Better OOV Translation with Bilingual Terminology Mining

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

Abstract

Unseen words, also called out-of-vocabulary words (OOVs), are difficult for machine translation. In neural machine translation, byte-pair encoding can be used to represent OOVs, but they are still often incorrectly translated. We improve the translation of OOVs in NMT using easy-to-obtain monolingual data. We look for OOVs in the text to be translated and translate them using simple-to-construct bilingual word embeddings (BWEs). In our MT experiments we take the 5−best candidates, which is motivated by intrinsic mining experiments. Using all five of the proposed target language words as queries we mine target-language sentences. We then back-translate, forcing the back-translation of each of the five proposed target-language OOV-translation-candidates to be the original source-language OOV. We show that by using this synthetic data to fine-tune our system the translation of OOVs can be dramatically improved. In our experiments we use a system trained on Europarl and mine sentences containing medical terms from monolingual data.

1 Introduction

Neural machine translation (NMT) systems achieved a breakthrough in translation quality recently, by learning an end-to-end system (Sutskever et al., 2014; Bahdanau et al., 2015). However, NMT systems have low quality when translating out-of-vocabulary words (OOVs), especially because they have a fixed modest sized vocabulary due to memory limitations. By splitting words into subword units the problem of representing OOVs can be solved (Sennrich et al., 2016b) but their translation is still problematic because by definition source-side OOVs were not seen in the training parallel data together with their translations. In this work, we evaluate a simple approach for improving the translation of OOVs using bilingual word embeddings (BWEs), which we hope will trigger more research on this interesting problem.

In previous approaches, to include words in the target sentence for which the translation is unknown the token `unk` is often used which can be handled by later steps. In many cases, such as named entities, it is possible to just copy the source token to the target side instead of translating it. Gulcehre et al. (2016) proposed a pointer network based (Vinyals et al., 2015) system which can learn when to translate and when to copy. On the other hand, it is not possible to always copy when the translation is unknown. If the alignment of the `unk` tokens to the source are known it is possible to translate source words using a large dictionary as a post-processing step. Although NMT systems do not rely on word alignments explicitly, it is possible to learn and output word alignments (Luong et al., 2015). It is also possible to use lexically-constrained decoders (Post and Vilar, 2018; Hasler et al., 2018) in order to force the network to output certain words or sequences. This way alignments are not needed and the system can decide the position of the constraints in the output. The disadvantage of the above methods is that the translation of words needed to be decided either as a pre- or post-processing step without the context which makes the translation of some words, such as polysemous words, difficult. In addition, lexically-constrained decoders require the target words to be observed in context at training time, or they will usually not be placed properly. In contrast, we fine-tune NMT systems for better translation of problematic words on the sentence level and are thus able to exploit the context instead of handling the problem on the word level.

In our approach, we rely on bilingual word embeddings (BWEs) which can be built using large

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5809–5815
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

monolingual data and a cheap bilingual signal. BWEs can easily cover a very large vocabulary. Given the sentences to translate we look for source language words not included in the parallel training set of our MT system (OOVs). We translate OOVs using BWE based dictionaries taking n-best candidates as opposed to previous work (e.g., (Luong et al., 2015)) where only the best translation is used during post-processing. In our experiments we take the $5-$best predictions of our BWEs, and retrieve sentences containing these target-language predictions from a monolingual corpus. As was shown before, NMT systems can be quickly and effectively fine-tuned using just a few sentences (Farajian et al., 2017, 2018; Wuebker et al., 2018). Based on the $5-$best translations of OOVs we mine sentences from target language monolingual data and generate a synthetic parallel corpus using back-translation (Sennrich et al., 2016a). We force the source-language translation of each OOV-translation-candidate to be the original OOV. We show that by using this synthetic data to fine-tune our system the translation of unseen words can be dramatically improved, despite the presence of wrong translations of each OOV in the synthetic data. We test our system on the translation of English medical terms to German and show significant improvements using our approach. In this paper, we study a domain adaptation task in order to show the advantages clearly, but our approach does not focus on this domain adaptation and it can also be directly applied generally with no modification (e.g., to an in-domain task).

