Measuring Semantic Abstraction of Multilingual NMT with Paraphrase Recognition and Generation Tasks

Jörg Tiedemann and **Yves Scherrer**
Department of Digital Humanities / HELDIG
University of Helsinki

Abstract

In this paper, we investigate whether multilingual neural translation models learn stronger semantic abstractions of sentences than bilingual ones. We test this hypotheses by measuring the perplexity of such models when applied to paraphrases of the source language. The intuition is that an encoder produces better representations if a decoder is capable of recognizing synonymous sentences in the same language even though the model is never trained for that task. In our setup, we add 16 different auxiliary languages to a bidirectional bilingual baseline model (English-French) and test it with in-domain and out-of-domain paraphrases in English. The results show that the perplexity is significantly reduced in each of the cases, indicating that meaning can be grounded in translation. This is further supported by a study on paraphrase generation that we also include at the end of the paper.

1 Introduction

An appealing property of encoder-decoder models for machine translation is the effect of compressing information into dense vector-based representations to map source language input onto adequate translations in the target language. However, it is not clear to what extent the model actually needs to model meaning to perform that task; especially for related languages, it is often not necessary to acquire a deep understanding of the input to translate in an adequate way. The intuition that we would like to explore in this paper is based on the assumption that an increasingly difficult training objective will enforce stronger abstractions. In particular, we would like to see whether multilingual machine translation models learn representations that are closer to language-independent meaning representations than bilingual models do. Hence, our hypothesis is that representations learned from multilingual data sets covering a larger linguistic diversity better reflect semantics than representations learned from less diverse material. This hypothesis is supported by the findings of related work focusing on universal sentence representation learning from multilingual data (Artetxe and Schwenk, 2018; Artetxe and Schwenk, 2018; Schwenk and Douze, 2017) to be used in natural language inference or other downstream tasks. In contrast to related work, we are not interested in fixed-size sentence representations that can be fed into external classifiers or regression models. Instead, we would like to fully explore the use of the encoded information in the attentive recurrent layers as they are produced by the seq2seq model.

Our basic framework consists of a standard attentional sequence-to-sequence model as commonly used for neural machine translation (Sennrich et al., 2017), with the multilingual extension proposed by Johnson et al. (2016). This extension allows a single system to learn machine translation for several language pairs, and crucially also for language pairs that have not been seen during training. We use Bible translations for training, in order to keep the genre and content of training data constant across languages, and to enable further studies on increasing levels of linguistic diversity. We propose different setups, all of which share the characteristics of having some source data in English and some target data in English. We can then evaluate these models on their capacity of recognizing and generating English paraphrases, i.e. translating English to English without explicitly learning that task. Starting with a base model using French–English and English–French training data, we select 16 additional languages as auxiliary information that are added to the base model, each of them separately.

There is a large body of related work on

Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, pages 35–42
Minneapolis, USA, June 6, 2019.

paraphrase generation using machine translation (Quirk et al., 2004; Finch et al., 2004; Prakash et al., 2016) based on parallel monolingual corpora (Lin et al., 2014; Fader et al., 2013), pivot-based translation (Bannard and Callison-Burch, 2005; Mallinson et al., 2017) and paraphrase databased extracted from parallel corpora (Ganitkevitch et al., 2013). Related work on multilingual sentence representation (Artetxe and Schwenk, 2018; Schwenk and Douze, 2017; Lample and Conneau, 2019) has focused on fixed-size vector representations that can be used in natural language inference (Conneau et al., 2018; Eriguchi et al., 2018) or other downstream tasks such as bitext mining (Artetxe and Schwenk, 2018) or (cross-lingual) document classification (Schwenk and Li, 2018).

2 Experimental Setup

For our experiments, we apply a standard attentional sequence-to-sequence model with BPE-based segmentation. We use the Nematus-style models (Sennrich et al., 2017) as implemented in MarianNMT (Junczys-Dowmunt et al., 2018). These models apply gated recurrent units (GRUs) in the encoder and decoder with a bi-directional RNN on the encoder side. The word embeddings have a dimensionality of 512 and the RNN dimensionality is set to 1,024. We enable layer normalization and we use one RNN layer in both, encoder and decoder.

