# Multi-Source Neural Machine Translation with Missing Data

Yuta Nishimura, *Nonmember, IEEE*, Katsuhito Sudoh, *Nonmember, IEEE*, Graham Neubig, *Nonmember, IEEE* and Satoshi Nakamura, *Fellow, IEEE*

***Abstract*—Machine translation is rife with ambiguities in word ordering and word choice, and even with the advent of machine-learning methods that learn to resolve this ambiguity based on statistics from large corpora, mistakes are frequent. *Multi-source translation* is an approach that attempts to resolve these ambiguities by exploiting multiple inputs (e.g. sentences in three different languages) to increase translation accuracy. These methods are trained on multilingual corpora, which include the multiple source languages and the target language, and then at test time uses information from both source languages while generating the target. While there are many of these multilingual corpora, such as multilingual translations of TED talks or European parliament proceedings, in practice, many multilingual corpora are not complete due to the difficulty to provide translations in *all* of the relevant languages. Existing studies on multi-source translation did not explicitly handle such situations, and thus are only applicable to complete corpora that have all of the languages of interest, severely limiting their practical applicability. In this paper, we examine approaches for multi-source neural machine translation (NMT) that can learn from and translate such incomplete corpora. Specifically, we propose methods to deal with incomplete corpora at both training time and test time. For training time, we examine two methods: (1) a simple method that simply replaces missing source translations with a special NULL symbol, and (2) a data augmentation approach that fills in incomplete parts with source translations created from multi-source NMT. For test-time, we examine methods that use multi-source translation even when only a single source is provided by first translating into an additional auxiliary language using standard NMT, then using multi-source translation on the original source and this generated auxiliary language sentence. Extensive experiments demonstrate that the proposed training-time and test-time methods both significantly improve translation performance.**

***Index Terms*—Neural machine translation, multilinguality, data augmentation.**

## I. Introduction

ONE of the major challenges in Machine Translation (MT) systems is the inherent ambiguity in translating across languages, and recent methods perform machine learning over large corpora to learn models to resolve this ambiguity. However, in many real situations, it is difficult to create large corpora, for a particular language of interest, and thus translation accuracy may suffer. One way to deal with this paucity of data is to use data from multiple languages to improve the translation accuracy [1], [2], [3], [4]. There are a significant number of multilingual document collections that can be used for this purpose, such as the proceedings of the European parliament [5] or the United Nations [6]. Documents in these corpora are manually translated into all official languages of the respective organizations. There are also voluntary translation effort multilingual captions such as those for talks [7] and movies [8]. In addition, OPUS (http://opus.nlpl.eu) provides many parallel corpora, and a large number of them contain multi-way translations: EMEA, software localizations (GNOME, PHP, etc.), the Bible, OpenSubtitles, and so on. However, in the case of such voluntary translation efforts, large portions of the corpus are not translated, especially into languages with a relatively small number of speakers.

In this paper, we focus on this sort of multilingual scenario, specifically using multi-source translation [9], [10], [11]. Multi-source translation is an approach to exploit multiple inputs (e.g. in two different languages) to decrease ambiguity and increase translation accuracy. Specifically, in the context of neural machine translation (NMT), there are several methods proposed to do so. For example, Zoph and Knight [10] proposed a method where multiple sentences are encoded and then passed into a single decoding process (the "multi-encoder" method). In addition, Garmash and Monz [11] proposed a method where NMT systems over multiple inputs are ensembled together to make a final prediction (the "mixture-of-NMT-experts" method).

These paradigms generally assume, and have been tested on the case where we have data in *all* of the source languages. However, it is unusual that translations in all of these languages are provided – other than parliament proceedings where it is legally mandated that translations in all languages be generated, most multi-lingual corpora have gaps where translators have not been able to generate translations. This paper focuses on *multi-source NMT with missing data*, proposing methods to leverage multiple languages to improve accuracy, while still maintaining the ability to train and test on corpora that are not complete.

First, we conducted a simple implementation of multi-source NMT using such an incomplete multilingual corpus by using a special symbol (arbitrarily, __NULL__) to represent the missing sentences. This makes it possible to both train and test using existing multi-source NMT implementations without any modifications by simply appending this symbol in place of missing sentences. At test time, we can expect the system to ignore the __NULL__ symbol, with the decoder choosing hypotheses using other input sentences. Experimental results with a real incomplete multilingual corpus of TED Talks show

Y. Nishimura, K. Sudoh and S. Nakamura are with Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan e-mail: (nishimura.yuta.nn9@is.naist.jp).

G. Neubig is with Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA.

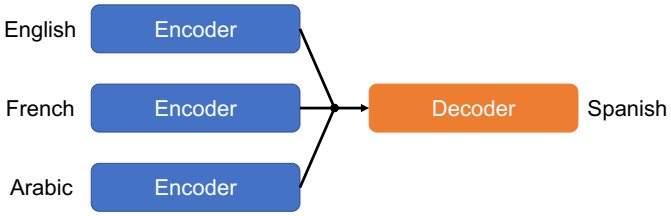



Fig. 1: Multi-encoder NMT

that the proposed method is effective for allowing training for multi-source NMT in situations where full multilingual corpora are not available.

