The LMU Munich Unsupervised Machine Translation Systems

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 unsupervised machine translation systems for English↔German translation. These systems were used to participate in the WMT18 news translation shared task and more specifically, for the unsupervised learning sub-track. The systems are trained on English and German monolingual data only and exploit and combine previously proposed techniques such as using word-by-word translated data based on bilingual word embeddings, denoising and on-the-fly backtranslation.

1 Introduction

The LMU Munich's Center for Information and Language Processing participated in the WMT 2018 news translation shared task for English↔German translation. Specifically, we participated in the unsupervised learning task which focuses on training MT models without access to any parallel data. The team has a strong track record at previous WMT shared tasks (Bojar et al., 2017, 2016, 2015, 2014, 2013) working on SMT systems (Cap et al., 2014, 2015; Weller et al., 2013; Sajjad et al., 2013; Huck et al., 2016; Peter et al., 2016; Tamchyna et al., 2016) and proposed a top scoring linguistically informed neural machine translation system (Huck et al., 2017) based on human evaluation at WMT17.

Neural machine translation (NMT) is state-of-the-art in automatic translation. Attention-based neural sequence-to-sequence models (Bahdanau et al., 2015) have been established as the basis for most recent work in MT and furthermore, have been used to obtain best scoring systems at WMT in recent years (Bojar et al., 2017, 2016). Previous work and the best scoring systems at WMT also showed that NMT can be scaled to millions of sentence pairs and even achieve human parity (Hassan et al., 2018). However, this comes under the caveat that we have access to a large amount of human-translated parallel data. Koehn and Knowles (2017) showed that NMT models cannot be properly trained under low resource conditions and are still behind phrase-based models. In extremely low resource scenarios, NMT fails completely which is a big obstacle if we want to enable automatic translation over a variety of languages. This motivates the unsupervised learning task at WMT this year. The task is run for three language pairs, but we only focus on English↔German translation. Although this language pair has an abundance of parallel data, we are constrained to only using monolingual data provided for the WMT18 news translation task, excluding Europarl and News Commentary because of content overlap.

The systems we use for our submissions are based on the recently proposed techniques for unsupervised machine translation by several studies (Artetxe et al., 2018; Lample et al., 2018a,b). The phrase-based unsupervised system uses bilingual word embeddings (BWEs) to create an initial phrase table and also utilizes a target-side n-gram language model. The backbone of the unsupervised NMT methods is denoising and on-the-fly backtranslation which enable a standard NMT architecture to be trained by only leveraging monolingual data. The model for our submission is mostly based on the work of Lample et al. (2018b). Additionally, we explore how word-by-word translated data based on BWEs can be utilized to improve the initial training and experiment with different ways of producing these translations. We also show that disabling denoising in the last stages of learning can provide for further improvements. We refer the reader to Huck et al. (2018) for our supervised systems for news and biomedical translation.

The remainder of the paper outlines the methods we used for generating BWEs, training a phrase-based and neural unsupervised machine translations systems. Moreover, it presents the obtained results as well as translation examples showcasing some of the strong and weak points of the NMT system.

2 Bilingual Word Embeddings

Both our phrase-based and neural unsupervised machine translation systems are based on bilingual word embeddings which represent source and target language words in a shared vector space. Recently, Conneau et al. (2017) showed that good quality bilingual embeddings can be produced by training monolingual models for both source and target languages and mapping them to a shared space without any bilingual signal. We follow this approach and use bilingual word embeddings, trained in an unsupervised fashion, to jump-start both of our systems.

As our baseline system we produce **word-by-word** translations relying only on the embeddings. For each word w_s in the source sentence we induce its translation:

$$tr_{wbw}(w_s) = \arg\max_{w \in V_t} cos(e(w_s), e(w))$$

where $e(w)$ is the vector representation of word w, $cos(x, y)$ is the cosine similarity of two vectors and V_t is the target vocabulary.

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 English words which causes errors when translating them word by word. The issue is also present when translating from English to German since multiple words should be transformed into one unit. To overcome this issue we experimented with **bigrams** in addition to unigrams. We tried a simple idea, namely, we looked for frequent bigrams in the English side of both the monolingual input data and the test set. We replaced bigrams with their concatenated forms in the sentences and also kept the original sentence. By training bilingual word embeddings on this data we automatically allow the word-by-word algorithm to translate compound words to bigrams and vice-versa.