In training we use dynamic mini-batches to automatically fit the allocated memory (3GB in our case) based on sentence length in the selected sample of data. The optimization procedure applies Adam (Kingma and Ba, 2015) with mean cross-entropy as the optimization criterion. We also enable length normalization, exponential smoothing, scaling dropout for the RNN layers with ratio 0.2 and also apply source and target word dropout with ratio 0.1. All of these values are recommended settings that have empirically been found in the related literature. For testing convergence, we use independent development data of roughly 1,000 test examples and BLEU scores to determine the stopping criterion, which is set to five subsequent failures of improving the validation score. The translations are done with a beam search decoder of size 12. The validation frequency is set to run each 2,500 mini-batches.

For the multilingual setup, we follow Johnson et al. (2016) by adding target language flags to the source text placing them as pseudo tokens in the beginning of each input sentence. We always train models in both directions enabling the model to read and generate the same language without explicitly training that task (i.e. paraphrasing is modeled as zero-shot translation). BPE (Sennrich et al., 2016) is used to avoid unknown words and to improve generalisations. Note that in our setup we need to ensure that subword-level segmentations are consistent for each language involved in several translation tasks. We opted for language-dependent BPE models with 10,000 merge operations for each code table. The total vocabulary size then depends on the combination of languages that we use in training but the vocabulary stays exactly the same for each language involved in all experiments.

Language	Transl.	Verses	Tokens
English	19	234,173	6,750,869
French	14	369,910	10,529,929
Afrikaans	5	75,974	2,329,773
Albanian	2	58,192	1,648,242
Breton	1	1,781	44,316
German	24	499,844	13,712,459
Greek	7	87,218	2,357,095
Frisian	1	29,173	852,582
Hindi	4	93,242	2,829,274
Italian	5	122,363	3,429,182
Dutch	3	87,460	2,596,298
Ossetian	2	37,807	936,533
Polish	5	52,668	1,248,108
Russian	5	75,904	1,727,536
Slovene	1	29,088	748,367
Spanish	8	236,830	6,607,932
Serbian	2	35,019	844,299
Swedish	1	29,088	833,983

Table 1: Statistics about the Bible data in our collection: number of individual Bible translations, number of verses and number of tokens per language in the training data sets.

2.1 Training data and configurations

The main data we use for our experiments comes from a collection of Bible translations (Mayer and Cysouw, 2014) that includes over a thousand languages. For high-density languages like English and French, various alternatives are available (see Table 1). Using the Bible makes it possible to easily extend our work with additional languages representing a wide range of linguistic variation, while at the same time keeping genre and content constant across languages.

For the sake of discussion, we selected English

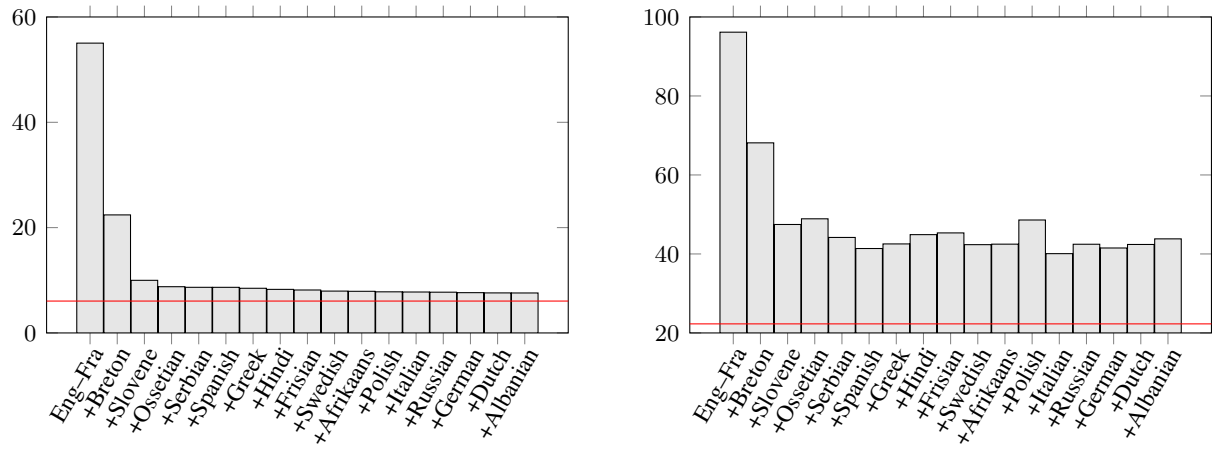

Figure 1: Paraphrase perplexity measured on Bible (left) and Tatoeba (right) test sentences (lower values are better). The figures show the effect of one auxiliary language added to the bilingual French-English model (leftmost bars). The lower red line represents the supervised model trained on English paraphrases. Languages are sorted by decreasing perplexity on the Bible data.