However, this first method essentially throws away the information provided by the utilization of multiple inputs when such inputs are not available. To alleviate this problem, we propose a second method that performs data augmentation at training time, creating a pseudo-parallel corpus that fills in the missing sentences with an output of a multi-encoder NMT system. Experimental results show that the proposed method is more effective than simply filling up the sentences with __NULL__.

Finally, we turn our eyes to methods that can utilize multi-source translation at *test* time. The __NULL__ augmentation method can technically be applied at test time but essentially reduces to single-source translation, and the pseudo-parallel corpus augmentation method cannot be used at test time because we do not know the gold reference translation. As an alternative, we propose a two-step method for multi-encoder NMT decoding motivated by pivot machine translation. Pivot machine translation uses an pivot language that has a large amount of data to get better translation accuracy on a language pair which has a small amount of data [12]. In other words, pivot machine translation gets better translation accuracy using a pivot language that is easier to machine-translate than the language pair that we actually want to translate. In multi-encoder NMT, we can consider the language in which we have missing translations as a pivot language, and create automatic translations. Specifically, our method creates multiple hypotheses in the missing language using normal one-to-one NMT, and chooses the appropriate hypothesis for multi-encoder NMT. We confirm the effectiveness of this proposed method in experiments over a test set which is missing one of the two sources.[1]

## II. Related Work

### A. Multi-Source NMT

At the present, there are two major approaches to multi-source NMT: multi-encoder NMT [10] and mixture of NMT experts [11]. We first review them in this section.

*1) Multi-Encoder NMT:* Multi-encoder NMT [10] is similar to the standard attentional NMT framework [15] but uses multiple encoders corresponding to the source languages and a single decoder, as shown in Figure 1.

[1]This journal article is based on content previously presented as two workshop proceedings [13], [14]. We have greatly expanded the explanation, and this article adds entirely new content related to test-time usage of multi-source methods, corresponding to the Sections IV and VI.

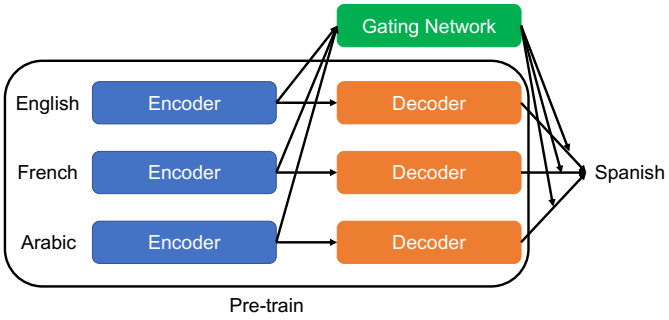



Fig. 2: Mixture of NMT Experts

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:

$$h = \tanh(W_c[h_1; h_2]) \tag{1}$$

$$c = c_1 + c_2 \tag{2}$$

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

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

Our implementation is based on Zoph and Knight [10], but we use global attention while they used local-p attention. *local-p* attention focuses only on a small subset of the source positions for each target word [16], while *global* attention attends to all words on the source side for each target word. We use global attention as it is the standard method used in the great majority of recent NMT work.

*2) Mixture of NMT Experts:* Garmash and Monz [11] 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 these independent models, as shown in Figure 2.

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

$$p_t = \sum_{j=1}^{m} g_t^j p_t^j \tag{4}$$

where $p_t^j$ and $g_t^j$ are the probability vector from the $j$-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:

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

where $f_t^j(x)$ is the input vector to the decoder of the $j$-th model. In this work, $x_t^j$ is the input vector at the time step $t$, which concatenates the embedding vector at the time step $t$ and the context vector of the $j$-th model at the previous time step $t$-1.

### B. Data Augmentation

Sennrich *et al.* proposed a method to use monolingual training data in the target language for training NMT systems, with no changes to the network architecture [17]. This method first trains a seed target-to-source NMT model using a parallel corpus and then translates the monolingual target language sentences into the source language to create a *synthetic* parallel corpus. It finally trains a source-to-target NMT model using the seed and synthetic parallel corpora. This simple method called *back-translation* makes effective use of available resources, and achieves substantial gains in accuracy. Imamura *et al.* proposed a method that generates multiple source sentences via sampling as an extension of the back-translation [18]. Firat *et al.* proposed a data augmentation framework for zero-shot translation that uses pseudo source language sentences translated from a pivot language [19].

There are also other approaches for data augmentation other than back-translation. Wang *et al.* proposed a method of randomly replacing words in both the source sentence and the target sentence with other random words from their corresponding vocabularies [20]. Kim and Rush proposed a sequence-level knowledge distillation in NMT that uses machine translation results by a large teacher model to train a small student model as well as ground-truth translations [21].

Our work is similar to the back-translation approach, but specifically tailored to multilingual and multi-source situations.