2 Approach

In order to fine-tune an NMT system we aim to generate a synthetic parallel corpus containing the translations of source OOVs on the target side. Our approach relies on a dictionary containing source-target word translations. We mine target language sentences using the $n-$best translations of OOVs from topic specific monolingual data. We back-translate these sentences and run a (fine-tuning) training step of the NMT system on the generated corpus. Even though many word translation candidates in the dictionary are incorrect, we show in our experiments that the NMT system can effectively filter out the noise in the synthetic corpus using the context.

2.1 Word Translation

To translate source language words we use a combination of BWE based cosine and orthographic similarity. BWEs represent source and target language words in a joint space and can be built by training monolingual spaces and projecting them to the common space. Initially, a small seed lexicon was used as the bilingual signal to learn a linear mapping (Mikolov et al., 2013) which was further improved by applying orthogonal transformations only (Xing et al., 2015). Recently, various techniques were developed to build BWEs without any bilingual signal (Conneau et al., 2018; Artetxe et al., 2018). In the work of Conneau et al. (2018) adversarial training is employed to generate an initial seed lexicon of frequent words which is then used for orthogonal mapping. Even though BWEs in general are of good quality the translation of various words types, such as named entities and rare words, could be further improved by using orthographic similarity (Braune et al., 2018; Riley and Gildea, 2018; Artetxe et al., 2019). Similarly to (Braune et al., 2018), we combine the BWE based cosine and orthographic similarity of word pairs to get the translations of source words. We generate a dictionary of source-target word pairs by taking the top n most similar target words for each source using both similarity measures. We define orthographic similarity as one minus normalized Levenshtein distance. Since orthographic similarity of close words are higher than their cosine, we weight the former with 0.2 (we found this value to work well on a different task and did not tune it further).

To build monolingual embeddings we use *fastText*'s skipgram model (Bojanowski et al., 2017) with dimension size 300 and minimum word frequency 3. For building unsupervised BWEs we use *MUSE* as the implementation of (Conneau et al., 2018). Note that we use unsupervised BWEs due to their good performance on the En-De language pair (see (Conneau et al., 2018)). But acquiring a small lexicon including frequent words is cheap for language pairs where unsupervised mapping has a lower performance than supervised mapping, and could be considered in future work.

2.2 NMT Fine-Tuning

We mine target language sentences from a monolingual corpus which contains the translations of source OOVs. Since the source sentences needed

	UFAL			UFAL+orth			EU+UFAL			EU+UFAL+orth		
n	$P@n$	$R@n$	$F_1@n$	$P@n$	$R@n$	$F_1@n$	$P@n$	$R@n$	$F_1@n$	$P@n$	$R@n$	$F_1@n$
1	58.19	13.58	22.02	58.13	25.28	35.24	68.65	37.56	48.55	69.59	41.87	52.28
5	44.46	26.10	32.89	50.05	43.82	46.73	54.33	48.46	51.22	51.13	51.71	51.41
10	35.80	29.84	32.55	41.04	47.64	44.09	42.94	53.41	47.61	44.45	56.34	49.70
20	29.54	33.58	31.43	34.43	50.16	40.83	36.42	58.78	44.98	37.42	61.30	46.47

Table 1: Quality of the mining procedure using different sizes of $n-$best translations. We use only sentences from UFAL or both EU and UFAL to build BWEs. We compare cosine only and cosine combined with orthography.

to be translated are available before running the decoder, it is possible to get a list of OOVs from them by using the word vocabulary of the parallel training data. We translate the OOVs of our development and test data using the dictionaries described above by taking $n-$best translations. We present experiments with different n values in our intrinsic experiments. These source words tend to be noisy, especially in the medical domain, thus we apply a filtering step by ignoring those words containing non-letter characters as more than one third of their characters. In addition, we also filter out translations that are stopwords. We then use the set of target language words to mine all sentences that contain any of them from the monolingual data. We filter out sentences longer than 50 tokens, since they tend to be listings of medical terms, and back-translate the rest to generate synthetic parallel data. We force the back-translation of each of the proposed target-language OOV-translation-candidates to be the original source-language OOV.