as our pivot language that we will use for evaluating the ability of the model to act as a paraphrase model. Furthermore, we took French as a second language to create a bilingual baseline model that can translate in both directions. As additional auxiliary languages, we then apply the ones listed in Table 1 together with some basic statistics of the training data. The idea behind the language selection is to create a somewhat diverse set of languages representing different amounts of coverage and typological relationships. The set is easy to extend but training requires extensive resources, which necessarily limits our selection at this point.

In the general setup, we do not include any pairs of English Bible translations as we do not want to evaluate a model that is specifically trained for a paraphrasing task. However, for comparison we also create a model comprising all pairs of English translation variants, which will serve as an upper bound (or rather, a lower bound in terms of perplexity) for models that are trained without explicit paraphrase data.

Exhaustively looking at all possible subsets of languages is not possible even with our small selection of 18 languages. Therefore, we restricted our study to the following test cases:

Bilingual model: A model trained on all combinations of English and French Bible translations. Each pair of aligned Bible verses represents two training instances, one for English-to-French and one for French-to-English. We also include French-to-French training instances using identical sentences in the input and output, in order to guide the model to correctly learn the semantics of the language flags.[1]

Trilingual models: Translation models trained on all bilingual combinations of Bibles in three languages – English, French and another auxiliary language (in both directions) + identical French verse pairs.

Multilingual model: One model that includes all languages in our test set with training data in both directions (translating from and to English or French) + identical French verse pairs.

Paraphrase model: A model trained on combinations of English Bible translations (the supervised upper bound).

Note that all models (including the bilingual one) cover the same English data including all Bible variants. We use exactly the same vocabulary for the English portion of each setup and no new English data is added at any point and any change that we observe when testing with English paraphrase tasks is due to the auxiliary languages that we add to the model as a translational training objective.

2.2 Test data

For our experiments, we apply test sets from two domains. One of them represents in-domain data from the Bible collection that covers 998 verses

[1] During our initial experiments, we realized that the language labels did not always pick up the information about the target language they are supposed to indicate. Especially in the bilingual case this makes sense as the model always sees the same language pair and identifying the source language is enough to determine what kind of output language it needs to generate. The label is not necessary and, therefore, ignored.

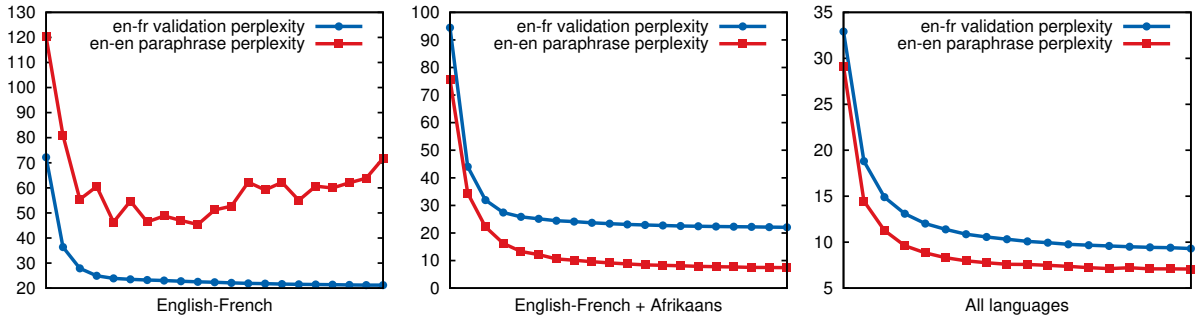

Figure 2: Learning curves from three models (the bilingual English-French model, a trilingual model and a multilingual one): Perplexity on Bible data, English-French in validation (blue) and English paraphrases in testing (red). Note the different scales.

from the New Testament that we held out of training and development sets. Our second test set comes from a very different source, namely data collected from user-contributed translations that are on-line in the Tatoeba database.[2] They include everyday expressions with translations in a large number of languages. As the collection includes translation alternatives, we can treat them as paraphrases of each other. We extracted altogether 3,873 pairs of synonymous sentences in English.