## III. Training-time Data Augmentation for Multi-source Translation

As mentioned in the introduction, multi-source NMT 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 English, Spanish, and French. However, it is unusual that translations in all of these languages are provided, particularly for all of the training data that we have available; there will usually be many sentences where we have only one of the sources. In this section, we propose two methods for dealing with this method at training time: *NULL augmentation* by just replacing missing sentences with a special symbol and *pseudo-translation augmentation* by filling machine translation outputs there.

### A. NULL Augmentation

The first proposed method is very simple; we just replace each missing input sentence with a special symbol. We refer to this special symbol as __NULL__, although the choice of this character string is arbitrary. By training the model using this method, we expect it to learn to ignore the __NULL__ symbol in multi-encoder NMT when it appears, and translate using the other translations. This method can be applied easily to any existing implementation of multi-encoder NMT without modification. Figure 3 provides a concrete example of this method. Here we suppose the source languages are Spanish, French, and Arabic and the target language is English, and that the Arabic input sentence is missing. In this case, the Spanish and French input sentences are passed into the corresponding encoders, and a special symbol __NULL__ is passed to the Arabic encoder. In the experiments described later, we applied this method for both multi-encoder NMT and mixture of NMT experts and compared them.

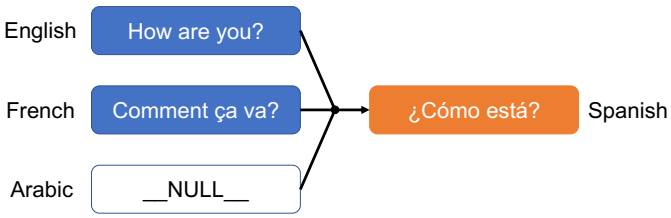



Fig. 3: Multi-encoder NMT with a missing input sentence. In this example, the Arabic input sentence is missing.

### B. Pseudo-translation Augmentation

*1) Motivation and overall framework:* While the first proposed method is simple, if the model is trained on corpora with a large number of __NULL__ symbols on the source side, a large number of training examples will be different from test time when we actually have multiple sources. Thus, these examples will presumably be less useful in training a model intended to do multi-source translation. We thus additionally propose an improved method for utilizing multi-source examples with missing data: using a pseudo-corpus whose missing translations are augmented with machine translation outputs using a trained multi-source NMT system as shown in Figure 3.

Here we use three languages to explain our proposed method; English, Spanish and French. Our goal is to generate a Spanish translation, and we suppose there are not any missing data on the English side, but Spanish and French translations have some missing data.

*2) Step-wise Augmentation Procedure:* Our augmentation procedure consists of the following three steps. The procedure is illustrated in Fig. 4

1) Train a multi-encoder NMT model (Source: English and Spanish, Target: French) to get French automatic translations using the baseline method, that replaces missing input sentences with a special symbol __NULL__.
2) Create French automatic translations using multi-encoder NMT, which was trained on the data generated in the first step. We conduct three types of augmentation in this step, the details of which we introduce later.
3) Switch the role of French and Spanish, in other words, we train a new multi-encoder NMT model (Source: English and French, Target: Spanish). At this time, we use French automatic translations on the source language side.

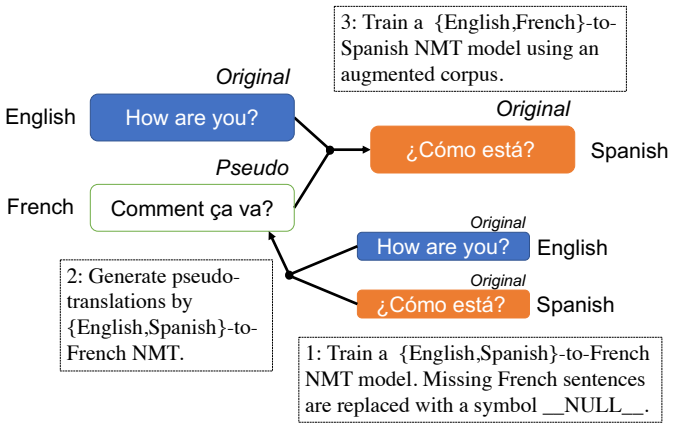



Fig. 4: Example of multi-encoder NMT with data augmentation using an incomplete corpus. The language pair is English, French-to-Spanish and the translation of French is missing.

This method is similar to back-translation but takes advantage of the fact that we have an additional source of knowledge (Spanish or French) when trying to augment the other language (French or Spanish respectively).

*3) Filling strategies:* We propose three specific types of data augmentation for multi-encoder NMT; "fill-in", "fill-in and replace" and "fill-in and add." Figure 5 illustrates examples in the {English, French}-to-{Spanish} case where one Spanish sentence is missing.

(a) **fill-in:** where only missing parts in the corpus are filled up with pseudo-translations.