To further improve the quality of our algorithm, we exploited **orthographic similarity** of words. Braune et al. (2018) showed that the performance of inducing word translations can be significantly improved using orthography. Following the approach there, we obtained improvements, especially when translating named entities, by using the following word translation function:

$$tr_{orth}(w_s) = \arg\max_{w \in V_t} \max \begin{pmatrix} cos(e(w_s), e(w)), \\ \lambda * orth(w_s, w)) \end{pmatrix}$$

where λ is a weighting constant and $orth(w_1, w_2)$ is the normalized Levenshtein distance of words w_1 and w_2.

As a contrastive set of experiments we added light supervision during the training of bilingual word embeddings in order to show performance differences compared to the fully unsupervised setup. To map monolingual spaces we used orthogonal mapping (Xing et al., 2015) with a seed lexicon of of 5000 word pairs, which was used as a baseline in (Conneau et al., 2017) as well.

2.1 Technical Details

To train monolingual word embeddings we used *fasttext* (Bojanowski et al., 2017) which employs subword information for better quality representations. We used 512 dimensional embeddings and default values for the rest of the parameters. For both unsupervised and lightly supervised mapping we used *MUSE* (Conneau et al., 2017) with default parameters. We fine-tuned λ on the test set of *WMT 2017* and used the method of (Mikolov et al., 2013) to mine frequent bigrams.

3 Unsupervised Phrase-based Translation

We have investigated unsupervised phrase-based translation (PBT). The results have been worse than with the neural model in our experiments. In this section, we therefore only give a short outline of the methods which we have explored in that area.

By means of a straightforward format conversion of the BWE lexicon, we can create a word-based "phrase table" that can be loaded into the Moses decoder (Koehn et al., 2007). The cosine similarities from the BWE model become feature scores in the phrase table. Note that we refrained from normalizing the cosine similarities, but wrote their values directly to the table.

Using Moses for decoding carries the advantage that an n-gram language model can be integrated

without any implementation effort. Once we have added a language model, we can also activate reordering. A distance-based distortion cost may then be added as a further feature.

An obvious difficulty is how to choose the weights for the features. If we assume that a small amount of bitext is actually available (say, a few hundred sentence pairs), then we can tune the weights with MERT or MIRA. We did the latter and built tuned unsupervised phrase-based systems in the outlined way for both translation directions.

With this initial system, we created synthetic training data. We translated around 50 M monolingual sentences from German to English. Not only the translations, but also the decoding word alignments were stored. Next, phrases can be extracted from the synthetic parallel corpus. We can use this new phrase table in the Moses decoder to build a better English→German unsupervised phrase-based system. The feature weights can be tuned again with MERT/MIRA. Word penalty and phrase penalty become useful with the phrase table from synthetic data. The new phrase table contains phrases of different lengths, not only words (or word bigrams).

We trained an English→German unsupervised phrase-based system according to the pipeline that we just described. Its output was uploaded as a contrastive submission, but we decided to not earmark it for manual evaluation.

4 Unsupervised Neural Translation

The system we used in this work builds on previous work on unsupervised neural machine translation (Artetxe et al., 2018; Lample et al., 2018a,b). We mostly make use of the techniques suggested in Lample et al. (2018b).

Before training the unsupervised NMT system proposed in Lample et al. (2018b), it is important to properly initialize certain key components which are otherwise randomly initialized. For that purpose, they propose to initialize the encoder and decoder embeddings with BPE-level embeddings trained using *fasttext* (Bojanowski et al., 2017). The BPE splitting is computed jointly on the German and English monolingual data. Given that these two languages are related and share surface forms, this technique is a reasonable choice.

The model proposed in Lample et al. (2018b) consists of two main components, a denoising and a translation component. The denoising part acts as a language model and is trained to produce fluent output in a given language based on a noisy version of the input. We follow the implementation of Artetxe et al. (2018) where the noisy version of the input sentence is obtained by making random swaps of contiguous words. Denoising helps to produce fluent output, but it's also used to enable reordering, and insertions and deletions of words. This is necessary since the model initially tends to do word-by-word translations while in German and English the word order is different.

The translation component works in a traditional way. However, given that the model doesn't have access to parallel data, it needs to make use of on-the-fly backtranslation. During training, the same model is used to backtranslate a sentence from the monolingual data and this pair of backtranslated sample/gold standard sample is used to train the model in a traditional fashion.

