Multi-Source Neural Machine Translation with Missing Data

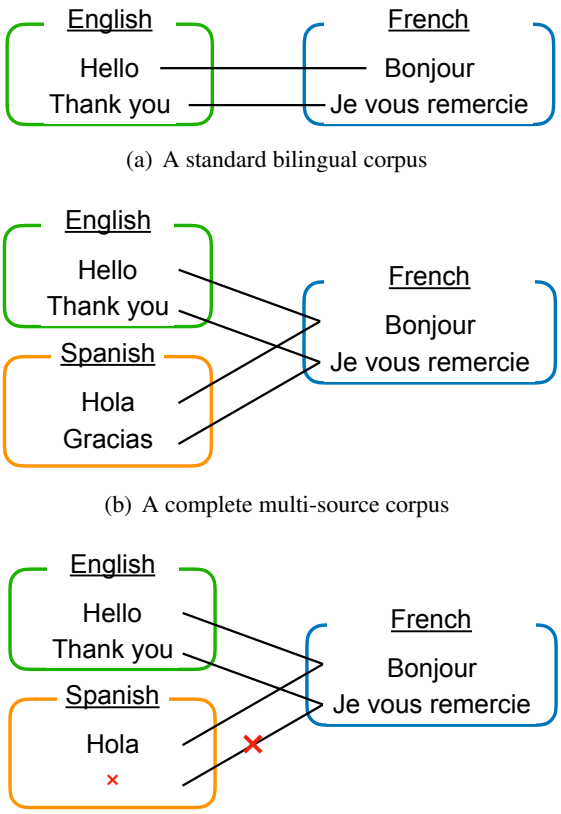
Yuta Nishimura[1] **, Katsuhito Sudoh**[1] **, Graham Neubig**[2,1]**, Satoshi Nakamura**[1]
[1]Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan
[2]Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
{nishimura.yuta.nn9, sudoh, s-nakamura}@is.naist.jp
gneubig@cs.cmu.edu

Abstract

Multi-source translation is an approach to exploit multiple inputs (e.g. in two different languages) to increase translation accuracy. In this paper, we examine approaches for multi-source neural machine translation (NMT) using an *incomplete* multilingual corpus in which some translations are missing. In practice, many multilingual corpora are not complete due to the difficulty to provide translations in *all* of the relevant languages (for example, in TED talks, most English talks only have subtitles for a small portion of the languages that TED supports). Existing studies on multi-source translation did not explicitly handle such situations. This study focuses on the use of incomplete multilingual corpora in multi-encoder NMT and mixture of NMT experts and examines a very simple implementation where missing source translations are replaced by a special symbol <NULL>. These methods allow us to use incomplete corpora both at training time and test time. In experiments with real incomplete multilingual corpora of TED Talks, the multi-source NMT with the <NULL> tokens achieved higher translation accuracies measured by BLEU than those by any one-to-one NMT systems.

1 Introduction

In general, machine translation systems translate from one source language to a target language. For example, we may translate a document or speech that was written in English to a new language such as French. However, in many real translation scenarios, there are cases where there are multiple languages involved in the translation process. For example, we may have an original document in English, that we want to translate into several languages such as French, Spanish, and Portuguese. Some examples of these scenarios are the creation of video captions for talks (Cettolo et al., 2012) or Movies (Tiedemann, 2009), or translation of official documents into all the languages of a governing body, such as the European parliament (Koehn, 2005) or UN (Ziemski et al., 2016). In these cases, we are very often faced with a situation where we *already* have good, manually cu-



(a) A standard bilingual corpus

(b) A complete multi-source corpus

(c) An incomplete multi-source corpus with missing data

Figure 1: Example of type of corpora.

rated translations in a number of languages, and we'd like to generate translations in the remaining languages for which we do not yet have translations.

In this work, we focus on this sort of multilingual scenario using multi-source translation (Och and Ney, 2001; Zoph and Knight, 2016; Garmash and Monz, 2016). Multi-source translation takes in multiple inputs, and references all of them when deciding which sentence to output. Specifically, in the context of neural machine translation (NMT), there are several methods proposed to do so. For example, Zoph and Knight (2016) propose a method where multiple sentences are each encoded separately, then all referenced during the decoding process (the "multi-encoder" method). In addition, Garmash and Monz (2016) propose a method where NMT systems over multiple inputs are ensembled together to make a final prediction (the "mixture-of-NMT-experts" method).