In our experiments we use an encoder-decoder NMT system (Sennrich et al., 2017) with attention, 500 dimensional embedding layer, 1024 dimensional GRU layer and we use Adam with a learning rate of 0.0001 to train the network. We apply word segmentation with BPE using 50K merge operations to the English text, and a linguistically informed pipeline to the target-side German text (Huck et al., 2017b). It is important to understand that OOVs for us are words, and we handle both the dictionary based OOV translation and sentence mining on the word level. BPEs are only used when using NMT to translate. We train two systems, one each for the forward and backward directions. We describe the used data in Section 3. During back-translation we force the OOV-translation-candidates to be back-translated to the original source-language OOV by changing the OOV-translation-candidate to a special token on the target side before translation and then substituting the special token in the source-language back-translated output with the original OOV. This way, we make sure the MT system sees the OOV and each of its OOV-translation-candidates in the correct target-language context for the particular OOV-translation-candidate being considered.

Finally, to improve the OOV translation of the forward system, we fine-tune it on the generated parallel data. We run only one training step over the whole synthetic corpus similarly to (Farajian et al., 2018), which makes the system learn newly seen words while not overwriting important knowledge previously learned from the truly parallel data the system was originally trained on. Since we mine target sentences based on multiple OOV-translation-candidates for each given OOV the system is tuned on different translations and their relevant contexts. This helps the network to correctly translate polysemous words, because the input context (which often disambiguates a polysemous word) will usually be most similar to the target-language context of the correct OOV-translation-candidate. Furthermore, this also makes our approach robust against incorrect OOV-translation-candidates in the used dictionary, since they are often used in very different contexts compared to the context of the source OOV we are translating.

3 Experiments

We translate medical English sentences to German. To train the baseline NMT system we used the *Europarl v7* (EU) parallel dataset containing $1.9M$ sentence pairs (Koehn, 2005). As medical data, we took $3.1M$ sentences from titles of medical Wikipedia articles, medical term-pairs, patents and documents from the European Medicines Agency which are part of the *UFAL Medical Corpus* (UFAL). Since the corpus is parallel, we split it and used even sentences for English and odd ones for German. We built BWEs not only on the monolingual medical data but on

		Acc_1	Acc_5
freq	(Braune et al., 2018)	38.6	47.4
	EU+UFAL+orth	25.9	40.6
rare	(Braune et al., 2018)	26.3	28.2
	EU+UFAL+orth	17.5	28.8

Table 2: Medical bilingual lexicon induction results showing the quality of the BWE based dictionaries using 1-best and 5-best translations.

the concatenation of all Europarl data and the monolingual medical data to improve the quality of BWEs (Hangya et al., 2018). We only mined sentences from the monolingual medical German corpus. The testing of our approach was done on the medical *Health In My Language* (HimL) corpora (Haddow et al., 2017) containing $1.9K$ sentence pairs in both development and test sets. All corpora were tokenized and truecased using *Moses* scripts (Koehn et al., 2007).

We ran two sets of experiments. First we show the translation quality of our dictionaries by looking at the OOVs and their translations using HimL development data. Then we show translation quality improvements on the HimL test data.

3.1 OOV Translation

The quality of our proposed method is highly dependent on that of the used dictionaries, since in order to mine useful sentences OOVs first needed to be translated correctly. Since we lack the gold translations of the OOVs, we measure the quality of the mined target language sentences using parallel data by following the approach presented for the fine-tuning of the NMT system. We translate source OOVs from the HimL development data using the $n-$best translations resulting a set of target language words. We mine sentences from the target side containing any of these translations. For each mined sentence we check if its source side pair contains the corresponding OOV, meaning that the correct translation of the OOV was contained by translation candidates, or not which means that the sentence was mined due to the translation of a different OOV. In addition, we also measure the number of missed sentences, i.e., in case a source sentence contains an OOV but its target reference was not mined due to no correct translation of the OOV in the candidates. We show precision, recall and F_1 scores indicating how precisely would our system mine sentences from the target side for the OOVs and the ratio of OOVs covered. We use dictionaries with different number of $n-$best translations built using only the medical sentences of UFAL or both Europarl and medical sentences in case of EU+UFAL. We also compare dictionaries using only cosine similarity with combined cosine and orthography (+orth).