In order to enable for the denoising, or language model effects to be transferred to the translation component, many parameters in the model are shared. The encoder is shared for German and English. This forces the model to produce a language-agnostic representation of the input sentence. It also enables for the decoder and the attention mechanism to be shared across both languages. Although the decoder is shared, a language identifier token is added at the beginning of each sentence only on the target side. In our experiments, we observed problems if we try to share the softmax layer, because the output tended to be a mixture of both German and English.

In the model used for our final submission, we use all of the outlined techniques from Lample et al. (2018b). However, we used additional data in the initial learning procedure and modified the training curricula in order to improve performance. In our experiments, we observed some initial training difficulties. As a result, in order to facilitate faster and easier learning, we make use of word-by-word translated synthetic parallel data, in addition to initializing the encoder and decoder embeddings. In our model, the training consists of alternative batches of monolingual data used for denoising and backtranslation and the word-by-word translated synthetic data. The word-by-word translations are obtained as described in Section 2. We also apply BPE splitting on this data before using it in training.

After a certain number of iterations, we stop with the training of the initial model and "unplug" two components of the previous training procedure. Namely, we remove the word-by-word translated data since this is useful to jump-start the learning, but later presumably will impede learning more nuanced translations. We also observe better results if we disable the denoising component and continue the training by only doing on-the-fly backtranslation. This improved results on both translation directions by more than 1 BLEU (Papineni et al., 2002). However, in subsequent experiments we observed that this can also lead to unstable learning and decrease the performance since bad translation decisions can be reinforced. As a result, the final training procedure should be carefully controlled.

As mentioned in Section 2, the model has problems translating named entities. This stems from the fact that it is dependent on BWEs, where two different named entities often mistakenly have similar representations, causing confusion. Following the improvements the word-by-word translation obtained by using orthographic similarity, we also try training a model with word-by-word translated data utilizing this similarity. We also use word-by-word translated data obtained by using bigrams and orthographic similarity.

5 Empirical Evaluation

The models in this work are trained on German and English NewsCrawl articles from 2007 to 2017. Since the total size of this data is very large, we randomly sampled 4M sentences for each language. Moreover, we study if there is any noticeable effect if we only utilize more recent data. As a result, we sampled 4M samples from NewsCrawl 2017 and report results with this dataset as well.

The datasets are tokenized and truecased with the standard scripts from the Moses toolkit (Koehn et al., 2007). When training the truecase models, we actually use all of the available NewsCrawl data, rather than our subsample. We also use BPE splitting. The BPE segmentation is computed jointly on all the NewsCrawl data available for both languages. Then, all sentences with more than 50 tokens are discarded. The NewsCrawl data is also used to train the BPE-level embeddings.

We implement our neural system on top of the code made available by Artetxe et al. (2018). The model is an attention-based encoder-decoder NMT with 2-layer GRU encoder and decoder. The number of hidden units is 600. We set the learning rate to 0.0002 and dropout in the encoder and decoder to 0.3. We checkpoint the model each 10K updates. The batch size is 32.

	BWE unsupervised	
	de-en	en-de
wbw	11.50	6.94
wbw+bigram	11.77	6.75
wbw+orth	12.37	7.92
wbw+orth+bigram	12.58	7.64
	BWE lightly supervised	
	de-en	en-de
wbw	10.99	7.28
wbw+bigram	11.28	7.08
wbw+orth	11.70	8.24
wbw+orth+bigram	11.98	7.93

Table 1: Baseline results (BLEU) with word-by-word translation on newstest2018. We indicate the use of bigrams and orthographic similarity with *bigram* and *orth* respectively.

5.1 BWE Baseline Experiments

We present our word-by-word translation baseline results in Table 1. Using bigrams on the English side helped for de-en but not for en-de. By analyzing translations we can conclude that 1) German compound words are correctly translated to multiple words in many cases and 2) the drop of en-de direction is caused by incorrectly translating bigrams, that are non-compounds on the target side, to one token units. On the other hand, using orthographic information gave significant improvements in both directions. The technique alone provided for improved translation of named entities without the use of a costly NER system. We got our best results by combining bigrams and orthographic similarity for German→English.

Comparing the results with the unsupervised and lightly supervised mapping it can be seen that the two systems are on par in performance, the former results higher BLEU points in case of de-en but lower for en-de. Our conjecture is that the multiple translations of the source words in the used lexicon helped tackle the morphological richness of the German language on the target side while it was not helpful otherwise.

