Strengthening Low-resource Neural Machine Translation through Joint Learning: The Case of Farsi-Spanish

Benyamin Ahmadnia[1], Raul Aranovich[1] and Bonnie J. Dorr[2]
[1]*Department of Linguistics, University of California, Davis, CA, U.S.A.*
[2]*Institute for Human and Machine Cognition (IHMC), Ocala, FL, U.S.A.*

Keywords: Computational Linguistics, Natural Language Processing, Neural Machine Translation, Low-Resource Languages, Joint Learning.

Abstract: This paper describes a systematic study of an approach to Farsi-Spanish low-resource Neural Machine Translation (NMT) that leverages monolingual data for joint learning of forward and backward translation models. As is standard for NMT systems, the training process begins using two pre-trained translation models that are iteratively updated by decreasing translation costs. In each iteration, either translation model is used to translate monolingual texts from one language to another, to generate synthetic datasets for the other translation model. Two new translation models are then learned from bilingual data along with the synthetic texts. The key distinguishing feature between our approach and standard NMT is an iterative learning process that improves the performance of both translation models, simultaneously producing a higher-quality synthetic training dataset upon each iteration. Our empirical results demonstrate that this approach outperforms baselines.

1 INTRODUCTION

A major difference between Neural Machine Translation (NMT) (Cho et al., 2014; Bahdanau et al., 2015) and Statistical Machine Translation (SMT) (Koehn et al., 2003; Chiang, 2007) is the way monolingual data are used in these two paradigms. SMT seamlessly integrates very large Language Models (LMs) trained on millions of sentences, while NMT is best supported by the generation of artificial (synthetic) parallel data (Ahmadnia et al., 2018). Making effective use of monolingual data is particularly critical under low-resource conditions, where the bilingual dataset is generally assumed to be small in comparison to available monolingual texts. Monolingual datasets are usually much easier than bilingual datasets to collect, and have been attractive resources for improving corpus-based MT models. Indeed, monolingual data play a key role in training data-driven MT systems. However, NMT systems rely heavily on high-quality bilingual datasets and, in fact, perform poorly when such datasets are small or unavailable.

This paper describes an approach that addresses this shortcoming by leveraging monolingual data from both source and target sides to jointly optimize forward and backward models. In an iterative process, each translation model helps the other, as in the process of back-translation (where translated text is interpreted back to the original language). Specifically, the backward model uses target monolingual data to generate synthetic data for the forward model, while the forward model employs source monolingual data to generate synthetic data for the backward model. A key advantage over prior work (Zhang et al., 2018) is that iterative training yields further enhancements and, after each iteration, both models are expected to improve with additional synthetic data. That is, each iteration yields better synthetic data with the two enhanced models than on the prior iteration.

Noisy translations are minimized through the use of a learning objective that assigns weights to the newly generated sentence pairs. Initial bilingual sentence pairs are all weighted as 1, while synthetic sentence pairs are weighted via the normalized model output probability. Weighting plays an important role in augmenting the final translation quality. The overall iterative training process essentially adds a joint Expectation-Maximization (EM) estimation over the monolingual data to the Maximum Likelihood Estimation (MLE) over bilingual data. For example, the E-step estimates the expectations of translations of the monolingual data, while the M-step updates model parameters with the smoothed translation probability

Ahmadnia, B., Aranovich, R. and Dorr, B.
Strengthening Low-resource Neural Machine Translation through Joint Learning: The Case of Farsi-Spanish.
DOI: 10.5220/0010362604750481
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 1*, pages 475-481
ISBN: 978-989-758-484-8

estimation. Our experimental results show that this joint learning approach not only outperforms baseline systems but also significantly strengthens translation quality of both the forward and backward model.

Our motivation for choosing Spanish and Farsi as the case-study is the linguistic differences between these languages, which are from different language families and have significant differences in their properties, may pose a challenge for MT. Following (Ahmadnia and Dorr, 2019), low-resource languages, also known as resource poor, are those that have fewer technologies and datasets relative to some measure of their international importance. In simple words, the languages for which parallel training data is extremely sparse, requiring recourse to techniques that are complementary to standard MT approaches. The biggest issue with low-resource languages is the extreme difficulty of obtaining sufficient resources. Natural Language Processing (NLP) methods that have been created for analysis of low-resource languages are likely to encounter similar issues to those faced by documentary and descriptive linguists whose primary endeavor is the study of minority languages. Lessons learned from such studies are highly informative to NLP researchers who seek to overcome analogous challenges in the computational processing of these types of languages.