However, this paradigm assumes that we have data in *all* of the languages that go into our multi-source system. For example, if we decide that English and Spanish are our input languages and that we would like to translate into French, we are limited to training and testing only on data that contains all of the source languages. However, it is unusual that translations in all of these languages are provided there will be many sentences where we have only one of the sources. In this work, we consider methods for multi-source NMT with missing data, such situations using an *incomplete* multilingual corpus in which some translations are missing, as shown in Figure 1. This incomplete multilingual scenario is useful in practice, such as when creating translations for incomplete multilingual corpora such as subtitles for TED Talks.

In this paper, we examine a simple implementation of multi-source NMT using such an incomplete multilingual corpus that uses a special symbol <NULL> to represent the missing sentences. This can be used with any existing multi-source NMT implementations without no special modifications. Experimental results with real incomplete multilingual corpora of TED Talks show that it is effective in allowing for multi-source NMT in situations where full multilingual corpora are not available, resulting in BLEU score gains of up to 2 points compared to standard bi-lingual NMT.

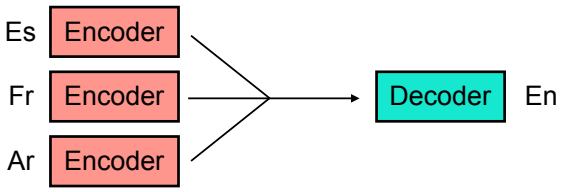


Figure 2: Multi-encoder NMT

2 Multi-Source NMT

At the present, there are two major approaches to multi-source NMT: multi-encoder NMT (Zoph and Knight, 2016) and mixture of NMT experts (Garmash and Monz, 2016). We first review them in this section.

2.1 Multi-Encoder NMT

Multi-encoder NMT (Zoph and Knight, 2016) is similar to the standard attentional NMT framework (Bahdanau et al., 2015) but uses multiple encoders corresponding to the source languages and a single decoder, as shown in Figure 2.

Suppose we have two LSTM-based encoders and their hidden states and cell states at the end of the inputs are h_1, h_2 and c_1, c_2, respectively. The multi-encoder NMT method initializes its decoder hidden states h and cell state c as follows:

$$h = \tanh(W_c[h_1; h_2]) \quad (1)$$

$$c = c_1 + c_2 \quad (2)$$

Attention is then defined over each encoder at each time step t and resulting context vectors c_t^1 and c_t^2, which are concatenated together with the corresponding decoder hidden state h_t to calculate the final context vector $\tilde{h_t}$.

$$\tilde{h_t} = \tanh(W_c[h_t; c_t^1; c_t^2]) \quad (3)$$

The method we base our work upon is largely similar to Zoph and Knight (2016), with the exception of a few details. Most notably, they used *local-p* attention, which focuses only on a small subset of the source positions for each target word (Luong et al., 2015). In this work, we used *global* attention, which attends to all words on the source side for each target word, as this is the standard method used in the great majority of recent NMT work.

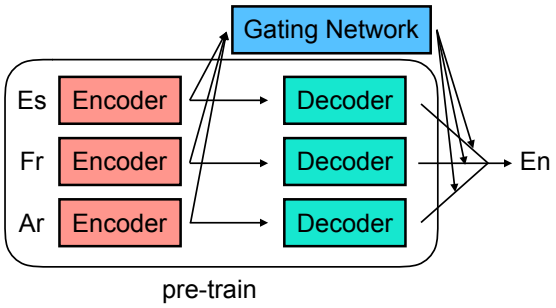


Figure 3: Mixture of NMT Experts

2.2 Mixture of NMT Experts

Garmash and Monz (2016) proposed another approach to multi-source NMT called *mixture of NMT experts*. This method ensembles together independently-trained encoder-decoder networks. Each NMT model is trained using a bilingual corpus with one source language and the target language, and the outputs from the one-to-one models are summed together, weighted according to a gating network to control contributions of the probabilities from each model, as shown in Figure 3.

The mixture of NMT experts determines an output symbol at each time step t from the final output vector p_t^ϵ, which is the weighted sum of the probability vectors from one-to-one models denoted as follows:

$$p_t^\epsilon = \sum_{j=1}^{m} g_t^j p_t^j \quad (4)$$

where p_t^j and g_t^j are the probability vector from j-th model and the corresponding weight at time step t, respectively. m is the number of one-to-one models. g_t is calculated by the gating network as follows:

$$g_t = \mathrm{softmax}(W_{gate} \tanh(W_{hid}[f_t^1(x); ...f_t^m(x)])) \quad (5)$$

where $f_t^j(x)$ is the input vector to the decoder of the j-th model, typically the embedding vector for the output symbol at the previous time step t-1.