5.2 Unsupervised PBT Results

The top half of Table 2 reports the translation quality that we achieved with the phrase-based un-

supervised approach (cf. Section 3), measured in case-sensitive BLEU. Our test set for these experiments is newstest2017 (whereas the BLEU scores in Table 1 are on newstest2018). The experiment in the first line of Table 2 is conceptually equivalent to the unsupervised "wbw" experiment from Table 1. We use the Moses decoder to perform monotonic word-by-word translation without a language model (LM) or any other feature functions except for the single translation model (TM) score that we obtain from the cosine similarities. If we add a 4-gram LM and heuristically weight the LM feature function with a scaling factor of 0.1 and the TM with 0.9 (second line in Table 2), the translation quality improves by more than 2.5 BLEU points in both of the two translation directions. By using a small parallel development set (newstest2016) to tune the two weights with MIRA (Cherry and Foster, 2012) (third line), we barely improve over our guessed scaling factors of 0.1 for the LM and 0.9 for the TM. Optimized scaling factors are however more relevant when we allow for reordering (fourth line), since we then activate a third feature function, namely a distance-based distortion cost. This adds another scaling factor, and a good informed guess of reasonable values for three weights becomes increasingly difficult. Activated reordering with tuned weights boosts our translation quality further.

We can go beyond simple word-by-word translation if we add our BWE bigrams to the TM, thus also enabling 1:2, 2:1, and 2:2 translation by means of new phrase table entries. Reordering and the 4-gram LM are kept active in the new configuration. But to give the system control over the lengths of the hypothesis translations (which now can differ from the input sentence lengths), we also activate the word penalty and phrase penalty feature functions, and we include three more binary indicator features for table entries that are 1:2, 2:1, and 2:2, respectively. The scaling factors are optimized on newstest2016 again. With bigrams, we observe higher translation quality in the German→English translation direction, but not in the English→German direction (fifth line in Table 2). This is consistent with what we noted above (cf. Table 1).

Finally, we created 50 M synthetic sentence pairs from German monolingual data with our best German→English phrase-based unsupervised system. With a phrase table extracted from the synthetic data, we achieve our best phrase-based unsupervised translation result in the English→German translation direction (sixth line).[1]

	unsup. PBT	
	de-en	en-de
wbw (Moses decoder)	7.92	4.49
+ 4-gram LM (weighted 0.1)	10.52	7.21
+ tuned weights	10.73	7.20
+ reordering	11.47	7.66
+ bigram	12.44	7.61
synthetic data	–	10.66
	unsup. NMT	
	de-en	en-de
baseline	13.77	10.45
fine-tune w/o denoising	15.03	12.08
w/ orth	16.06	12.38
w/ orth + bigram	16.98	13.13
NewsCrawl 2017	16.42	12.46

Table 2: BLEU scores with the unsupervised systems on newstest2017.

5.3 Unsupervised NMT Results

We show the results from our unsupervised neural systems (cf. Section 4) in the bottom half of Table 2. The translation quality still lags behind supervised translation systems. Only one other team (RWTH Aachen University) competed in the WMT18 unsupervised learning sub-track, and the performance of their unsupervised systems is roughly comparable to our submissions.

Our final submission system was trained on a subsample of NewsCrawl from 2007 to 2017. We did not include any of the orthographic similarity or bigram word-by-word translated data. The model selection was done based on the newstest2017 test set and we use the same model checkpoint for both translation directions. For the final submission model, we removed the word-by-word translated data after 6K iterations and subsequently trained the model for a total of 300K iterations. This model was able to obtain 13.77 on the de-en and 10.45 on en-de translation task. Subsequently, we disabled denoising and continued the training just with on-the-fly backtranslation which managed to provide for further gains of 1.26 for de-en and 1.63 for en-de. In subse-

[1] In consideration of the computational cost, we decided to try synthetic data in only one of the two translation directions.

quent experiments we observed that removing the word-by-word translated data does not change the performance and for the contrastive experiments, for simplicity, we remove it at the same time as disabling denoising.

Our contrastive experiments show that the choice of data can have some effect on the translation performance. Training a model on a subsample of NewsCrawl 2017, showed to be more beneficial. Using more recent data can provide for better correlation between the training and test sets. However, it is difficult to pinpoint whether this is because of better general content overlap or because of the recency of the data.

In the word-by-word translations, the use of orthographic similarity proved to be very helpful. Some of those effects are transfered when we use that data in the neural system. For de-en it provided for an improvement of 1.03 BLEU, while for en-de only 0.30 BLEU.

