Self-Supervised Neural Machine Translation

Dana Ruiter
Saarland University

Cristina España-Bonet
Saarland University
DFKI GmbH

Josef van Genabith
Saarland University
DFKI GmbH

druiter@lsv.uni-saarland.de
{cristinae,Josef.Van_Genabith}@dfki.de

Abstract

We present a simple new method where an emergent NMT system is used for simultaneously selecting training data and learning internal NMT representations. This is done in a self-supervised way without parallel data, in such a way that both tasks enhance each other during training. The method is language independent, introduces no additional hyper-parameters, and achieves BLEU scores of 29.21 ($en2fr$) and 27.36 ($fr2en$) on *newstest2014* using English and French Wikipedia data for training.

1 Introduction

Neural machine translation (NMT) has brought major improvements in translation quality (Cho et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017). Until recently, these relied on the availability of high-quality parallel corpora. As such corpora exist only for a few high-resource language combinations, overcoming this constraint by either extracting parallel data from non-parallel sources or developing unsupervised techniques in NMT is crucial to cover all languages.

Obtaining comparable corpora is becoming easier (Paramita et al., 2019) and **extracting parallel sentences** from them a wide research field. Most of the methods estimate similarities between fragments to select pairs. Here we focus on similarities estimated from NMT representations. The strength of **NMT embeddings as semantic representations** was first shown qualitatively in Sutskever et al. (2014); Ha et al. (2016) and Johnson et al. (2017), and used for estimating semantic similarities at sentence level in España-Bonet and Barrón-Cedeño (2017) for example. In a systematic study, España-Bonet et al. (2017) show that cosine similarities between context vectors discriminate between parallel and non-parallel sentences already in the first stages of training. Other approaches perform max-pooling over encoder outputs (Schwenk, 2018; Artetxe and Schwenk, 2018) or calculate the mean of word embeddings (Bouamor and Sajjad, 2018) to extract pairs.

On the other hand, **unsupervised NMT** is now achieving impressive results using large amounts of monolingual data and small parallel lexicons (Lample et al., 2018a; Artetxe et al., 2018b; Yang et al., 2018). These systems rely on very strong language models and back-translation, and build complex architectures that combine denoising autoencoders, back-translation steps and shared encoders among languages. The most successful architectures also use SMT phrase tables, standalone or in combination with NMT (Lample et al., 2018b; Artetxe et al., 2018a).

In **our approach**, we propose a new and simpler method without *a priori* parallel corpora. Our premise is that NMT systems —either sequence to sequence models with RNNs, transformers, or any architecture based on encoder–decoder models— already learn strong enough representations of words and sentences to judge on-line if an input sentence pair is useful or not. Our approach resembles **self-supervised learning** (Raina et al., 2007; Bengio et al., 2013), i.e. learning a primary task where labelled data is not directly available but where the data itself provides a supervision signal for another auxiliary task which lets the network learn the primary one. In our case this comes with a twist: we find cross-lingually close sentences as an auxiliary task for learning MT and learning MT as an auxiliary task for finding cross-lingually close sentences in a mutually self-supervised loop: in effect a doubly virtuous circle.

Our approach is also related to unsupervised NMT but differs in important aspects: since in our case there is no back-translation involved, the original corpus must contain similar sentences,

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1828–1834
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

therefore the use of comparable corpora is recommended to speed up the training.

In the following, we describe the approach (Section 2) and the experiments in which it is going to be tested (Section 3). Section 4 reviews the results and, finally, we summarise and sketch future work in Section 5.

2 Joint Model Architecture

Without loss of generality, we consider a bidirectional NMT system {L1, L2}→{L1, L2} where the encoder and decoder have the information of both languages L1 and L2. The bidirectionality is simply achieved by tagging the source sentence with the target language as done by Johnson et al. (2017) in their multilingual systems and inputting sentence pairs in both directions. Two dimensions determine our architectures: (*i*) the specific representation of an input sentence, and (*ii*) the similarity or score function for an input sentence pair.

We focus on two different embedding spaces in the encoder to build **semantic sentence representations**: the sum of word embeddings (C_e) and the hidden states of an RNN or the encoder outputs of a transformer (C_h). We define:

$$C_e = \sum_{t=1}^{T} e_t, \qquad C_h = \sum_{t=1}^{T} h_t, \qquad (1)$$

where e_t is the word embedding at time step t and h_t its hidden state (RNN) or encoder output (transformer). In case h_t is an RNN hidden state, it is further defined by the concatenation of its forward and backward component $h_t^{\mathrm{RNN}} = [\overrightarrow{h}_t; \overleftarrow{h}_t]$.

