised translation such as denoising autoencoding and online back-translation. We bootstrap the model with masked language model pretraining and enhance it with back-translations from an unsupervised phrase-based system which is itself bootstrapped using unsupervised bilingual word embeddings.

1 Introduction

In this paper we describe the system we developed at the LMU Munich Center for Information and Language Processing, which we used to participate in the unsupervised track of the news translation task at WMT19. The system builds on our last year's submission to the unsupervised shared task (Stojanovski et al., 2018) and previous work on unsupervised machine translation (Lample et al., 2018a; Artetxe et al., 2018c; Lample et al., 2018b; Lample and Conneau, 2019). We submitted system runs for the German→Czech translation direction. The goal of the unsupervised track is to train machine translation models without access to any bilingual or comparable monolingual data.

Supervised neural machine translation (NMT) has achieved state-of-the-art results (Bahdanau et al., 2015). With the introduction of the Transformer (Vaswani et al., 2017) the quality of automatic translations has been significantly improved. However, a prerequisite for high performance has been access to large scale bilingual data. Naturally, this is not available for many language pairs and specific domains. Moreover, Koehn and Knowles (2017) also show that in low-resource setups neural models fail to match traditional phrase-based systems in terms of quality. This is the motivation for the unsupervised track at WMT19.

The system we use to participate in the shared task is multipart and borrows on existing techniques for unsupervised learning. We make use of bilingual word embeddings (BWE), phrase-based translation (PBT), cross-lingual masked language models (MLM) and NMT models, all trained in an unsupervised way. Lample et al. (2018a) and Artetxe et al. (2018c) showed that, given proper bootstrapping, it is possible to train unsupervised NMT models by making use of two general techniques, denoising autoencoding and online back-translation. Lample et al. (2018b) and Artetxe et al. (2018b) further showed that this is also possible for phrase-based statistical machine translation. A key technique that enables this is obtaining word-by-word translations by utilizing unsupervised bilingual word embeddings. Lample et al. (2018b) further simplified the bootstrapping step by showing that jointly trained BPE-level (Sennrich et al., 2016) embeddings are a better alternative, assuming closely related languages that potentially share surface forms. Lample et al. (2018b) also showed that a single shared encoder and decoder are sufficient for learning both translation directions. A general trend in NLP recently has been unsupervised masked language model pretraining. Devlin et al. (2018) showed that a wide range of NLP tasks are significantly improved by fine-tuning large MLM. They propose a way to train a Transformer language model which has access to left and right context as opposed to traditional LM which only have left context access. Lample and Conneau (2019) extended the approach to a multilingual setting and showed that this vastly outperforms the previous approaches for bootstrapping NMT models.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 393–399
Florence, Italy, August 1-2, 2019.

The model we used to participate in the shared task makes use of several of the aforementioned techniques. We train unsupervised BWEs and use them to bootstrap an unsupervised PBT model. We use large scale German and Czech monolingual NewsCrawl data to train a cross-lingual masked language model in order to bootstrap our unsupervised NMT model which itself is trained using denoising autoencoding and online back-translation. We combine all of these techniques and obtain competitive results in the shared task.

2 Bilingual Word Embeddings

Recently, many works showed that good quality bilingual word embeddings can be produced by using only monolingual resources (Conneau et al., 2017; Artetxe et al., 2018a; Dou et al., 2018). Most of these techniques follow a two-step approach involving (i) training monolingual vector spaces for both languages using large amount of monolingual data and (ii) projecting them to a shared bilingual space. We use the approach of (Conneau et al., 2017) which employs adversarial training to build bilingual word embeddings for the initialization of the phrase table used by our PBT system.

A general approach to measure word similarity in embedding spaces is to calculate their *cosine* similarity. A disadvantage of this approach is caused by the so called hubness problem of high dimensional spaces (Dinu et al., 2015), i.e., some words are similar to a high proportion of other words although their meaning is not necessarily close. To overcome the problem, the cosine similarity based *Cross-Domain Similarity Local Scaling* (CSLS) metric was proposed (Conneau et al., 2017). In short, this metric adjusts the similarity values of a word based on the density of the area where it lies, i.e., it increases similarity values for a word lying in a sparse area and decreases values for a word in a dense area. We use CSLS to create a dictionary of the 100 nearest target words for each source language word with their similarities which we convert to a phrase table. For more details on phrase-table creation see Section 3.