(b) **fill-in and replace:** where we both augment the missing parts and replace original translations with automatic translations in the source languages other than English (as English has no missing data). This method has the potential to override wrongly-aligned or non-literal translations existing in the original data similarly to work by Morishita et al. [22], which demonstrated the effectiveness of applying back-translation for an unreliable part of a provided corpus. Translations of TED talks are from many independent volunteers, so there may be some differences between translations other than original English, or they may even include some free or over-simplified translations. We aim to fill such a gap using data augmentation.

(c) **fill-in and add:** where we both augment the missing sentences and add automatic translations to sentences that already have original translations in the source language other than English. This helps prevent introduction of too much noise due to the complete replacement of original translations with automatic translations in the second method.

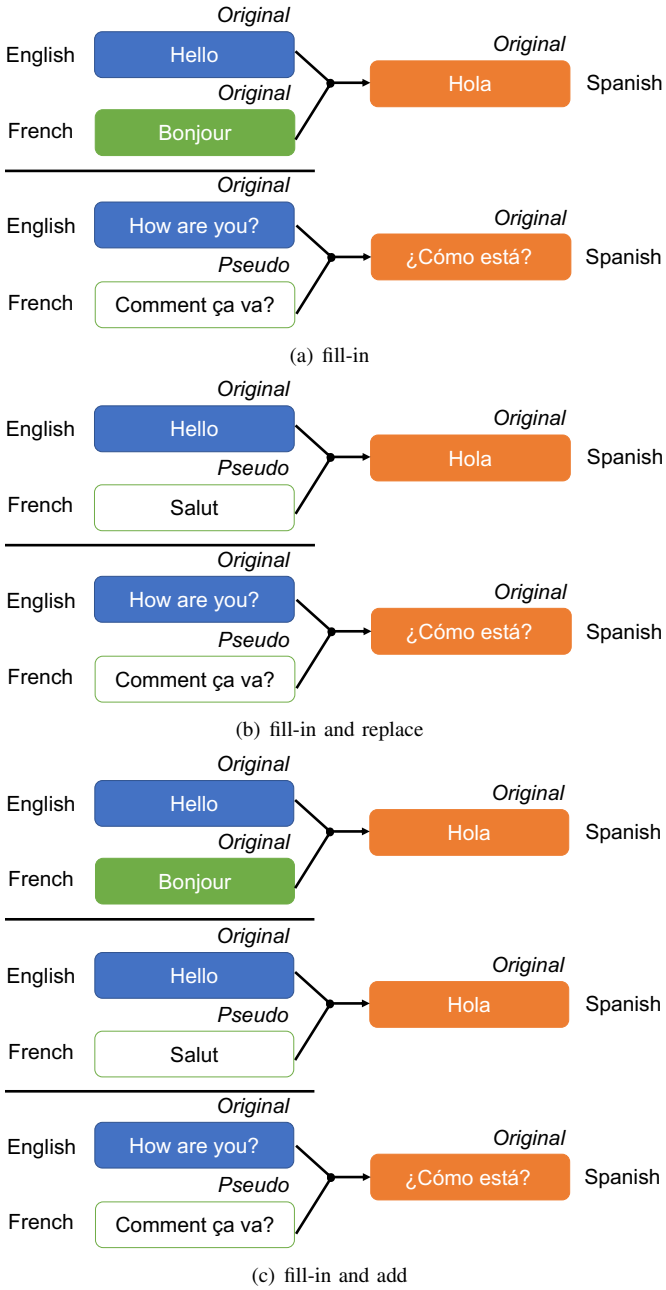



Fig. 5: Example of three types of augmentation. The language Pair is {English, French}-to-{Spanish} and French translation corresponding to "How are you?" is missing. In this example, the white background indicates the pseudo-translation produced from multi-source NMT and the colored background means the original translation.

## IV. Test-time Data Augmentation for Multi-source Translation

Our previous two proposed methods provide a better way to utilize missing multilingual corpora at training time, but they cannot be used trivially if there are missing source sentences at test time. This is because the first proposed method can be applied at testing time but simply ignores the missing sentence, reverting back to one-to-one translation (or worse), and the second proposed method uses translations of a target language when filling up the missing source translations, and these target translations are not available at test time. Thus we propose the third method to deal with an incomplete corpus at test time, specifically a multi-target machine translation method, which generates multiple languages at the same time to increase translation accuracy [23]. Our specific approach is motivated

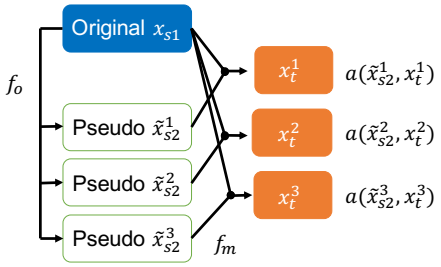



Fig. 6: Example of considering n-best source translations

by work on pivot machine translation, which uses a pivot language that has a large amount of data to get better translation accuracy on language pairs that have a small amount of data [12]. In other words, pivot translation gets better translation accuracy using a pivot language where machine translation is easier than the language pair in which we actually want to do machine translation. In multi-encoder NMT, we can consider a language which has a missing translation as a pivot language, and create an automatic translation for this language. We propose the method to create multiple source translations of a missing language using normal one-to-one NMT, and choose an appropriate one from multiple translation hypotheses for multi-encoder NMT, as shown in Fig 6. Below, we describe the proposed method in detail.