Adding bigrams did not provide for consistent improvements in the word-by-word translations. However, the neural system managed to make use of these translations better, most likely from the additional reordering that is contained in this data. Furthermore, compound words in German are handled better in this way, since we have a more direct mapping between them and English words. We only present results with translations obtained with the combination of orthographic similarity and bigrams. Adding bigrams, improved upon the orthographic similarity translations by 0.92 for de-en and 0.75 for en-de. Using this technique, we obtain the highest performance on both translations directions.

We also extracted pseudo parallel sentences by mining NewsCrawl 2015. The similarity of a sentence pair is computed by calculating the average similarity between all source-target pairwise word similarities. The similarity between a source and target word is computed based on the BWEs and the orthographic similarity. We extracted ≈8K sentences. We oversampled the dataset to the size of the monolingual data and used it at the beginning of the training. We also attempted to use the original 8K sentences as a last fine-tuning step. Both approaches did not provide for improvements over our best scoring system.

6 Analysis

In Table 3 we present examples and we compare German→English translations with the different contrastive setups we outline in the experimental results. We show the phenomena that we observed and discuss some of the challenges that the systems are still not able to overcome. This can be a useful analysis that can provide insight into where future work should focus on.

In the first example we see that the models are to some extent able to do simple reorderings and insertions. We can see that most models were able to properly reorder "wollte die 45-Jährige" to "the 45-year-old wanted". The *Orth. + bigram* and *NewsCrawl 2017* were able to move "beruhigen" (calm) in front of "their brother" and furthermore inserted the preposition "to".

In the second example, we can observe that the models were again able to infer that the phrase "tot aufgefunden" should be reordered to "found dead". Additionally, the whole phrase was inserted at a much more appropriate place in the English sentence rather than at the end. Another interesting phenomenon is that the *NewsCrawl 2017* model was able to do a 2-1 mapping by translating "Einkaufzentrums" to "shopping centre". On the other hand, this example shows the challenges our models encounter. Given the relatively unintuitive mapping between "Koch" and "Hopkinson" that we have from the BWEs, the models had difficulty properly translating this word. Furthermore, most of them were not able to infer that "nach" in combination with "gezogen" translates to "moved to" and we see some more literal translations.

The third example shows some of the issues we had with translating named entities. Models without the orthographic similarity extension had trouble finding a suitable translation of "Erdogans". Furthermore most of the models inferred that adding the preposition "of" is necessary in this case.

The last example shows the importance of the dataset being used. The first three systems are trained on the same data and didn't translate "Kalendar" as opposed to the one trained on a subsample of NewsCrawl 2017. Although not necessarily related to the dataset being more recent, it shows that it most likely contained sentences that enabled proper translation to "calendar".

source	Gemeinsam mit ihrem Lebensgefährten wollte die 45-Jährige ihren Bruder *beruhigen*.
reference	The 45-year-old and her partner wanted *to calm* down her brother.
Final submission	Met with her boyfriend, the 45-year-old wanted their brother calming.
Orthographic	Watching her boyfriend, the 45-year-old didn't have handled their brother.
Orth. + bigram	Together with her boyfriend, the 45-year-old wanted *to calm* their brother.
NewsCrawl 2017	Together with her boyfriend, the 45-year-old wanted *to calming* their brother.
source	Ein 28-jähriger Koch, der vor kurzem nach San Francisco gezogen ist, wurde im Treppenhaus eines örtlichen *Einkaufzentrums* tot aufgefunden.
reference	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
Final submission	A 28-yard Koch, who was pulled before he was pulled after San Francisco, was found in the stairwell of a local outlet dead Province.
Orthographic	A 28-year-old Reid, who has ever been relocated after San Francisco, was found dead in the hallway of a local crop.
Orth. + bigram	A 28-yard Koch, who recently moved after San Francisco, was found dead in the hallway of its local outlet.
NewsCrawl 2017	A 28-year-old Koch, who was given her home to San Francisco, was found dead in the stairwell of a local *shopping centre*.
source	Der Sport ist - wie das ganze Land - gespalten in Anhänger und Gegner Erdogans.
reference	The sport - like the entire country - is divided into those who support Erdogan, and those who do not.
Final submission	The sport is - like the whole country - divided in supporters and opponents Drogba.
Orthographic	The BBC is - like the whole country - divided in supporters and opponents of Erdogan.
Orth. + bigram	The sports is - like the whole country - divided in supporters and opponents of Erdogan.
NewsCrawl 2017	The sport is - like the whole country - divided in supporters and opponents of Mrs. May.
source	Das Treats Magazin arbeitet mit dem Fotografen David Bellemere zusammen, um einen 1970er Jahre Pirelli-inspirierten Kalendar für 2017 herauszubringen.
reference	Treats magazine is partnering with photographer David Bellemere to launch a 1970s' Pirelli-inspired calendar for 2017.
Final submission	The Treats magazine works with the photographers David Bellemere together, when a 1970s Pirelli-inspiring Kalendar for 2017 dates.
Orthographic	The Treats magazine works with the photographers David Bellemere together to bring a 1970s Pirelli-inspiring Kalendar for 2017.
Orth. + bigram	The Treats magazine works with the photographers David Bellemere together to bring a 1970s Pirelli-inspected Kalendar for 2017.
NewsCrawl 2017	The Treats magazine works with the brains David Bellemere together to attribute a 1970s Pirelli inspires calendar for 2017.