These representations are used to **score input sentence pairs**. We study two functions for sentence selection with the aim of exploring whether a threshold-free selection method is viable.

Let S_{L1} and S_{L2} be the vector representations for each sentence of a pair (either C_e or C_h). The **cosine similarity** of a sentence pair is calculated as the dot product of their representations:

$$\mathrm{sim}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{S_{\mathrm{L1}} \cdot S_{\mathrm{L2}}}{\|S_{\mathrm{L1}}\| \, \|S_{\mathrm{L2}}\|}, \qquad (2)$$

which is bounded in the [-1, 1] range. However, the threshold to decide when to accept a pair is not straightforward and might depend on the language pair and the corpus (España-Bonet et al., 2017; Artetxe and Schwenk, 2018). Besides, even if the measure does not depend on the length of the sentences, it might be scaled differently for different sentences. To solve this, Artetxe and Schwenk (2018) proposed a **margin-based** function:

$$\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{\mathrm{sim}(S_{\mathrm{L1}}, S_{\mathrm{L2}})}{\mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 + \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2}, \qquad (3)$$

where $\mathrm{avr}_{\mathrm{kNN}}(X, Y_k)$ corresponds to the average similarity between a sentence X and $k\mathrm{NN}(X)$, its k nearest neighbors Y_k in the other language:

$$\mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\mathrm{sim}(X, Y)}{k}. \qquad (4)$$

This scoring method penalises sentences which have a generally high cosine similarity with several candidates. Following Artetxe and Schwenk (2018), we use $k = 4$ in our experiments.

In the **selection process** that follows, we consider four strategies. In all of them, $\mathrm{sim}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ and $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ can be used for scoring.

(i) Threshold dependent. We find the highest scoring target sentence for each source sentence (pair i) as well as the highest scoring source for each target sentence (pair j) for either representation S=C_h *or* S=C_e (systems H and E respectively in the experiments). Since often $i \neq j$, the process is not symmetric and only pairs that have been matched during selection in both language directions are accepted to the candidate list. A threshold is empirically determined to filter out false positives.

(ii) High precision, medium recall. (system P) We apply the same methodology as before, but we use both representations S=C_h *and* S=C_e. Only pairs that have been matched during selection in both language directions *and* both representation types are accepted to the candidate list. C_h and C_e turn out to be complementary and this further restriction allows us to get rid of the threshold, and the sentence selection becomes parameter-free.

(iii) Medium precision, high recall. (system R) The combination of representations is a key point for a threshold-free method, but the final selection becomes very restrictive. In order to increase recall, we are more permissive with the way we select pairs and instead of taking only the highest scoring target sentence for each source sentence we take the top-n (n=2 in our experiments). We still use both representations and extend the number of candidates considered only for S=C_h, which is the most restrictive factor at the beginning of training.

(iv) Low precision, high recall. Generalisation of the previous strategy where we make the method symmetric in source–target and C_h–C_e.

3 Experimental Setting

Data. We use Wikipedia (WP) dumps[1] in English (en) and French (fr), and pre-process the articles and split the text into sentences using the Wikitailor toolkit[2] (Barrón-Cedeño et al., 2015). We further tokenise and truecase them using standard Moses scripts (Koehn et al., 2007) and apply a byte-pair encoding (Sennrich et al., 2016) of 100 k merge operations trained on the concatenation of English and French data. We also remove duplicates and discard sentences with more than 50 tokens for training the MT systems. We fix these settings as a comparison point for all the experiments even though smaller vocabularies and longer sentences might imply the extraction of more parallel sentences (see Section 4). We use *newstest2012* for validation and *newstest2014* for testing.

WP dumps are used for two different purposes in our systems: (*i*) to calculate initial word embeddings and (*ii*) as training corpus. In the first case, we use the complete editions (92 M sentences / 2.247 M tokens in en and 27 M / 652 M in fr). In the second case, we select only the subset of articles that can be linked among languages using Wikipedia's *langlinks* with Wikitailor, i.e., we only take an article if there is the equivalent article in the other language. For this, the total amount of sentences (tokens) is 12 M (318 M) for en and 8 M (207 M) for fr.