MT has proven successful for a number of language pairs. However, each language comes with its own challenges, and Farsi is no exception. Farsi suffers significantly from shortage of digitally available parallel and monolingual texts. It is morphologically rich, with many characteristics shared only by Arabic. It makes no use of articles (*a*, *an*, *the*) and no distinction between capital and lower-case letters. Symbols and abbreviations are rarely used. As a consequence of being written in the Arabic script, Farsi uses a set of diacritic marks to indicate vowels, which are generally omitted except in infant writing or in texts for those who are learning the language. Sentence structure is also different from that of English. Farsi places parts of speech such as nouns, subjects, adverbs and verbs in different locations in the sentence, and sometimes even omits them altogether. Some Farsi words have many different accepted spellings, and it is not uncommon for translators to invent new words. This can result in OOV words.

Spanish utilizes the Latin alphabet, with a few special letters, vowels with an acute accent *(á, ú, é, ó, í)*, *u* with an umlaut (*ü*), and an *n* with a tilde (*ñ*). Due to a number of reforms, the Spanish spelling system is almost perfectly phonemic and, therefore, easier to learn than the majority of languages. Spanish is pronounced phonetically, but includes the trilled *r* which is somewhat complex to reproduce. In the Spanish IPA, the letters *b* and *v* correspond to the same symbol *b* and the distinction only exists in regional dialects. The letter *h* is silent except in conjunction with *c*, *ch*, which changes the sound into *tf*. Spanish language punctuation is very close to English. There are a few significant differences. For example, in Spanish, exclaim and interrogative sentences are preceded by inverted question and exclamation marks. Also, in a Spanish conversation, a change in speakers is indicated by a dash, while in English, each speaker's remark is placed in separate paragraphs. Formal and informal translations address several different characteristics. Inflection, declination and grammatical gender are important features of Spanish language.

A number of *divergences* (Dorr, 1994; Dorr et al., 2002) between low-resource (e.g., Farsi) and high-resource (e.g., Spanish) languages pose many challenges in translation. In Farsi, the modifier precedes the word it modifies, and in Spanish the modifier follows the head word (although it may precede the head word under certain conditions). In Farsi, sentences follow a "Subject", "Object", "Verb" (SOV) order, and in Spanish, the sentences follow the "Subject", "Verb", "Object" (SVO) order (Ahmadnia et al., 2017). Such distinctions are exceedingly prevalent and thus pose many challenges for machine translation.

2 RELATED WORK

Prior approaches to monolingual data-driven NMT fall into three categories: (1) integration of LMs trained with monolingual data; (2) generation of pseudo-sentence pairs from monolingual data; and (3) joint training of both source↔target translation models by minimizing reconstruction errors of monolingual sentences.

In the first category, a LM is separately trained with monolingual data and then integrated into the NMT model. In the work of (Gülçehre et al., 2015), monolingual LMs are trained independently, and then integrated during decoding through rescoring or by adding the LM's recurrent hidden state to the decoder state of the encoder-decoder network. An additional controller mechanism is used, to control the magnitude of the LM signal.

In the second category, translation models trained from bilingual sentence pairs are applied to monolingual data. Sentences from the original monolingual data are then paired with their translated counterparts to form a pseudo parallel corpus for a larger training set. A successful approach is that of (Sen-

nrich et al., 2016), wherein target monolingual data are leveraged to generate artificial parallel data via back-translation. This approach has proven effective, but generated pseudo bilingual sentence pairs yield limited performance gains over the use of monolingual data alone.

In the third category, monolingual data are reconstructed with both source-to-target and target-to-source translation models, and the two models are jointly trained (Ahmadnia and Dorr, 2019). (He et al., 2016) treats the forward and backward models as the primal and dual tasks, respectively. (Cheng et al., 2016) uses both source and target monolingual data for semi-supervised reconstruction where two NMTs are employed, one that translates the source monolingual data into target translations, and the other that reconstructs source monolingual data from target translations.

(Currey et al., 2017) trained a NMT system to both translate source text and copy target text, thereby exploiting monolingual corpora in the target language. Specifically, they created a bilingual corpus from the monolingual text in the target language so that each source sentence is identical to the target sentence. This copied data is then mixed with the parallel corpus and the NMT system is trained like normal, with no metadata to distinguish the two input languages. In fact, their method proves to be an effective way of incorporating monolingual data into low-resource NMT.