We assume we have a multi-encoder NMT system $f_m$ whose source language sentences are $x_{s1}$ and $x_{s2}$, and a target language sentence is $x_t$. $x_{s2}$ is a missing sentence. We additionally assume a trained one-to-one NMT system $f_o$ whose source language is $s_1$ and target language is $s_2$. Then we output n-best translations $\tilde{X}_{s2} = \{\tilde{x}^1_{s2}, ..., \tilde{x}^n_{s2}\}$ using the trained one-to-one NMT with beam search to fill in the missing sentence $x_{s2}$. The sentence probabilities of these n-best translations are $P_s = \{p(\tilde{x}^1_{s2}|f_o, x_{s1}), ..., p(\tilde{x}^n_{s2}|f_o, x_{s2})\}$. Then we output multi-encoder NMT translations $X_t = \{x^1_t, ..., x^n_t\}$ using n-best translations into the additional source sentence. The sentence probabilities using n-best source translations are $P_t = \{p(x^1_t|f_m, x_{s1}, \tilde{x}^1_{s2}), ..., p(x^n_t|f_m, x_{s1}, \tilde{x}^n_{s2})\}$. Finally, we choose the most appropriate translation $x_t$ defined by the following equations.

$$a(\tilde{x}^n_{s2}, x^n_t) = \lambda \ln p(\tilde{x}^n_{s2}|f_o, x_{s1}) + (1-\lambda) \ln p(x^n_t|f_m, x_{s1}, \tilde{x}^n_{s2}) \quad (6)$$

$$\tilde{x_{s2}}, x_t = \underset{x^n_{s2}, x^n_t \in X_{s2}, X_t}{\arg\max} a(\tilde{x}^n_{s2}, x^n_t) \quad (7)$$

$\lambda$ in the equation 6 is a hyperparameter that we use to consider which probability between the source translation and the multi-encoder translation is more important.

## V. Training-time Data Augmentation Experiments

### A. Experiment Outline

In this section, we examine the efficacy of our methods for training-time data augmentation, including both "NULL augmentation" (Section III-A) and "pseudo-corpus augmentation" (Section III-B). We first validate our general approach of utilizing missing data in multi-source NMT training, focusing on the comparison between NULL augmentation and other baseline methods. Then we move on to compare pseudo-corpus augmentation with NULL augmentation and other natural baselines focusing on data augmentation.

### B. NMT settings

NMT[2] settings are the same for all the methods in the experiments. We use bidirectional LSTM encoders [15] and global attention with input feeding for the NMT model [16]. The number of dimensions is set to 512 for the hidden and embedding layers. Subword segmentation was applied using SentencePiece [24]. We trained one subword segmentation model for the training corpus that concatenated all of the languages in a language set. For parameter optimization, we used Adam [25] with gradient clipping of 5. The number of hidden state units in the gating network for the mixture of NMT experts experiments was 256. We used BLEU [26] as the evaluation metric and SacreBLEU[3] [27] as the evaluation tool. We performed early stopping, saving parameter values that had the best log likelihoods on the validation data and used them when decoding the test data. However, in the experiments with mixture of NMT experts, we found this likely to result in over-fitting, so we train our models for only 1 epoch and saved parameter values.

### C. Experiment: NULL Augmentation on a 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. The purpose of creating a pseudo-incomplete corpus is to extensively compare the performance in complete and incomplete situations.

*1) Data:* We used UN6WAY [6] 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 800,000 sentences from the UN6WAY corpus whose sentence lengths were less than 40 words. We excluded 200,000 sentences for each language except English for the pseudo-incomplete multilingual corpus as shown in Table I. "Sentence No." in Table I 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.

*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 sub-models in the mixture of NMT experts, were also compared for reference.

First, we conducted experiments using all of the 800,000 sentences in the complete multilingual corpus, *Complete*

[2]We used pytorch as a neural network toolkit. https://github.com/pytorch/pytorch.git

[3]https://github.com/awslabs/sockeye/tree/master/sockeye_contrib/sacrebleu

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

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

*(0.8M)*. In case of the mixture of NMT experts, the gating network was trained using the 800,000 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 system 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.

*3) Results:* From results in Table II, we can see that the multi-source approaches achieved consistent improvements over the one-to-one NMT models in the all conditions. This is in concert with 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 corpus, even if just through the simple modification of replacing missing sentences with __NULL__.

With respect to the difference between multi-encoder NMT and mixture of NMT experts, multi-encoder NMT achieved much higher BLEU in all conditions. One possible reason here is the model complexity; the multi-encoder NMT model uses a large single model while one-to-one sub-models in the mixture of NMT experts can be trained independently.

### *D. Experiment: NULL Augmentation on a Real Incomplete Corpus (TED Talks)*

In addition to the experiments with a pseudo-incomplete multilingual corpus, we also checked our proposed method with an actual incomplete multilingual corpus.