One problem with the approach arises when translating German compound words which are combinations of two or more words that function as a single unit of meaning. In most of the cases, these words should be translated into multiple Czech words, but our generated dictionary contains only 1-to-1 translations. In our previous work (Stojanovski et al., 2018), we experimented with bigrams in addition to unigrams in order to overcome this issue. We looked for frequent bigrams in the non-German side of the monolingual input data and trained separate embeddings for bigrams. Similarly, in the system of Artetxe et al. (2018c) embeddings for word n-grams are learned. The disadvantage of this approach is the lack of ability to represent previously unseen n-grams. It also significantly increases the size of the vocabulary. Since new compounds are constantly created in the German language, this could cause problems when using the system in the long run. To tackle the problem we applied the inverse of the approach and used compound splitting on all the German data. In this way we kept the vocabulary size relatively low and our system can handle novel compound words. A negative aspect of our approach is that non-compositional nouns could be incorrectly translated.

3 Unsupervised Phrase-based Translation

We build on the BWEs to create an unsupervised phrase-based translation system using the Moses decoder (Koehn et al., 2007).

In an initial step (*iteration 0*), a bilingual word-based translation lexicon is obtained from the embeddings space and stored in a format compatible with Moses' phrase table. The BWE cosine similarities serve as translation feature scores. We include multiple single-word target-side translation candidates per source-side token, given as the nearest neighbors in the bilingual embeddings space. An n-gram language model trained on target-side monolingual data is provided to Moses as another feature function. Moses then decodes with a variant of a beam search algorithm. We tune scaling factors to combine the feature functions.[1]

In a next step (*iteration 1*), synthetic parallel data is produced in order to acquire multi-word phrase table entries and improve over the initial simple word-based Moses translation system. To this end, we prepare an *iteration 0* Moses setup for the inverse translation direction (cs→de) as well and use it to translate a larger-sized Czech monolingual corpus (NewsCrawl 2018) into German. The Czech side of the resulting synthetic bitext is

[1]Note that a small parallel corpus (newstest2009) is utilized to tune the scaling factors.

original human-created data, whereas the German side is noisy machine translation output from our *iteration 0* Czech→German unsupervised PBT engine. When machine-translating the monolingual corpus, we let the Moses decoder also write out the word alignment information. From this synthetic German-Czech bitext, a phrase table for the German→Czech translation direction can be extracted and a new German→Czech Moses PBT system can be built in the usual manner. We opted to switch off reordering in the *iteration 0* setup, but now allow for reordering in *iteration 1*. We also add word penalty, phrase penalty, and distance-based distortion cost feature functions and tune the scaling factors again.

The process of producing synthetic parallel data can be repeated, which we do for one more step (*iteration 2*). The idea here is to also improve the inverse translation system by means of building an *iteration 1* system for the Czech→German direction as well through machine-translating German monolingual training data (the German NewsCrawl 2018 corpus) to Czech using the initial German→Czech unsupervised PBT engine. The improved inverse-direction system is then applied to back-translate the Czech monolingual training corpus once again and achieve better quality of the synthetic bitext. The *iteration 2* German→Czech is trained with a phrase table extracted from that higher-quality synthetic bitext. The systems in the two translation directions can benefit from each other in the course of the reciprocal re-training procedure. Translation quality in both directions is gradually improved.

4 Unsupervised Neural Translation

4.1 Masked Language Model Pretraining

We use the MLM approach proposed in Lample and Conneau (2019) to pretrain our NMT model. The MLM is trained by masking a percentage of the tokens which then the model is tasked to predict. Lample and Conneau (2019) extend MLM in a multilingual context by adding language-specific embeddings and using monolingual data from multiple languages. We train a MLM with German and Czech monolingual data. We randomly sample 15% of the input tokens and mask 80% of those with [MASK], swap 10% with a random token and in 10% of cases we keep the original token. We train a 6-layer Transformer with 8 attention heads, and an embedding and layer size of 1024. The size of the position-wise feed-forward neural network is 4096. We use dropout of 0.1, GELU activations (Hendrycks and Gimpel, 2017) and learned positional embeddings. The model is trained with batches of 32 streams of continuous sentences composed of 256 tokens. For further details, we refer to Lample and Conneau (2019). The model was trained for 7 days and subsequently used to initialize the encoder and decoder of the NMT model.

4.2 Denosing Autoencoding and Online Back-translation

As with previous work (Artetxe et al., 2018c; Lample et al., 2018a,b; Lample and Conneau, 2019) we train an unsupervised NMT model with denoising autoencoding and online back-translation. It is important to properly bootstrap this model in order to enable the model to get off the ground. In previous work this was made possible by using word-by-word translations or jointly trained BPE-level word embeddings. We bootstrap the model with the pretrained cross-lingual MLM as in Lample and Conneau (2019).

Although we initialize the model with a pretrained cross-lingual MLM, it is still necessary to use denoising autoencoding. Since the LM is trained with the masked LM objective, it is reasonable to assume that it has not learned language-specific reorderings which are key for machine translation. The denoising autoencoding is trained by feeding in a noisy version of a sentence and trying to reconstruct the original version. The noisy sentences are created by dropping words with probability of 0.1, shuffling words within a range of 3 and masking them with a probability of 0.1. In this way, the model is trained to produce fluent output. Furthermore, denoising autoencoding enables the model to learn important reorderings, insertions and deletions.