(Luong et al., 2015) adopted a simple auto-encoder or skip-thought method (Kiros et al., 2015) to exploit the source monolingual data, but no significant BLEU gains are reported. Also, (Zhang and Zong, 2016) investigated the usage of the source large-scale monolingual data in NMT and they aimed at greatly enhancing its encoder network so that they could obtain high-quality context vector representations. They proposed the self-learning algorithm to generate the synthetic large-scale parallel data for NMT training as well as the multi-task learning framework using two NMTs to predict the translation and the reordered source-side monolingual sentences simultaneously.

Our work transcends issues described above by using source monolingual data to augment reverse NMT models. We adopt EM to iteratively update bidirectional NMT models. We exploit either source and target monolingual data and demonstrates improvements over the use of target monolingual data alone.

(Ramachandran et al., 2017) adopted pre-trained weights of two language models to initial the encoder and decoder of a sequence-to-sequence NMT model, and then fine-tune it with labeled data. Their approach is complementary to ours by leveraging pre-trained language model to initial bidirectional NMT models, and it may lead to additional gains.

3 METHOD DESCRIPTION

Joint learning expands the task setting from solely enhancing the forward NMT model training with a target monolingual dataset to enhancing the model with a paired dataset. This approach aims at jointly optimizing either a forward or a backward NMT model with the help of a monolingual dataset from both source and target languages.

Given a set of sentences $Y = y_1, y_2, ..., y_n$ in target language, and a set of sentences $X = x_1, x_2, ..., x_n$ in source language. First, the initial forward and backward neural TMs are pre-trained with a bilingual dataset D, defined as:

$$\left\{x^{(n)}, y^{(n)}\right\}_{n=1}^{N}$$

where N denotes the number of sentences in D. At the beginning of the next iteration, the two TMs are used to translate monolingual datasets X and Y, yielding two synthetic training datasets (Y' and X'). Either the forward or the backward model is then trained on the updated training dataset by combining Y' and X' with D.

The k-best translations from a NMT system are weighted with the translation probabilities from the NMT model. In the next iteration, the aforementioned process is iterated. However, the synthetic training dataset is re-generated through the updated forward and backward models. The learned forward and backward models are enhanced over the first iteration (iteration 0). The joint learning approach adds an EM process over the monolingual data in both source and target languages. However, the training criteria on D still uses MLE.

Let $\hat{Y}$ be monolingual target-language corpus:

$$\left\{\hat{y}^{(z)}\right\}_{z=1}^{Z}$$

We derive the new learning objective for joint learning e.g., the learning objective is to maximize the likelihood of the monolingual dataset as well as the bilingual dataset as follows:

$$C = \sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)}) + \sum_{z=1}^{Z} \log P(\hat{y}^{(z)}) \qquad (1)$$

where

$$\sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)})$$

denotes the likelihood of bilingual dataset, and

$$\sum_{z=1}^{Z} \log P(\hat{y}^{(z)})$$

represents target monolingual dataset likelihood.

We define the source translations as hidden states for the corresponding target sentences and decompose $\log P(y^{(z)})$ as follows:

$$\log P(\hat{y}^{(z)}) = \log \sum_{x} P(x, \hat{y}^{(z)}) = \log \sum_{x} W(x) \frac{P(x, \hat{y}^{(z)})}{W(x)} \quad (2)$$

where x represents a latent variable of the source translation of target sentence, $W(x)$ is the approximated probability distribution of x, $P(x)$ represents the marginal distribution of sentence x. $W(x)$ must satisfy the following condition:

$$f = \frac{P(x, \hat{y}^{(z)})}{Q(x)}$$

where f is a constant and does not depend on y. Given $\sum_x W(x) = 1$, $W(x)$ is defined as follow:

$$W(x) = \frac{P(x, \hat{y}^{(z)})}{\sum_x P(x, \hat{y}^{(z)})} \quad (3)$$

We use $P(x|\hat{y}^{(z)})$ given by backward TM as $Q(x)$ and combine Equations (1) and (2):

$$C_{forward} = \sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)}) + \sum_{z=1}^{Z} \sum_{x} \log P(x|\hat{y}^{(z)}) \log P(\hat{y}^{(z)}|x) \quad (4)$$

where

$$\sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)})$$

is the same as MLE training estimated in the E-step, and

$$\sum_{z=1}^{Z} \sum_{x} \log P(x|\hat{y}^{(z)}) \log P(\hat{y}^{(z)}|x)$$

is optimized via EM and maximized in the M-step.