From both test data sources, we create a single-reference test set for paraphrase recognition and a multi-reference test set for paraphrase generation. The single-reference Bible test set uses the *Standard* English Bible as the source, and the *Common* English Bible[3] as the reference. The multi-reference Bible test set uses the *Amplified* Bible as the source (the first one on our list), and all 18 other English Bibles as the references.

The Tatoeba single-reference test set contains all 3,873 synonymous sentence pairs. For the multi-reference test set, we filtered the data to exclude near-identical sentence pairs by expanding contractions (like "I'm" to "I am") that are quite common in the data and removed all pairs that differ only in punctuation after that procedure. Furthermore, we merged alternatives of the same sentence into synonym sets and created, thus, a multi-reference corpus for testing containing a total of 2,444 sentences with their references.

3 Results

We evaluate the models on two tasks: (1) paraphrase recognition and (2) paraphrase generation. The following sections summarize our main findings in relation to these two tasks. We also evaluated the actual translation performance to ensure that the models are properly trained. The results of that test are listed in the supplementary material.

3.1 Paraphrase Recognition

First of all, we would like to know how well our translation models are capable of handling paraphrased sentences. For this, we compute perplexity scores of the various models when observing English output sentences for given English input coming from the two paraphrase test sets. The intuition is that models with a higher level of semantic abstraction in the encoder should be less surprised by seeing paraphrased sentences on the decoder side, which will result in a lower perplexity.

Let us first look at the in-domain data from our Bible test set. Figure 1 (left half) illustrates the reduction in perplexity when adding languages to our bilingual model. The figure is sorted by decreasing perplexities. While the picture does not reveal any clear pattern about the languages that help the most, we can see that they all contribute to an improved perplexity in comparison to the bidirectional English-French model. Breton is clearly the least useful language, without doubt due to the size of that language in our collection. Note that a further 5% perplexity reduction over the best trilingual model is achieved by the model that combines all languages (perplexity of 7.23, which is very close to the lower bound of 6.05).

The picture is similar but with a slightly different pattern on out-of-domain data. Figure 1 (right half) shows the same plot for the Tatoeba test set with languages sorted in the same order as in the previous figure. Adding languages helps again,

[2] https://tatoeba.org/eng/

[3] CEB is an ambitious new translation rather than a revision of other translations (https://www.biblegateway.com).

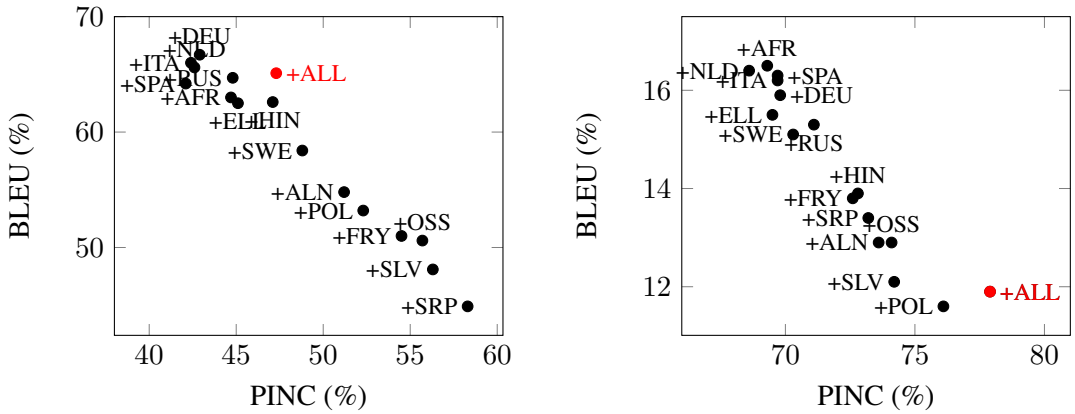

Figure 3: Paraphrase BLEU vs. PINC scores for the Bible test set (left) and the Tatoeba test set (right).

which is re-assuring, but the amount is less pronounced and further away from the lower bound (which is, however, to be expected in this setup). Again, Breton is not helping as much. Furthermore, in the out-of-domain case, the model combining all languages actually does not improve the perplexity any further (the value of 42.63 is similar to other trilingual models), which is most probably due to the strong domain mismatch that influences the scores significantly.