Model Specifications. We implemented[3] the architecture described in Section 2 within the OpenNMT toolkit (Klein et al., 2017) both for RNN and Transformer encoders, and trained:

LSTM$_{\mathbf{simP}}$: 1-layer bidirectional encoder with LSTM units, additive attention, 512-dim word embeddings and hidden states, and an initial learning rate (λ) of 0.5 with SGD. C_e and C_h are both used as representations in the high precision mode and $\mathrm{sim}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ as scoring function.

LSTM$_{\mathbf{margP}}$: The same as *LSTM*$_{\mathrm{simP}}$ but $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ as scoring function.

LSTM$_{\mathbf{margR}}$: The same as *LSTM*$_{\mathrm{margP}}$ but C_e and C_h are used in the high recall mode.

LSTM$_{\mathbf{margH}}$: As *LSTM*$_{\mathrm{margP}}$ with C_h as only representation. A hard threshold of 1.0 is used.

LSTM$_{\mathbf{margE}}$: As *LSTM*$_{\mathrm{margP}}$ with C_e as only representation. A hard threshold of 1.2 is used.

Transformer: Transformer base as defined in Vaswani et al. (2017) with 6-layer encoder–decoder with 8-head self-attention, 512-dim word embeddings and a 2048-dim hidden feed-forward. Adam optimisation with λ=2 and $beta2$=0.998; `noam` λ decay with 8000 warm-up steps. Labels are smoothed (ϵ=0.1) and a dropout mask (p=0.1) is applied.

The five models described in the LSTM category have transformer counterparts which follow the same transformer base architecture.

All systems are trained on a single GPU GTX TITAN using a batch size of 64 (LSTM) or 50 (transformer) sentences.

4 Results and Discussion

In order to train the 10 NMT systems, we initialise the word embeddings following Artetxe et al. (2017) using a seed dictionary of 2.591 numerals automatically extracted from our Wikipedia editions, and feed the system directly with comparable articles. This avoids the $n \times m$ explosion of possible combinations of sentences, where n is the number of sentences in L1 and m in L2. In our approach, we input $\sum_{\mathrm{article}} n_i \times m_j$ sentence pairs, that is, only all possible source–target sentence combinations within two articles linked by Wikipedia's langlinks. Hence we miss the parallel sentences in non-linked articles but we win in speed.

Articles are input in lots[4]. For them, the appropriate representation and scoring function are applied. Sentence pairs accepted by the selection method within a lot are extracted. Whenever enough parallel sentences are available to create a training batch, a training step is performed. Embeddings are modified by back-propagation and

[1] We use WP editions downloaded in Jan. 2015 from https://dumps.wikimedia.org/

[2] https://github.com/cristinae/WikiTailor

[3] https://github.com/ruitedk6/comparableNMT

[4] Since $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ takes into account the k-nearest neighbors of each sentence, small input lots lead to scarce information when selecting pairs. Considering lots with more than 15 sentences avoids the problem.

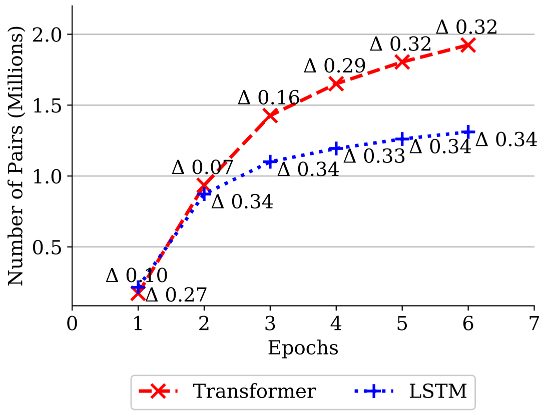

Figure 1: Number of unique accepted sentence pairs over the first 6 epochs for both margP systems. Points are labeled with the difference between the average margin scores of accepted and rejected pairs.

the next lot of articles is processed with the improved representations. Notice that the extracted pairs may therefore differ through iterations, since it is the sentence representation at the specific training step that is responsible for the selection.