The E-step uses the forward NMT model to generate the source translations as hidden variables, which are paired with the target sentences to build a new distribution of training data together with D. Thus, maximization of C is approximated by maximizing the log-likelihood on the new training data. The translation probability is utilized as the weight of the synthetic sentence pairs, which helps with filtering out low-quality translations.

Back-translation (Sennrich et al., 2016) is a successful exploitation method of monolingual data where an NMT system is first trained in the reverse direction (backward) and is then used to translate target monolingual data back into the source language. The resulting sentence pairs constitute a pseudo bilingual dataset to be added to the initial training data to learn a forward model.

It is easy to verify that back-translation is a special case of the formulation of $C_{forward}$ in which $P(x|\hat{y}^{(z)}) = 1$ because only the best translation from the backward NMT model is used

$$C_{forward} = \sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)}) + \sum_{z=1}^{Z} \log P(\hat{y}^{(z)}|\hat{y}^{(z)}_{backward}) \quad (5)$$

Similarly, the likelihood of the backward model can be derived as follows:

$$C_{backward} = \sum_{n=1}^{N} \log P(x^{(n)}|y^{(n)}) + \sum_{k=1}^{K} \sum_{y} P(y|x^{(k)}) \log P(x^{(k)}|y) \quad (6)$$

where y is a target translation (hidden state) of $x^{(k)}$. The overall learning objective is the sum of likelihood in both directions ($C_{total} = C_{forward} + C_{backward}$). During the derivation of $C_{forward}$, we use the translation probability from the backward model as the approximation of $P'(x|\hat{y}^{(z)})$. When $P(x|\hat{y}^{(z)})$ gets closer to $P'(x|\hat{y}^{(z)})$, we get a tighter lower bound of $C'_{forward}$, gaining more opportunities to improve the forward model.

4 EXPERIMENTAL RESULTS

We applied joint learning to Farsi↔Spanish translation. We selected the training data from *Tanzil* collection (Tiedemann, 2012), which consists of 50K parallel sentence pairs. We randomly selected 0.5M Farsi sentences as well as 0.5M Spanish sentences extracted from *Opensubtitles2018* (Lison and Tiedemann, 2016) as the monolingual datasets. In all cases, any sentence longer than 50 words is removed from the training dataset. As the validation dataset, we used 5K parallel sentences from *Tanzil* corpus. We also used the 10K parallel sentences from *Tanzil* corpus as our test dataset. We limited the vocabulary size to contain up to 50K most frequent words, and convert remaining words into the <UNK> token.

Source sentence	وقتی سوت داور به صدا در آمد ، پایتخت اسپانیا مادرید غرش کرد
Reference	cuando sonó el silbato del árbitro, la capital Española de Madrid rugió
Translations	[iteration 0]: la capital española de madrid estaba rugiendo con el madrid ---------- [iteration 2]: la capital española de madrid estaba rugiendo con el sonido del final de la puerta ---------- [iteration 4]: cuando sonó el silbato del árbitro, la capital española de madrid estaba rugiendo

Figure 1: Example translations of a Farsi sentence in different iterations.

For the implementation we utilize *Transformer* (Vaswani et al., 2017) on top of PyTorch, which uses a 6-layer LSTM encoder-decoder and the hidden layer of 1024 in our experiments. The training uses a mini-batch of 256 and the Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951) with an initial learning rate of 0.01. We set the size of word embeddings layer to 512. We also set dropout to 0.1. We use a maximum sentence length of 50 words. We also set a beam size of 8, and the model continues for 20 epochs (in both training and test steps) on a single GPU. We employed Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2001) (higher is better) and Translation Error Rate (TER) (Snover. et al., 2006) (lower is better) as the evaluation metrics.

For building the synthetic bilingual texts, we set the beam size to 4 to speed up the decoding process. We first sort all monolingual data according to the sentence length and then 128 sentences are simultaneously translated with parallel decoding implementation. As for model training, 10 EM iterations are found to be sufficient for convergence.

Joint learning (NMT-JL) is compared to standard attention-based NMT system trained on bilingual corpora (NMT-baseline), round-tripping (NMT-RT) (Ahmadnia and Dorr, 2019), and unsupervised-learning (NMT-UL) (Artetxe et al., 2019). Tables 1 and 2 show performance results on Farsi↔Spanish translations.

It is worth noting that more iterations lead to better evaluation results consistently, which validates the hypothesis that joint training of NMT models in two directions boosts translation quality.