The translation component of the network is trained by first using the model in inference mode to produce back-translations. The back-translations are coupled with the original sentences to create pseudo-parallel data and train the model in a traditional fashion.

We train a single joint model using both techniques on both language directions. The goal is to end up with a model capable of translating from German into Czech. However, since the model de-

pends on having quality German→Czech translations, it is important to be able to produce German back-translations from Czech. As a result, we train the model in both language directions.

The model has a single shared encoder and decoder, each equipped with 6 layers and 8 attention heads per layer. The batch size is 1600 tokens. We apply dropout of 0.1. We share the source, target and output embeddings and also share them across the two languages.

4.3 Incorporating PBT Synthetic Data

The training curriculum to enable this model to work is to first pretrain a cross-lingual MLM. Subsequently, one can further bootstrap this model with back-translations from an unsupervised phrase-based system and finally, fine-tune this model with the unsupervised neural criteria. However, due to time constraints we first fine-tune the pretrained MLM with the NMT system. After several iterations of training, we include additional back-translations from the phrase-based system. We only used pseudo-parallel German→Czech translations. We continue using online back-translation during this fine-tuning stage, but not denoising autoencoding. For the primary submission at WMT19, we used back-translations from *iteration 0* from the phrase-based system. In subsequent experiments, we also trained a model with data from *iteration 1*.

5 Experiments and Empirical Evaluation

5.1 Data and Preprocessing

As monolingual data in this work we used German and Czech NewsCrawl articles from 2007 to 2018. In the case of both languages the corpora contained a small set of sentences coming from foreign languages which we filtered out using a language detection tool[2]. The datasets were tokenized and truecased with the standard scripts from the Moses toolkit (Koehn et al., 2007).

For the bilingual word embeddings used by our PBT system we compound split the German corpus using `compound-splitter.perl` from the Moses toolkit with the following parameters: minimum word size 4; minimum count 5; maximum count 1000. To train monolingual word embeddings we used *fasttext* (Bojanowski et al., 2017), instead of *word2vec* (Mikolov et al., 2013), which performs better on morphological rich languages by employing subword information. We used 300 dimensional embeddings and default values for the rest of the parameters. For the unsupervised mapping we used *MUSE* (Conneau et al., 2017) with default parameters, but restricting the vocabulary size for both source and target languages to the most frequent $200K$ words due to memory considerations.

We used BPE segments in the case of our neural system. The segmentation was computed jointly on all the NewsCrawl data available for both languages using $32K$ merge operations. We train the cross-lingual MLM with German NewsCrawl 2017-2018, and Czech NewsCrawl 2007-2018 monolingual data. For the unsupervised NMT model, we use NewsCrawl 2018 for German and NewsCrawl 2013-2018 for Czech. In this way, both models are trained with roughly equal amounts of German and Czech data. Details on the training data is in Table 1. For the NMT experiments, we use the code from (Lample and Conneau, 2019)[3].

In the following we perform evaluation for both our unsupervised phrase-based and neural machine translation systems. We report BLEU scores on the detokenized translations of newstest2013 and newstest2019 using *sacreBLEU*[4] (Post, 2018).

model	de	cs
BWE	270M	67M
MLM	75M	67M
PBT	270M	67M
NMT	37M	41M

Table 1: Training data sizes in number of sentences.

5.2 PBT Experiments

As mentioned earlier we initialize our PBT system with BWEs trained on compound split data. In Table 2 we show baseline word-by-word (*wbw*) results, i.e., we greedily translate each source word independently of the others using the most similar target word, according to the BWE-based dictionary, without any reordering. We compare BWEs trained with and without compound split data. The results of both approaches are low, which is due to the morphological richness of the target lan-

[2] https://github.com/indix/whatthelang

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

[4] https://github.com/mjpost/sacreBLEU

	newstest2013 de→cs
wbw	4.2
wbw+comp. split	4.3
unsup. PBT iter. 0	6.0
unsup. PBT iter. 1	7.9
unsup. PBT iter. 2	8.4

Table 2: Baseline results (BLEU) with word-by-word translations (wbw) and unsupervised phrase-based translations (PBT) on newstest2013. We compare wbw results with and without compound splitting on the German language side. For the unsupervised PBT experiments, German is compound-split.

	newstest2013 de→cs
unsup. NMT	17.0
unsup. NMT + PBT iter. 0	18.5
+ fine-tune no PBT	18.3
+ fine-tune PBT iter. 1	18.8
unsup. NMT + PBT iter. 1	19.1

Table 3: BLEU scores with the unsupervised NMT systems on newstest2013.