3 Multi-Source NMT with Missing Data

In this work, we examine methods to use incomplete multilingual corpora to improve NMT in a specific language pair. This allows multi-source techniques to be applied, reaping the benefits of other additional languages even if some translations in these additional languages are missing.

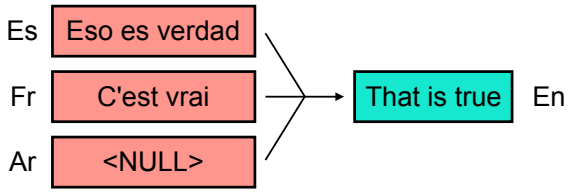


Figure 4: Multi-encoder NMT with a missing input sentence

Specifically, we attempt to extend the methods in the previous section to use an incomplete multilingual corpus in this work.

3.1 Multi-encoder NMT

In multi-encoder NMT, each encoder must be provide with an input sentence, so incomplete multilingual corpora cannot be used as-is.

In this work, we employ a very simple modification that helps resolve this issue: replacing each missing input sentence with a special symbol <NULL>. The special symbol <NULL> can be expected to be basically ignored in multi-encoder NMT, with the decoder choosing word hypotheses using other input sentences. Note that this method can be applied easily to any existing implementation of the multi-encoder NMT with no modification of the codes.

Figure 4 illustrates the modified multi-encoder NMT method. Here the source languages are Spanish, French, and Arabic and the target language is English, and the Arabic input sentence is missing. Here, the Spanish and French input sentences are passed into the corresponding encoders and <NULL> is input to the Arabic encoder.

3.2 Mixture of NMT Experts

In the mixture of NMT experts method, each one-to-one NMT model can be trained independently using incomplete multilingual corpora. However, we still need a complete multilingual corpus to train the gating network.

We also employ a special symbol <NULL> in the mixture of NMT experts to deal with missing input sentences in the same way as the multi-encoder NMT described above. The gating network can also be expected to learn to ignore the outputs from the missing inputs.

4 Experiments

We conducted two experiments with different incomplete multilingual corpora. One is an experiment with a pseudo-incomplete multilingual corpus, the other is an experiment with an actual incomplete multilingual corpus.

4.1 NMT settings

We describe the settings of common parts for all NMT models: multi-encoder NMT, mixture of NMT experts, and one-to-one NMT. We used global attention and attention feeding (Luong et al., 2015) for the NMT models and used a bidirectional encoder (Bahdanau et al., 2015) in their encoders. The number of units was 512 for the hidden and embedding layers. Vocabulary size was the most frequent 30,000 words in the training data for each source and target languages. The parameter optimization algorithm was Adam (Kingma and Ba, 2015) and gradient clipping was set to 5. The number of hidden state units in the gating network for the mixture of NMT experts experiments was 256. We used BLEU (Papineni et al., 2002) as the evaluation metric. We performed early stopping, saving parameter values that had the smallest log perplexities on the validation data and used them when decoding test data.

4.2 Pseudo-incomplete multilingual corpus (UN6WAY)

First, we conducted experiments using a *complete* multilingual corpus and a *pseudo-incomplete* corpus derived by excluding some sentences from the complete corpus, to compare the performance in complete and incomplete situations.

4.2.1 Data

We used UN6WAY (Ziemski et al., 2016) as the complete multilingual corpus. We chose Spanish (Es), French (Fr), and Arabic (Ar) as source languages and English (En) as a target language The training data in the experiments were the 0.8 million sentences from the UN6WAY corpus whose sentence lengths were less than or equal to 40 words. We excluded 200,000 sentences for each language except English for the pseudo-incomplete multilingual corpus as shown in Table 1. "Sentence number" in Table 1 represents the line number in the corpus, and the *x* means the part removed for the incomplete multilingual corpus. We also chose 1,000 and 4,000 sentences for validation and test from the UN6WAY corpus, apart from the training data. Note that the validation and test data here had no missing translations.

Sentence No.	Es	Fr	Ar	En
1-200,000	x			
200,001-400,000			x	
400,001-600,000		x		
600,001-800,000				

Table 1: Settings of the pseudo-incomplete UN multilingual corpus (x means that this part was deleted)

4.2.2 Setup

We compared multi-encoder NMT and the mixture of NMT experts in the complete and incomplete situations. The three one-to-one NMT systems, Es-En, Fr-En, and Ar-En, which were used as submodels in the mixture of NMT experts, were also compared for reference.