Figure 1 shows Farsi-Spanish translation results in different iterations. Specifically, iteration 0 corresponds to the scored baseline from Table 1, and obviously, the first few iterations gain most, especially for Iteration 2.

After three iterations (0, 2, and 4), no significant improvements are observed. As the target-source model approaches the ideal translation probability, the lower bound of the cost is closer to the true cost and there is a smaller potential for gain. Since there is a lot of uncertainty during the training, the performance sometimes drops a little, generally yielding little (or no) net gain.

Table 1: The Farsi-to-Spanish translation results (the "Fa" denotes Farsi and the "Es" denotes Spanish).

Translation	**Model**	**BLEU**	**TER**
Fa-Es	NMT-baseline	33.12	53.04
Fa-Es	NMT-RT	35.66	50.83
Fa-Es	NMT-UL	36.19	49.39
Fa-Es	NMT-JL	**38.53**	**47.29**

Table 2: Spanish-to-Farsi translation results (the "Fa" denotes Farsi and the "Es" denotes Spanish).

Translation	**Model**	**BLEU**	**TER**
Es-Fa	NMT-baseline	31.02	55.64
Es-Fa	NMT-RT	33.97	52.35
Es-Fa	NMT-UL	35.06	51.24
Es-Fa	NMT-JL	**35.88**	**49.41**

NMT-JL can be considered a general version of NMT-RT where any pseudo sentence pair is weighted as 1. NMT-JL slightly surpasses NMT-RT on all test datasets, which confirms that the weight can lead to better performance. This approach assigns a low weight to synthetic sentence pairs with poor translations, so as to punish their effect on model updates. The translation is improved in subsequent iterations.

5 CONCLUSIONS AND FUTURE WORK

We have applied a joint learning approach to integrating the training of a pair of TMs in a unified learning process with the help of monolingual data from both source and target sides. A joint-EM learning technique is employed to optimize two TMs cooperatively. The resulting framework enables two models to jointly boost each other's translation performance. Translation probabilities associated with each model are used to compute weights that estimate the translation accuracy and punish the low-quality translations.

As a future work, we are interested in extending the present method to jointly learn multiple NMT systems for several languages employing massive amount of monolingual datasets.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their valuable feedback and discussions. We also would like to acknowledge the financial support received from the Linguistics Department at UC Davis (USA).

REFERENCES

Ahmadnia, B. and Dorr, B. J. (2019). Augmenting neural machine translation through round-trip training approach. *Open Computer Science*, 9(1):268–278.

Ahmadnia, B., Kordjamshidi, P., and Haffari, G. (2018). Neural machine translation advised by statistical machine translation: The case of farsi-spanish bilingually low-resource scenario. In *Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1209–1213.

Ahmadnia, B., Serrano, J., and Haffari, G. (2017). Persian-Spanish low-resource statistical machine translation through english as pivot language. In *Proceedings of Recent Advances in Natural Language Processing*, pages 24–30.

Artetxe, M., Labaka, G., and Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., and Liu, Y. (2016). Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1965–1974.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Cho, K., merrienboer, B. V., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Currey, A., Barone, M., and andK. Heafield, A. V. (2017). Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Dorr, B. J. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Dorr, B. J., Pearl, L., Hwa, R., and Habash, N. (2002). Duster: A method for unraveling cross-language divergences for statistical word-level alignment. In *Proceedings of the 5th conference of the Association for Machine Translation in the Americas*.

Gülçehre, Ç., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *ArXiv*, abs/1503.03535.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W. (2016). Dual learning for machine translation. In *Proceedings of the 30th Conference on Neural Information Processing Systems*.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Proceedings of the 29th Conference on Advances in Neural Information Processing Systems*, pages 3294–3302.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*.

Luong, T., Sutskever, I., Le, Q., Vinyals, O., and Zaremba, W. (2015). Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 11–19.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2001). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Ramachandran, P., Liu, P., and Le, Q. (2017). Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 383–391.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of Association for Computational Linguistics*, pages 86–96.

Snover., M., Dorr, B. J., Schwartz, R., Micciulla, L., and Weischedel, R. (2006). A study of translation error rate with targeted human annotation. In *Proceedings of the Association for Machine Transaltion in the Americas*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Advances in Neural Information Processing Systems*, pages 5998–6008.

Zhang, J. and Zong, C. (2016). Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Zhang, Z., Liu, S., Li, M., Zhou, M., and Chen, E. (2018). Joint training for neural machine translation models with monolingual data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 555–562.