Table 3: Example translations obtained using the different neural systems.

7 Conclusion

Corpus-based machine translation approaches typically require parallel training data. In this work, we have investigated methods which allow for unsupervised learning of translation models, i.e., we have examined how machine translation systems can be trained without any parallel data.

LMU Munich is one of two teams who participated in the WMT18 unsupervised learning subtrack for machine translation of news articles between German and English. Our shared task submission consists of an unsupervised phrase-based translation system and an unsupervised neural machine translation system.

We have shown how bigrams and orthographic similarity in the underlying bilingual word embeddings benefit the results. We have presented effective unsupervised learning techniques for both the phrase-based and the neural paradigm and have demonstrated how an effective training curriculum improves translation quality.

Acknowledgments

We thank Helmut Schmid for helpful discussions and comments. 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, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Ma-

chine 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.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

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

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46.

Fabienne Braune, Viktor Hangya, Tobias Eder, and Alexander Fraser. 2018. Evaluating Bilingual Word Embeddings on the Long Tail. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 188–193.

Fabienne Cap, Marion Weller, Anita Ramm, and Alexander Fraser. 2014. CimS – The CIS and IMS Joint Submission to WMT 2014 Translating from English into German . In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 71–78.

Fabienne Cap, Marion Weller, Anita Ramm, and Alexander Fraser. 2015. CimS - The CIS and IMS Joint Submission to WMT 2015 Addressing Morphological and Syntactic Differences in English to German SMT . In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 84–91.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada. Association for Computational Linguistics.

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

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint arXiv:1803.05567*.

Matthias Huck, Fabienne Braune, and Alexander Fraser. 2017. LMU Munich's Neural Machine Translation Systems for News Articles and Health Information Texts. In *Proceedings of the Second Conference on Machine Translation*, pages 315–322.

Matthias Huck, Alexander Fraser, and Barry Haddow. 2016. The Edinburgh/LMU Hierarchical Machine Translation System for WMT 2016. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 311–318.

Matthias Huck, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2018. LMU Munich's Neural Machine Translation Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Volume 2, Shared Task Papers*.

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, 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA. Association for Computational Linguistics.

Jan-Thorsten Peter, Tamer Alkhouli, Hermann Ney, Matthias Huck, Fabienne Braune, Alexander Fraser, Aleš Tamchyna, Ondřej Bojar, Barry Haddow, Rico Sennrich, Frédéric Blain, Lucia Specia, Jan Niehues, Alex Waibel, Alexandre Allauzen, Lauriane Aufrant, Franck Burlot, elena knyazeva, Thomas Lavergne, François Yvon, Mārcis Pinnis, and Stella Frank. 2016. The QT21/HimL Combined Machine Translation System. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 344–355.

Hassan Sajjad, Svetlana Smekalova, Nadir Durrani, Alexander Fraser, and Helmut Schmid. 2013. QCRI-MES Submission at WMT13: Using Transliteration Mining to Improve Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 219–224.

Aleš Tamchyna, Roman Sudarikov, Ondřej Bojar, and Alexander Fraser. 2016. CUNI-LMU Submissions in WMT2016: Chimera Constrained and Beaten. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 385–390.

Marion Weller, Max Kisselew, Svetlana Smekalova, Alexander Fraser, Helmut Schmid, Nadir Durrani, Hassan Sajjad, and Richárd Farkas. 2013. Munich-Edinburgh-Stuttgart Submissions at WMT13: Morphological and Syntactic Processing for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 232–239.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation . In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.