First, we conducted experiments using all of the 0.8 million sentences in the complete multilingual corpus, *Complete (0.8M)*. In case of the mixture of NMT experts, the gating network was trained using the one million sentences.

Then, we tested in the incomplete data situation. Here there were just 200,000 complete multilingual sentences (sentence No. 600,001-800,000), *Complete (0.2M)*. Here, a standard multi-encoder NMT and mixture of NMT experts could be trained using this complete data. On the other hand, the multi-source NMT with <NULL> could be trained using 800,000 sentences (sentence No. 1-800,000), *Pseudo-incomplete (0.8M)*. Each one-to-one NMT could be trained using these 800,000 sentences, but the missing sentences replaced with the <NULL> tokens were excluded so resulting 600,000 sentences were actually used.

4.2.3 Results

Table 2 shows the results in BLEU. The multi-source approaches achieved consistent improvements over the one-to-one NMTs in the all conditions, as demonstrated in previous multi-source NMT studies. Our main focus here is Pseudo-incomplete (0.8M), in which the multi-source results were slightly worse than those in Complete (0.8M) but better than those in Complete (0.2M). This suggests the additional use of incomplete corpora is beneficial in multi-source NMT compared to the use of only the complete parts of the cor-

Condition	One-to-one			Multi-encoder	Mix. NMT Experts
	Es-En	Fr-En	Ar-En		
Complete (0.8M)	31.87	25.78	23.08	37.55 (+5.68)*	33.28 (+1.41)
Complete (0.2M)	27.62	22.01	17.88	31.24 (+3.62)	32.16 (+4.54)
Pseudo-incomplete (0.8M)	30.98	25.62	22.02	**36.43** (+5.45)*	32.44 (+1.47)

Table 2: Results in BLEU for one-to-one and multi-source ({Es, Fr, Ar}-to-En) translation on UN6WAY data (parentheses are BLEU gains against the best one-to-one results). * indicates the difference from mixture of NMT experts is statistically significant ($p < 0.01$).

Source	Training	Valid.	Test
{En, Fr, Pt (br)}-to-Es			
English	189,062	4,076	5,451
French	170,607	3,719	4,686
Portuguese (br)	166,205	3,623	4,647
{En, Es, Pt (br)}-to-Fr			
English	185,405	4,164	4,753
Spanish	170,607	3,719	4,686
Portuguese (br)	164,630	3,668	4,289
{En, Es, Fr}-to-Pt (br)			
English	177,895	3,880	4,742
Spanish	166,205	3,623	4,647
French	164,630	3,668	4,289

Table 3: Data statistics in the tasks on TED data (in the number of sentences). Note that the number of target sentences is equal to that of English for each task.

Target	Training	Valid.	Test
Spanish	83.4	85.0	78.2
French	85.0	83.2	89.7
Portuguese (br)	88.6	89.3	90.0

Table 4: The percentage of data without missing sentences on TED data.

pus, even if just through the simple modification of replacing missing sentences with <NULL>.

With respect to the difference between the multi-encoder NMT and mixture of NMT experts, the multi-encoder achieved much higher BLEU in Pseudo-incomplete (0.8M) and Complete (0.8M), but this was not the case in Complete (0.2M). One possible reason here is the model complexity; the multi-encoder NMT uses a large single model while one-to-one sub-models in the mixture of NMT experts can be trained independently.

4.3 An actual incomplete multilingual corpus (TED Talks)

4.3.1 Data

We used a collection of transcriptions of TED Talks and their multilingual translations. Because these translations are created by volunteers, and the number of translations for each language is dependent on the number of volunteers who created them, this collection is an actual incomplete multilingual corpus. The great majority of the talks are basically in English, so we chose English as a source language. We used three translations in other languages for our multi-source scenario: Spanish, French, Brazilian Portuguese. We prepared three tasks choosing one of each of these three languages as the target language and the others as the additional source languages. Table 3 shows the number of available sentences in these tasks, chosen so that their lengths are less than or equal to 40 words.

4.3.2 Setup

We compared multi-encoder NMT, mixture of NMT experts and one-to-one NMT with English as the source language. The validation and test data for these experiments were also incomplete. This is in contrast to the experiments on UN6WAY where the test and validation data were complete, and thus this setting is arguable of more practical use.