*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 originally in English, so we chose English as a source language. We used three translations in other languages for our multi-source scenario: Spanish, French and Brazilian Portuguese. We prepared three tasks choosing one of these three languages as the target language and the others as the additional source languages. Table III shows the number of available sentences in these tasks, chosen so that their lengths are less than 40 words, and Table IV shows the percentage of available sentences compared to the number of English sentences in these tasks.

*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 arguably of more practical use.

*3) Results:* Table V shows the results in BLEU and BLEU gains with respect to the one-to-one results. The multi-source NMT systems 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.

We analyzed the results using the TED data in detail to investigate the mixed results above. Figure 7 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 convincingly better performance than mixture of NMT experts. However, when the input is incomplete, mixture of NMT experts achieves performance better except on {En,Es,Pt(br)}-to-Fr. From this result, we can assume that mixture of NMT experts, with its explicit gating network, is better at ignoring the irrelevant missing sentences. This indicates the need for a better test-time strategy than NULL augmentation such as the one that we propose in Section IV and evaluate in Section VI.

*4) Translation examples:* Table VI shows a couple of translation examples in the {English, Spanish, French}-to-Brazilian Portuguese experiment.

In Example (1), there is only the English sentence in the source sentences. We can see that sentences which are generated from all models are the same as the reference sentences, even though French and Brazilian Portuguese sentences are missing. Therefore multi-source NMT models work properly even if there are missing sentences.

In Example (2), BLEU+1 of mixture of NMT experts and multi-encoder NMT are larger than one-to-one (English-to-Brazilian Portuguese) because of the Spanish sentence, even though the source sentence of French is missing. In mixture of NMT experts and one-to-one NMT output “Bem,” means “Well,” in English, but this was not output when using multi-encoder NMT. Furthermore, the Spanish source sentence doesn’t have a word which means “Well,”. It is assumed that the multi-encoder NMT could use information of the other languages more effectively than mixture of NMT experts.

### *E. Experiment: Augmentation with Pseudo-translations*

Finally, we examine the efficacy of our Pseudo-translation Augmentation method.

*1) Data:* Similarly to above, we used a collection of transcriptions of TED Talks and their multilingual translations. The numbers of these voluntary translations differs significantly by language. We chose three different language sets for

TABLE II: 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).

| Condition | One-to-one | | | Multi-encoder | Mix. NMT Experts |
|---|---|---|---|---|---|
| | Es-En | Fr-En | Ar-En | | |
| Complete (0.8M) | 51.11 | 47.18 | 43.29 | 59.45 (+8.34) | 53.16 (+2.05) |
| Complete (0.2M) | 48.92 | 44.82 | 39.46 | 55.87 (+6.95) | 50.76 (+1.84) |
| *Pseudo-incomplete* (0.8M) | 51.02 | 46.63 | 43.43 | **57.94** (+6.92) | 52.09 (+1.07) |

TABLE III: 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.

| Source | Training | Valid. | Test |
|---|---|---|---|
| {En, Fr, Pt(br)}-to-Es | | | |
| English | 179,784 | 3,880 | 5,200 |
| French | 162,764 | 3,547 | 4,484 |
| Brazilian Portuguese | 158,807 | 3,451 | 4,459 |
| {En, Es, Pt(br)}-to-Fr | | | |
| English | 176,473 | 3,972 | 4,546 |
| Spanish | 162,764 | 3,547 | 4,484 |
| Brazilian Portuguese | 157,384 | 3,490 | 4,113 |
| {En, Es, Fr}-to-Pt(br) | | | |
| English | 169,791 | 3,689 | 4,544 |
| Spanish | 158,807 | 3,451 | 4,459 |
| French | 157,384 | 3,490 | 4,113 |

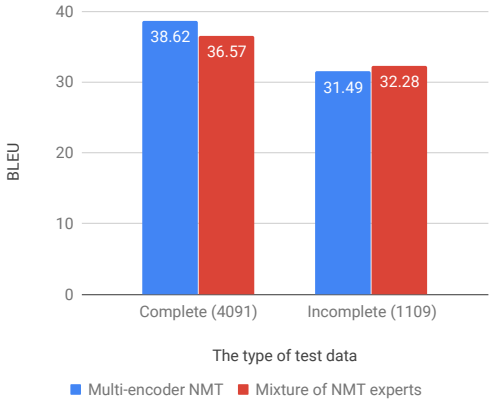
TABLE IV: The percentage of sentences without missing sentences on TED data.

| Target | Training | Valid. | Test |
|---|---|---|---|
| Spanish | 83.83 | 85.08 | 78.67 |
| French | 85.40 | 83.11 | 89.99 |
| Brazilian Portuguese | 88.77 | 89.48 | 90.03 |