	Cochrane	NHS24
baseline	22.4	20.2
copy	23.4	20.5
fine-tuned	27.2	22.5

Table 3: BLEU scores on the HimL test sets comparing the baseline systems and our OOV specific fine-tuning.

We present results in Table 1. By comparing dictionaries it can be seen that by using the additional EU data to build embeddings the translation performance could be improved. As it was shown in (Hangya et al., 2018) as well, the use of additional general knowledge monolingual embeddings have higher quality. In addition, although the parallelism in the EU data is not exploited explicitly, it effects mapping due to higher monolingual space isomorphism (Søgaard et al., 2018). Using orthographic similarity in addition to cosine further improves quality since a lot of medical terms have similar surface forms across languages.

The precision using the most similar translation of OOVs indicates good dictionary quality for all setups. On the other hand, it misses a lot of OOVs. By increasing translation candidates recall could be improved to the detriment of precision. Looking at F_1 scores we found that $5-$best translations gives best results 3 out of 4 times, thus we chose this value for the MT experiments.

We also compare the quality of our best dictionary (EU+UFAL+orth) to previous work by running bilingual lexicon induction using the test lexicons of Braune et al. (2018) containing frequent and rare medical words respectively. Accuracies of 1-best and 5-best translations in Table 2 show comparable word translation quality to previous work, although we do not employ any task specific steps in contrast to Braune et al. (2018). Note that our dictionary does not contain some of the rare words of the test lexicons which we ignore during evaluation.

3.2 Machine Translation

We present the improvements of our approach in terms of translation quality in the following. As the baseline, we used the English to German NMT

source	regular **nosebleeds**
reference	regelmäßige **Nasenbluten**
baseline	Regelmäßige **Misskredite** (discredits)
fine-tuned	Regelmäßige **Nasenbluten**
source	dizziness or **lightheadedness**
reference	Schwindel oder **Benommenheit**
baseline	Schwindelerregend (dizzying) oder **zurückhaltend** (reluctant).
fine-tuned	Schwindel oder **Schwächegefühl** (feeling of faintness)
source	A coronary **angioplasty** may not be technically possible [. . .]
reference	Eine **Koronarangioplastie** ist wahrscheinlich technisch nicht möglich [. . .]
baseline	Ein **Herzinfarkt** (heart attack) ist vielleicht technisch nicht möglich [. . .]
fine-tuned	Eine koronare **Angioplastie** ist möglicherweise nicht technisch möglich [. . .]
source	Four different alpha blockers were tested (**alfuzosin**, **tamsulosin**, **doxazosin** and **silodosin**).
reference	Vier verschiedene Alphablocker wurden getestet (**Alfuzosin**, **Tamsulosin**, **Doxazosin** und **Silodosin**).
baseline	Vier verschiedene Alphablocker wurden getestet (**alfuzos**, **tasuloin**, **doxasa** und **silodosin**).
fine-tuned	Vier unterschiedliche Alphablocker wurden untersucht (**Alfuzosin**, **Tamsulosin**, **Doxazosin** und **Tigecyclin**).

Table 4: Example translations comparing the baseline with our fine-tuned model. OOVs and their translations are highlighted in bold. For convenience, we provide the English meaning of a selected set of German translations (small font in parentheses).

system detailed earlier without fine-tuning, i.e., trained only on Europarl data. We also compare our system to an approach which simply copies source OOVs to the target side. Similarly to our back-translation approach, we change OOVs to a special token on the source side before translation which we substitute with the original OOV on the target side. If multiple OOVs appear in a sentence we use the order as they appear on the source side. Based on the experiments in the previous section, we used the EU+UFAL+orth dictionary with 5−best translations resulting in 95K mined target sentences from the monolingual corpus. We present case-sensitive BLEU scores calculated with the `mteval-v13a.pl` script from the Moses toolkit on the two parts of HimL test set separately: Cochrane and NHS24.