	newstest2019 de→cs
unsup. NMT	16.2
*unsup. NMT + PBT iter. 0	17.0
‡unsup. NMT + PBT iter. 0	17.6
+ fine-tune no PBT	17.4
+ fine-tune PBT iter. 1	17.8
unsup. NMT + PBT iter. 1	17.8

Table 4: BLEU scores with the unsupervised NMT systems on newstest2019. * - primary submission, trained for ∼12h. ‡- trained for ∼62h.

guage. On one hand, based on manual investigation[5] of the BWE-based dictionary and the sentence translations, we conclude that the various inflected forms of the correct Czech stems are often the most similar translations of given German words. On the other hand, without the context it is much harder to pick the right form as opposed to some other language pairs such as German and English. Compound splitting resulted in performance increase of the system which is due to the translation of German compounds to multiple Czech words. In addition, it also helped lowering the number of Out-Of-Vocabulary (OOV) words which is partly due to limiting the size of the vocabulary.

Table 2 also presents the results from our PBT system. At *iteration 0* the model obtains 6.0 BLEU on newstest2013. The score increased to 7.9 BLEU at *iteration 1* and to 8.4 at *iteration 2*.

5.3 NMT Experiments

We show the results from our unsupervised neural model and the combination with synthetic data from the phrase-based system. Our primary submission at WMT19 has achieved competitive results despite using a single model with no ensembling. The model for the primary submission was trained for ∼12h due to time constraints. For the contrastive experiments we present in Table 3 we further trained this model for ∼62h overall. We train the models on 8 Nvidia GTX 1080 Ti with 12 GB RAM.

We present results on newstest2013. For model selection we used newstest2009. The first row in Table 3 shows our baseline unsupervised neural system. This model achieves significant improvements over the word-by-word approach and PBT system. All results except for the *unsup. NMT* baseline are obtained by applying compound splitting to the German input from newstest2013. We present the result for the baseline without compound splitting because the initial cross-lingual MLM and unsupervised NMT system were trained with German monolingual data which was not compound split. However, the BWEs and PBT system were trained with compound split German monolingual data and as a result the German back-translations we obtain from the PBT system were compound split. Consequently, all contrastive models where we fine-tune the original unsupervised NMT system are trained with compound split German monolingual data. However, we do not observe any adverse effects on translation quality. Furthermore, the results from the fine-tuned models show that very similar results are obtained with both versions of the test set.

When fine-tuning our model with PBT synthetic data, we disable denoising autoencoding, but continue to do online back-translation. Even though we used PBT synthetic data from *iteration 0*, we observe significant improvements. We fine-tune the model for ∼62h and BLEU score was improved from 17.0 to 18.5. We use this model for

[5]Note that none of the authors speak the target language.

the primary submission, but a version which was trained for ∼12h only. We intuitively assumed that removing this data and continuing training with online back-translation only would further improve performance. However, we observe that BLEU score decreased to 18.3.

We also experimented with adding PBT synthetic data from *iteration 1*. We tried adding this data as we did with the back-translations from *iteration 0*. Furthermore, we also tried fine-tuning the model trained on *iteration 0* data with data from *iteration 1*. For this setup, the data from *iteration 0* was removed. It is interesting that fine-tuning the initial unsupervised NMT obtains better performance than fine-tuning the model trained with *iteration 0* data. The best score we managed to obtain was 19.1 by fine-tuning the initial unsupervised NMT with *iteration 1* data and translating a compound split version of newstest2013.

In Table 4 we show the results on newstest2019. Our primary submission obtained 17.0 BLEU. Further training and including synthetic data from *iteration 1* increased the score to 17.8 BLEU.

6 Conclusion

In this work, we present LMU Munich's unsupervised system for German→Czech news translations. We developed unsupervised BWEs, phrase-based and neural systems and studied different ways of combining them. We show that an unsupervised neural model pretrained with large cross-lingual masked language model is superior to unsupervised phrase-based model for this language pair. Despite working on a Germanic-Slavic language pair, the unsupervised methods for machine translation work well and provide for a relatively good translation quality.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable input. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 640550).

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)*, pages 789–798.

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*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised Neural Machine Translation. In *International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15. ArXiv: 1409.0473.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *CoRR*, abs/1710.04087.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving Zero-Shot Learning by Mitigating the Hubness Problem. In *Proceedings of the International Conference on Learning Representations: Workshop Track*.

Zi-Yi Dou, Zhi-Hao Zhou, and Shujian Huang. 2018. Unsupervised Bilingual Lexicon Induction via Latent Variable Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 621–626.

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

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

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-Based & Neural Unsupervised Machine Translation. *arXiv preprint arXiv:1804.07755*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser. 2018. The LMU Munich Unsupervised Machine Translation Systems. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 517–525, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.