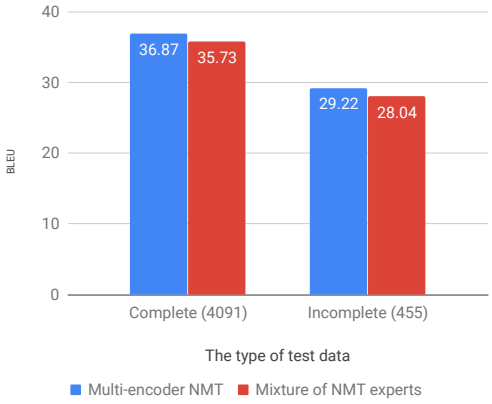
TABLE V: Results in BLEU (and BLEU gains) by one-to-one and multi-source NMT on TED data. Note that the target language in each row differs so the results in different rows cannot be compared directly.

| Task | One-to-one (En-to-Trg) | Multi-encoder | Mix. NMT Experts |
|---|---|---|---|
| {En, Fr, Pt(br)}-to-Es | 33.93 | **37.21** (+3.28) | 35.62 (+1.69) |
| {En, Es, Pt(br)}-to-Fr | 33.47 | **36.19** (+2.72) | 35.07 (+1.60) |
| {En, Es, Fr}-to-Pt(br) | 35.31 | **38.79** (+3.48) | 36.87 (+1.56) |

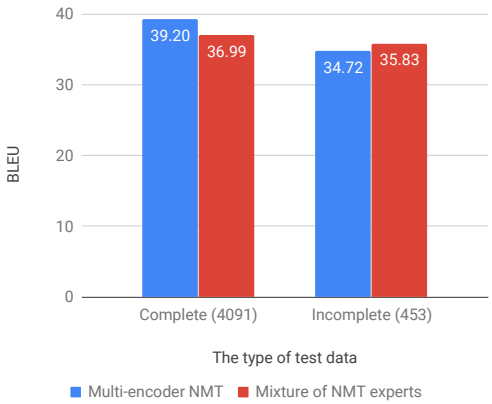
the experiments: {English (en), Croatian (hr), Serbian (sr)}, {English (en), Slovak (sk), Czech (cs)}, and {English (en), Vietnamese (vi), Indonesian (id)}. The data in the experiments was chosen from the TED Talks corpus, and we limited to sentences for which lengths were less than 40 words. The experiments were designed for the translation from English to another language with the help of the other language in the language set, because the great majority of TED talks are in English, there are not any missing portions in the English sentences. Table VII shows the number of training sentences for each language set. At test time, we conducted the experiment with a complete corpus where both source sentences are represented, as this method cannot easily handle sentences where one of the sources is missing at test time.




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




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




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

Fig. 7: Detailed comparison of BLEU on the TED test data. *Complete* indicates the part of test data in which there is no missing translation, and *Incomplete* indicates the part in which there are some missing translation. The number in parentheses is the number of translations.

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

| Type | Sentence | BLEU+1 |
|---|---|---|
| Example (1) | | |
| Source (En) | Here's one of my favorite pictures of Mars. | |
| Source (Es) | __NULL__ | |
| Source (Fr) | __NULL__ | |
| Reference | Eis uma de minhas fotos favoritas de Marte. | |
| En-to-Pt(br) | Aqui est uma das minhas fotos favoritas de Marte. | 0.54 |
| Multi-encoder | Aqui est uma das minhas fotos favoritas de Marte. | 0.54 |
| Mix. NMT experts | Aqui est uma das minhas fotos favoritas de Marte. | 0.54 |
| Example (2) | | |
| Source (En) | Well, it means everything's fine until this happens. | |
| Source (Es) | Significa que todo est bien hasta que sucede esto. | |
| Source (Fr) | __NULL__ | |
| Reference | Significa que tudo vai bem at que isso acontea. | |
| En-to-Pt(br) | Bem, significa tudo isso tudo bem at isso acontecer. | 0.21 |
| Multi-encoder | Significa que tudo est bem at que isso acontece. | 0.52 |
| Mix. NMT experts | Bem, significa que tudo est bem at que isso acontece. | 0.40 |

TABLE VII: "train" shows the number of available training sentences, and "missing" shows the number and the fraction of missing sentences in comparison with English ones.

| Pair | Trg | train | missing |
|---|---|---|---|
| en-hr/sr | hr | 115,127 | 34,116 (29.6%) |
| | sr | 129,461 | 48,450 (37.4%) |
| en-sk/cs | sk | 58,109 | 16,772 (28.9%) |
| | cs | 97,488 | 56,151 (57.6%) |
| en-vi/id | vi | 150,829 | 81,945 (54.3%) |
| | id | 77,936 | 9,052 (11.6%) |

*2) Baseline Methods:* We compared the proposed methods with the following three baseline methods.

**One-to-one NMT** A standard NMT model from one source language to another target language. The source language is English or the remaining language in the language set. If the target language part is missing in the parallel corpus, such sentences pairs cannot be used in training so they are excluded from the training set.

**Multi-encoder NMT with back-translation** A multi-encoder NMT system using English-to-X NMT to fill up the missing parts in the other source language X.[4]

**Multi-encoder NMT with __NULL__** A multi-encoder NMT system using a special symbol __NULL__ to fill up the missing parts in the other source language X.