Results are in Table 3. The performance of the baseline system is poor on both parts of the test set due to the many OOVs in the source sentences which were not seen in the parallel Europarl. The system is also out of domain which causes an additional detriment. (Cf. Huck et al. (2017a, 2018) for descriptions of state-of-the-art health domain translation systems that are trained on large in-domain parallel data.) A simple source-to-target OOV token copying strategy improves over the baseline, but not by a large margin. The fine-tuned system, by contrast, performs considerably better, achieving an increase of $+4.8$ and $+2.3$ BLEU points on Cochrane and NHS24, respectively. By looking at examples (Table 4) we see that, on top of the domain adaptation effect of the back-translated data, the translation of OOVs is improved, especially of medical terminology, showing the effectiveness of the approach.

4 Conclusions

Although OOVs can be represented in NMT systems, translation is difficult. In this paper we proposed a method for better translation of OOVs. Our approach relies on bilingual word embeddings based dictionaries which are simple to construct but cover a large vocabulary. We mine target-language sentences containing the 5−best translations of OOVs according to our BWEs. We then back-translate. Using this noisy synthetic parallel data we fine-tune the initial NMT system. We showed the performance of our approach on the translation of medical terms using a system trained on Europarl parallel data. Our results showed that having both source OOVs and their translations in the sentence pairs results in improvements in BLEU. Our method of term mining followed by back-translation and fine-tuning can easily be applied to any NMT task including non-domain-adaptation tasks.

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. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proc. ACL*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An Effective Approach to Unsupervised Machine Translation. *CoRR*, abs/1902.01313.

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation By Jointly Learning To Align and Translate. In *Proc. ICLR*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Fabienne Braune, Viktor Hangya, Tobias Eder, and Alexander Fraser. 2018. Evaluating bilingual word embeddings on the long tail. In *Proc. NAACL-HLT*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *Proc. ICLR*.

M. Amin Farajian, Nicola Bertoldi, Matteo Negri, Marco Turchi, Marcello Federico, and Fondazione Bruno Kessler. 2018. Evaluation of Terminology Translation in Instance-Based Neural MT Adaptation. In *Proc. EAMT*.

M. Amin Farajian, Marco Turchi, Matteo Negri, Marcello Federico, and Fondazione Bruno Kessler. 2017. Multi-Domain Neural Machine Translation through Unsupervised Adaptation. In *Proc. WMT*.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the Unknown Words. In *Proc. ACL*.

Barry Haddow, Alexandra Birch, Ondrej Bojar, Fabienne Braune, Colin Davenport, Alexander Fraser, Matthias Huck, Michal Kaspar, Kvetoslava Kovaríková, Josef Plch, Anita Ramm, Juliane Ried, James Sheary, Ales Tamchyna, Dusan Varis, Marion Weller, and Phil Williams. 2017. HimL: Health in my Language. In *Proc. EAMT*.

Viktor Hangya, Fabienne Braune, Alexander Fraser, and Hinrich Schütze. 2018. Two Methods for Domain Adaptation of Bilingual Tasks : Delightfully Simple and Broadly Applicable. In *Proc. ACL*.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural Machine Translation Decoding with Terminology Constraints. In *Proc. NAACL-HLT*.

Matthias Huck, Fabienne Braune, and Alexander Fraser. 2017a. LMU Munich's Neural Machine Translation Systems for News Articles and Health Information Texts. In *Proc. WMT*.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017b. Target-side Word Segmentation Strategies for Neural Machine Translation. In *Proc. WMT*.

Matthias Huck, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2018. LMU Munich's Neural Machine Translation Systems at WMT 2018. In *Proc. WMT*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. MT Summit*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcelo Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL: Interactive Poster and Demonstration Sessions*.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the Rare Word Problem in Neural Machine Translation. In *Proc. ACL*.

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

Matt Post and David Vilar. 2018. Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. In *Proc. NAACL-HLT*.

Parker Riley and Daniel Gildea. 2018. Orthographic Features for Bilingual Lexicon Induction. In *Proc. ACL*.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proc. EACL, Software Demonstrations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proc. ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proc. ACL*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proc. ACL*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proc. NIPS*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Proc. NIPS*.

Joern Wuebker, Patrick Simianer, and John DeNero. 2018. Compact Personalized Models for Neural Machine Translation. In *Proc. EMNLP*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proc. NAACL-HLT*.