4.3.3 Results

Table 5 shows the results in BLEU and BLEU gains with respect to the one-to-one results. All the differences are statistically significant ($p < 0.01$) by significance tests with bootstrap resampling (Koehn, 2004). The multi-source NMTs achieved consistent improvements over the one-to-one baseline as expected, but the BLEU gains were smaller than those in the previous experiments using the UN6WAY data. This is possibly

because the baseline performance was relatively low compared with the previous experiments and the size of available resources was also smaller.

In comparison between the multi-source NMT and the mixture of NMT experts, results were mixed; the mixture of NMT experts was better in the task to French.

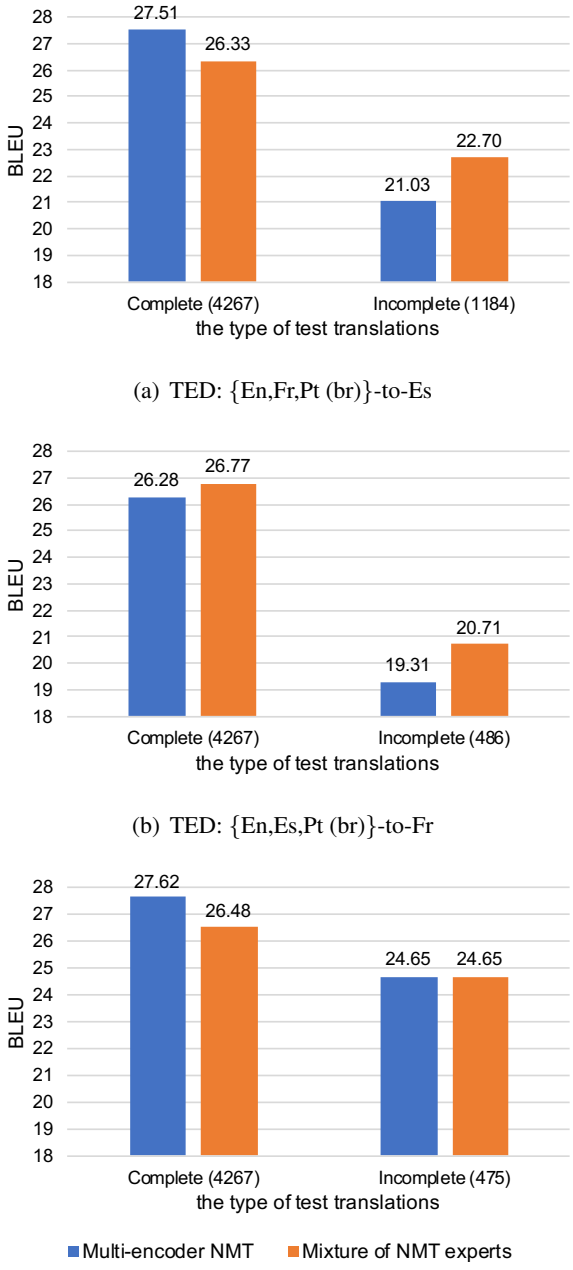
4.3.4 Discussion

We analyzed the results using the TED data in detail to investigate the mixed results above. Figure 5 shows the breakdown of BLEU in the test data, separating the results for complete and incomplete multilingual inputs. When all source sentences are present in the test data, multi-encoder NMT has better performance than mixture of NMT experts except for {En, Es, Pt (br)}-to-Fr. However, when the input is incomplete, mixture of NMT experts achieves performance better than or equal to multi-encoder NMT. From this result, we can assume that mixture of NMT experts, with its explicit gating network, is better at ignoring the irrelevant missing sentences. It's possible that if we designed a better attention strategy for multi-encoder NMT we may be able to resolve this problem. These analyses would support the results using the pseudo incomplete data shown in Table 2, where the validation and test data were complete.

4.3.5 Translation examples

Table 6 shows a couple of translation examples in the {English, French, Brazilian Portuguese}-to-Spanish experiment. In Example(1), BLEU+1 of mixture of NMT Experts is larger than one-to-one (English-to-Spanish) because of the French sentence, although the source sentence of Brazilian Portuguese is missing. BLEU+1 of multi-encoder is same as one-to-one, but the generation word is different. The word of "minar" is generated from multi-encoder, and "estudiar" is generated from one-to-one. "minar" means "look" in English, and "estudiar" means "study", so the meaning of sentence which was generated from multi-encoder is close to the reference one than that from one-to-one. Besides the word of "ver" which is generated from mixture of NMT experts meas "see" in English, so the sentence of multi-encoder is more appropriate than the reference sentence.