To further demonstrate the problems of the bilingual model to learn proper semantic representations that can be used for paraphrase detection, we can also have a look at the learning curves in Figure 2. The first plot nicely shows that the perplexity scores on paraphrase data do not follow the smooth line of the validation data in English and French whereas the models that include auxiliary languages have the capability to improve the model with respect to paraphrase recognition throughout the training procedure in a similar way as the main objective (translation) is optimized. The model that combines all languages achieves by far the lowest paraphrase perplexity. Learning curves of other trilingual models look very similar to the one included here.

3.2 Paraphrase Generation

This second experiment aims at testing the capacity of the NMT models to generate paraphrases of the input instead of translations. The hypothesis is that the generated sentences will preserve the meaning of the input, but not necessarily the same form, such that the generated sentences can be viewed as genuine paraphrases of the input sentences.

Good paraphrase models should produce sentences that are as close as possible to one of the references, yet as different as possible from the source. The first part can be measured by common machine translation metrics such as BLEU (Papineni et al., 2002), which supports multiple references. The second part can be measured by specific paraphrase quality metrics such as PINC (Chen and Dolan, 2011), which computes the proportion of non-overlapping n-grams between the source and the generated paraphrase. Good paraphrases should thus obtain high BLEU as well as high PINC scores on some paraphrase test set.

Figure 3 plots BLEU scores against PINC scores for the two test sets (lowercased and ignoring punctuations), the alternative English translations in the heldout data from the Bible and the Tatoeba paraphrase set. We exclude the bilingual model and the Breton model from the graphs, as they have BLEU scores close to 0 and PINC scores close to 100% due to the output being generated in the wrong language.

The figures show a more or less linear correlation between BLEU and PINC. This is expected to a certain extent, as there is a clear trade-off between producing varied sentences (higher PINC) and preserving the meaning of the source sentence (higher BLEU). However, we find that the model containing all languages shows the overall best performance (e.g., according to the arithmetic mean of PINC and BLEU). This suggests that a highly multilingual model provides indeed more abstract internal representations that eventually lead to higher-quality paraphrases. We also conclude that additional languages with large and diverse (i.e., many different Bibles) datasets are better at preserving the meaning of the source sentence. However, there is no obvious language fam-

Source	But even as he was on the road going down, his servants met him and reported, saying, Your son lives!
+NLD	And as he was on the road, his servants went down with him, and reported, saying, Thy son lives!
+SPA	But as it was on the road, his servants came to him and told him, "Your own Son lives!"
+ALL	And while he was on the way, his servants came to him, saying, "Your son lives!"
Source	Give attention to this! Behold, a sower went out to sow.
+AFR	Pay attention to this! Behold, the sower went out to sow.
+ALL	Take care of this. Behold, a sower went out to sow.
+BRE	Give attention to this! For, look! un semeur sortit pour semer.
+DEU	Listen to this! Behold, a sower went out to sow.

Source	He slept soundly.
Eng-Fra	Et il se prosterna devant soi.
+BRE	And, behold, he rose up quickly.
+DEU	And he began to sleep.
+ELL	He was sleeping.
+ALL	And when he had died, he was asleep.
Source	She has no brothers.
Eng-Fra	Elle n'a point de frères.
+BRE	Or, elle n'a pas de frères.
+DEU	For she has no brothers.
+OSS	No, brothers.
+ALL	You have no brothers.

Table 2: Examples of generated Bible (left) and Tatoeba (right) paraphrases.

ily or similarity effect.

The Tatoeba test set yields much lower BLEU scores than the Bible test set, due to the large number of unseen words and constructions, and also because the Tatoeba test set has only an average of 1.1 reference paraphrases per sentence, whereas the Bible test set has 18 references for each verse. This is most probably also the reason why the multilingual model including all languages (*ALL*) performs worse than most other models in terms of BLEU scores for the Tatoeba paraphrase test. It is highly likely that plausible paraphrases are not part of the test set if it only includes one or very few references like it is the case with Tatoeba, which is obviously a short-coming of BLEU as a metric for paraphrase evaluation.