Figure 1 shows the number of unique pairs selected during the first six epochs of training for both $\text{LSTM}_{\text{margP}}$ and $\text{Transformer}_{\text{margP}}$. The number of accepted sentences increases throughout the epochs, and so does the number of unique sentences used in training. Especially the first iteration over the data set is vital for improving and adapting the representations to the data itself. This quadruples the number of unique sentences accepted in the second pass over the data. While sentences are still able to pass from *rejected* to *accepted* as training advances, the two distributions are pushed apart and the gap in average margin scores between the two distributions (Δ) increases as the representations get better at discriminating. We observe curriculum learning in the process: at the beginning (epoch 1) simple sentences with *anchors* (mostly homographs such as numbers, named entities, acronyms...) are selected but as training progresses, complex semantically equivalent sentences are extracted too. Curriculum learning is important since once the capacity of a neural architecture is exhausted, more data does not improve the performance. This self-supervised architecture not only selects the data but it does it in the most useful way for the learning. It remains to be checked whether smaller vocabularies and therefore a larger number of common BPE sub-units modifies the distribution of selected sentences especially at the beginning of training.

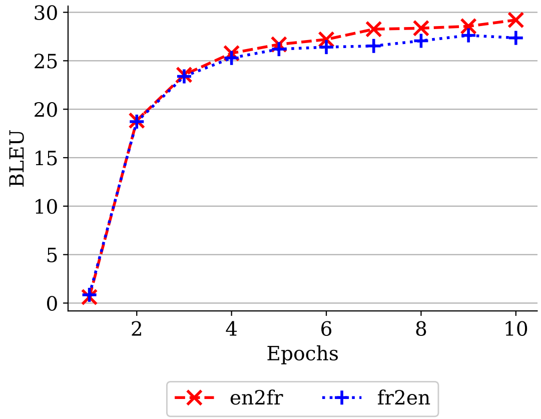

Figure 2: BLEU scores of $\text{Transformer}_{\text{margP}}$ on *newstest2014* as training progresses.

These trends are common to all our models with small nuances due to the concrete architectures. Transformers generally accumulate more unique pairs before convergence than their LSTM counterparts for example, but other than this the behaviour is the same. To validate our method, we carry out a control experiment on parallel data (Europarl) where we scramble the target sentences, creating pseudo-comparable data with a ratio of 1:5 between parallel and unrelated sentences. On this data, we can measure precision and recall and we observe how our approach progresses towards high values for these scores in both margP and margR systems. These experiments also validate the nomenclature used in Section 2: $\text{Transformer}_{\text{margR}}$ reaches higher levels of recall than $\text{Transformer}_{\text{margP}}$ (98.4% vs. 95.3%) at the cost of a lower precision (73.9% vs. 94.7%). The major increment in data through training leads to a higher translation quality as measured by BLEU, so extraction and training in a loop enhance each other's performance. Figure 2 shows the progressive improvement in translation performance throughout the training process of system $\text{Transformer}_{\text{margP}}$ and, again, the trend is general.

Table 1 summarises the final performance of our 10 systems according to BLEU. The first thing to point out is that the difference between $\text{sim}(S_{\text{L1}}, S_{\text{L2}})$ and $\text{margin}(S_{\text{L1}}, S_{\text{L2}})$ is clear and margin outperforms sim by more than 13 and 4 BLEU points for the LSTM and Transformer models respectively. The differences among the representations used with the same scoring function are not so big but still relevant. Single representation

Reference	Corpus, en+fr sent. (in millions)	BLEU $en2fr$	 $fr2en$
Unsupervised NMT			
Artetxe et al. (2018b)	NCr13, 99+32	15.13	15.56
Lample et al. (2018a)	WMT, 16+16	15.05	14.31
Yang et al. (2018)	WMT, 16+16	16.97	15.58
Self-supervised NMT			
LSTM$_{\text{simP}}$	WP, 12+8	10.48	10.97
LSTM$_{\text{margE}}$	WP, 12+8	13.71	14.26
LSTM$_{\text{margH}}$	WP, 12+8	21.50	20.84
LSTM$_{\text{margP}}$	WP, 12+8	23.64	22.95
LSTM$_{\text{margR}}$	WP, 12+8	20.05	19.45
Transformer$_{\text{simP}}$	WP, 12+8	25.21	24.96
Transformer$_{\text{margE}}$	WP, 12+8	27.33	25.87
Transformer$_{\text{margH}}$	WP, 12+8	24.45	23.83
Transformer$_{\text{margP}}$	WP, 12+8	**29.21**	**27.36**
Transformer$_{\text{margR}}$	WP, 12+8	28.01	26.78
Unsupervised NMT+SMT			
Artetxe et al. (2018a)	NCr13, 99+32	26.22	25.87
Lample et al. (2018b)	NCr17,358+69	28.10	27.20