In Example(2), there is only the English sentence in the source sentences. We can see that sentences which are generated from all models are same as the reference sentences, although French and Brazilian Portuguese sentences are missing. Therefore multi-source NMT models work properly even if there are missing sentences.



(a) TED: {En,Fr,Pt (br)}-to-Es

(b) TED: {En,Es,Pt (br)}-to-Fr

(c) TED: {En,Es,Fr}-to-Pt (br)

Figure 5: Detailed comparison of BLEU in TED test data. *Complete* means the part of test data, in which there is no missing translation, and *incomplete* means that, in which there are some missing translation. The number in a parenthesis is the number of translations.

Task	One-to-one (En-to-target)	Multi-encoder	Mix. NMT Experts
{En, Fr, Pt (br)}-to-Es	24.32	**26.01** (+1.69)	25.51 (+1.19)
{En, Es, Pt (br)}-to-Fr	24.54	25.62 (+1.08)	**26.23** (+1.69)
{En, Es, Fr}-to-Pt (br)	25.14	**27.36** (+2.22)	26.39 (+1.25)

Table 5: Results in BLEU (and BLEU gains) by one-to-one and multi-source NMT on TED data. Note that the target language in each row differs so the results in different rows cannot be compared directly. All the differences are statistically significant ($p < 0.01$).

Type	Sentence	BLEU+1
Example (1)		
Source (En)	Then I started looking at the business model.	
Source (Fr)	Puis j'ai regard le modle conomique.	
Source (Pt (br))	<NULL>	
Reference	Despus empec a ver el modelo de negocio.	
En-to-Es	Luego empec a estudiar el modelo empresarial.	0.266
Multi-encoder	Luego empec a mirar el modelo empresarial.	0.266
Mix. NMT experts	Luego empec a ver el modelo de negocios.	0.726
Example (2)		
Source (En)	Sometimes they agree.	
Source (Fr)	<NULL>	
Source (Pt (br))	<NULL>	
Reference	A veces estn de acuerdo.	
En-to-Es	A veces estn de acuerdo.	1.000
Multi-encoder	A veces estn de acuerdo.	1.000
Mix. NMT experts	A veces estn de acuerdo.	1.000

Table 6: Translation examples in {English, French, Brazilian Portuguese}-to-Spanish translation.

5 Related Work

In this paper, we examined strategies for multi-source NMT. On the other hand, there are other strategies for multilingual NMT that do not use multiple source sentences as their input. Dong et al. (2015) proposed a method for multi-target NMT. Their method is using one sharing encoder and decoders corresponding to the number of target languages. Firat et al. (2016) proposed a method for multi-source multi-target NMT using multiple encoders and decoders with a shared attention mechanism. Johonson et al. (2017) and Ha et al. (2016) proposed multi-source and multi-target NMT using one encoder and one decoder, and sharing all parameters with all languages. Notably, these methods use multilingual data to better train one-to-one NMT systems. However, our motivation of this study is to improve NMT further by the help of other translations that are available on the source side at test time, and thus their approaches are different from ours.

6 Conclusion

In this paper, we examined approaches for multi-source NMT using *incomplete* multilingual corpus in which each missing input sentences is replaced by a special symbol <NULL>. The experimental results with simulated and actual incomplete multilingual corpora show that this simple modification allows us to effectively use all available translations at both training and test time.

The performance of multi-source NMT depends on source and target languages, and the size of missing data. As future work, we will investigate the relation of the languages included in the multiple sources and the number of missing inputs to the translation accuracy in multi-source scenarios.

7 Acknowledgement

Part of this work was supported by JSPS KAKENHI Grant Numbers and JP16H05873 and JP17H06101.

References

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

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th EAMT Conference*, pages 261–268.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-Task Learning for Multiple Language Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Ekaterina Garmash and Christof Monz. 2016. Ensemble Learning for Multi-Source Neural Machine Translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation*, Seattle, Washington.

Melvin Johonson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vigas, 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 Linguistics, vol. 5*, pages 339–351.

Diederik P. Kingma Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2001. Statistical Multi-Source Translation. In *Proceedings of the eighth Machine Translation Summit (MT Summit VIII)*, pages 253–258.

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

Jörg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Micha Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Barret Zoph and Kevin Knight. 2016. Multi-Source Neural Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.