Table 2 shows some examples of paraphrases generated from the Bible and Tatoeba test set. One can see that different models tend to produce different paraphrases while preserving the general meaning of the source sentence at least in the case of the Bible data. Tatoeba is more problematic due to the domain mismatch and we will come back to that issue in the discussions further down.

One caveat is that paraphrase generation could trivially be achieved by copying the input to the output especially when evaluating the results using BLEU. Therefore, we also measured the percentage of identical copies that each model produces leaving out punctuations and lowercasing the data. The results show that copying is a rare case for the multilingual models and the input is only matched in at most 1.4% of the cases (for Bible data) and at most 5.1% of the cases in the Tatoeba test set. However, adding English-English training data changes this behaviour dramatically, increasing the copying effect to over 70% of the cases in both test sets, which breaks the use of such models as a paraphrase generator. This happens even though we train on pairs of different Bible translations into English, effectively training a paraphrase model with supervised learning. Details of this evaluation are given in the supplementary material.

Source	Have you never eaten a kiwi?
+AFR	Have you not eaten sour grapes?
Source	Do you have a cellphone?
+HIN	Do you have a scorpion?
Source	Do your children speak French?
+SPA	Do your children speak Greek?
Source	Could I park my car here?
+ITA	Do I get up here with my cavalry?
Source	Birds fly.
+DEU	The flying creatures shall fly away .

Figure 4: Examples of generated Tatoeba paraphrases.

Finally, we can also observe the effect of domain mismatch between the training data and the Tatoeba test set. A considerable proportion of the test vocabulary refers to contemporary objects which obviously do not appear in the Bible training corpus, and it will, thus, be difficult for the model to generate adequate paraphrases. A few examples of sentences containing out-of-vocabulary words are shown in Figure 4. They indicate that the models are able to partially grasp the semantics of concepts and sentences often trying to replace unknown expressions with creative but reasonable alternatives coming from the context of the Bible. However, this observation calls for a more systematic evaluation of the semantic similarity of paraphrases than it is done by n-gram overlap with reference paraphrases, which is, unfortunately, out of the scope of this paper.

4 Conclusions

We have presented a study on the meaning representations that can be learned from multilingual data sets. We show that additional linguistic diversity lead to stronger abstractions and we verify our intuitions with a paraphrase scoring task that measures perplexity of multilingual sequence-to-sequence models. We also investigate the ability of translation models to generate paraphrases and conclude that this is indeed possible with promising results even without diversified decoders. In the future, we will try to push the model further to approach truly language-independent meaning representation based on massively parallel data sets as additional translational grounding. We will also study the model with bigger and less homogenous data sets and compare it to other approaches to paraphrase generation including pivot-based back-translation models. Furthermore, we will test sentence representations obtained by multilingual NMT models with additional downstream tasks to further support the main claims of the paper.

Acknowledgments

The work in this paper is supported by the Academy of Finland through project 314062 from the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113). We would also like to acknowledge NVIDIA and their GPU grant.

References

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *CoRR*, abs/1811.01136.

Mikel Artetxe and Holger Schwenk. 2018. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *arXiv e-prints*, page arXiv:1812.10464.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the ACL 2005*, pages 597–604, Ann Arbor, Michigan.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of ACL 2011*, pages 190–200, Portland, Oregon, USA.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. *CoRR*, abs/1809.05053.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *CoRR*, abs/1809.04686.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of ACL 2013*, pages 1608–1618, Sofia, Bulgaria.

Andrew Finch, Taro Watanabe, Yasuhiro Akiba, and Eiichiro Sumita. 2004. Paraphrasing as machine translation. 11:87–111.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL 2013*, pages 758–764, Atlanta, Georgia.

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. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. *arXiv preprint arXiv:1804.00344*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of EACL 2017*, pages 881–893, Valencia, Spain.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proc. of LREC*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, Pennsylvania, USA.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *Proceedings of COLING 2016*, pages 2923–2934.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP 2004*, pages 142–149, Barcelona, Spain.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *ACL workshop on Representation Learning for NLP*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *LREC*, pages 3548–3551.

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. *CoRR*, abs/1703.04357.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL 2016*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.