Table 1: BLEU scores achieved on *newstest2014* with `multi-bleu.perl`. Training corpora differ by various authors: News Crawl 2007–2013 (NCr13), 2007–2017 (NCr17), the full WMT data and Wikipedia (WP).

models margE and margH (only word embeddings or encoder outputs) are 2–10 BLEU points below systems that combine both representations. It should be noted that such single representation systems *can* perform comparatively well (see Transformer$_{\text{margH}}$) if the threshold is optimally set. However, this is not guaranteed even with a preceding exploration of the threshold parameter. In margP and margR, the combinations of representations do not need such hyper-parameters and achieve the best translation quality. The best system, Transformer$_{\text{margP}}$, focuses on extracting parallel sentences with high precision and obtains BLEU scores of 29.21 ($en2fr$) and 27.36 ($fr2en$) with a total of 2.4 M selected unique sentence pairs. When increasing recall, too few new parallel sentences are gained as compared to the new false positives to improve the final translation, and Transformer$_{\text{margR}}$ and LSTM$_{\text{margR}}$ are $\sim$1–3 BLEU points below their medium recall counterparts. Notice that we do not include the *Low precision, high recall* strategy since the effect is even more pronounced.

Table 1 also presents a comparison with related work on unsupervised NMT. The comparison is delicate because training corpora and methodology differ. If we compare the final performance, we observe that we achieve similar results with less data (us vs. Lample et al. (2018b)); and when the same order of magnitude of sentences is used we obtain significantly better results (us vs. Lample et al. (2018a) and Yang et al. (2018)). The crucial difference here is that in one case one needs monolingual data, whereas we are using comparable corpora.

5 Conclusions and Future Work

We present a joint architecture to select data and train NMT systems simultaneously using the emerging NMT system itself to select the data. This is a form of self-supervision alternating between two tasks that support each other in an incremental fashion. We focus on data representation, an adequate function for the selection process, and studying how to avoid additional hyper-parameters that depend on the input corpus. The key point of our approach is the combination of a margin-based score with the intersection of sentence representations for filtering the input corpus.

As future work, we will apply our methodology to domain adaptation. In this setting, word embeddings and hidden layers are already initialised via standard NMT training on parallel data and training is continued with an in-domain monolingual or comparable corpus. Our architecture is also useful for data selection in data rich language pairs and we will perform experiments on cleaning noisy parallel corpora.

In the same vain as unsupervised MT, we want to continue our research by using back translation for rejected pairs and dealing with phrases instead of full sentences. That will allow us to extract more parallel text from a corpus and facilitate using these approaches for low-resourced languages. Existing approaches make use of huge amounts of monolingual ($\sim$100 M, references in Table 1) or comparable ($\sim$10 M, this work) sentences and these numbers are still far from what one can gather in a truly low-resource scenario.

Acknowledgments

The project on which this paper is based was partially funded by the German Federal Ministry of Education and Research under the funding code 01IW17001 (Deeplee) and by the Leibniz Gemeinschaft via the SAW-2016-ZPID-2 project (CLuBS). Responsibility for the content of this publication is with the authors.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

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

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA.

Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, and Lluís Màrquez. 2015. A factory of comparable corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Houda Bouamor and Hassan Sajjad. 2018. H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings. In *11th Workshop on Building and Using Comparable Corpora*, page 43.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2017. Lump at SemEval-2017 Task 1: Towards an Interlingua Semantic Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 144–149, Vancouver, Canada. Association for Computational Linguistics.

Cristina España-Bonet, Adám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An empirical analysis of NMT-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the International Workshop on Spoken Language Translation*, Seattle, WA.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguist*, 5:339–351.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations, ICLR*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049. Association for Computational Linguistics.

Monica Lestari Paramita, Ahmet Aker, Paul Clough, Robert Gaizauskas, Nikos Glaros, Nikos Mastropavlos, Olga Yannoutsou, Radu Ion, Dan Ştefănescu, Alexandru Ceauşu, Dan Tufiş, and Judita Preiss. 2019. *Using Comparable Corpora for Under-Resourced Areas of Machine Translation*, chapter Collecting Comparable Corpora. Springer, Cham.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, ICML'07, pages 759–766, New York, NY, USA. ACM.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234. Association for Computational Linguistics.

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, (ACL 2016), Volume 1: Long Papers*, pages 1715–1725, Berlin, Germany.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55. Association for Computational Linguistics.