**Shared-encoder NMT** A single shared-encoder multilingual NMT system as Johnson *et al.* [3] trained using all possible language directions (e.g., en-hr, en-sr, hr-sr, hr-en, sr-hr, and sr-en for en-hr/sr) with target language tags.

*3) Main Results:* Table VIII shows the results in BLEU [26]. We can see that our proposed methods demonstrate larger gains in BLEU than baseline methods in all language settings. On these pairs, this demonstrates that the proposed method is an effective way for using incomplete multilingual corpora, exceeding other reasonable baselines. On the other hand, in {English, Vietnamese, Indonesian}, our proposed methods obtained smaller gains in BLEU than baseline methods compared to the other two language sets: {English, Croatian, Serbian} and {English, Slovak, Czech}. We observed that the improvement by the proposed method is correlated to the performance of the one-to-one NMT with non-English source languages. This suggests the benefit of the proposed method comes from the effectiveness of the non-English source languages. The pseudo translations help the multi-source NMT to improve further especially when a similar language is included in the source languages as with {English, Croatian, Serbian} and {English, Slovak, Czech}. In {English, Vietnamese, Indonesian} case, Vietnamese and Indonesian do not help each other so much as suggested by the one-to-one results, so the additional pseudo translations helped just a little. Our proposed method is affected by which languages it uses, and the proposed method is likely more effective for similar language pairs because the expected accuracy of the data augmentation gets better by the help of lexical and syntactic similarity including shared subword entries. The shared-encoder systems showed similar or slightly better performance in BLEU than the one-to-one models but were much worse than the multi-encoder methods in en-hr/sr and en-sk/cs. It is possibly due to the difference in the input information. The shared-encoder approach takes a single input sentence while the multi-encoder approach utilizes multiple input sentences in different languages.

*4) Different Types of Augmentation:* We examined three types of augmentation: "fill-in", "fill-in and replace", "fill-in and add". In Table VIII, we can see that there were no significant differences among them, despite the fact that their training data were very different from each other. We conducted additional experiments using incomplete corpora with lower quality augmentation by one-to-one NMT to investigate the differences of the three types of augmentation. We created three types of pseudo-multilingual corpora using back-translation from one-to-one NMT and trained multi-encoder NMT models using them. Our expectation here was that the aggressive use of relatively low quality augmented translations may contaminate the training data and decrease the translation accuracy.

Table IX shows the results. In {English, Croatian, Serbian}

[4]This is not exactly *back*-translation because the source translations are not from the target language but from the other source language (English) in our multi-source condition. But we use this familiar term here for simplicity.

TABLE VIII: Main results in BLEU for English-Croatian/Serbian (en-hr/sr), English-Slovak/Czech (en-sk/cs), and English-Vietnamese/Indonesian (en-vi/id).

| Pair | Trg | baseline single-source | | | baseline multi-source | | proposed method | | |
|---|---|---|---|---|---|---|---|---|---|
| | | shared-encoder | one-to-one (En to Trg) | one-to-one (Other to Trg) | multi-encoder (fill up with symbol) | multi-encoder (back translation) | fill-in | fill-in and replace | fill-in and add |
| en-hr/sr | hr | 21.42 | 21.50 | 27.30 | 27.58 | 27.49 | 30.21 | 29.67 | **30.33** |
| | sr | 17.65 | 17.43 | 23.31 | 25.03 | 23.73 | 25.16 | **25.52** | 25.16 |
| en-sk/cs | sk | 16.13 | 14.55 | 18.75 | 20.66 | 20.00 | 22.03 | 21.50 | **22.12** |
| | cs | 15.52 | 15.49 | 18.06 | 20.93 | 19.87 | **23.21** | 23.16 | 23.13 |
| en-vi/id | vi | 25.13 | 25.00 | 17.61 | 24.88 | 25.02 | **25.33** | 24.68 | 25.13 |
| | id | 23.67 | 25.85 | 15.96 | 25.86 | 25.91 | 25.55 | 24.36 | **26.00** |

TABLE IX: The difference of three types of augmentation in BLEU for English-Croatian/Serbian (en-hr/sr), English-Slovak/Czech (en-sk/cs), and English-Vietnamese/Indonesian (en-vi/id). We used one-to-one model to produce pseudo-translations.

| Pair | Trg | multi-encoder NMT (back-translation) | | |
|---|---|---|---|---|
| | | fill-in | fill-in and replace | fill-in and add |
| en-hr/sr | hr | **27.49** | 24.22 | 25.96 |
| | sr | **23.73** | 20.09 | 22.27 |
| en-sk/cs | sk | **20.00** | 15.94 | 18.06 |
| | cs | **19.87** | 16.89 | 18.91 |
| en-vi/id | vi | 25.02 | 24.36 | **25.22** |
| | id | 25.91 | 25.08 | **26.26** |

and {English, Slovak, Czech}, we obtained significant drops in BLEU scores with the aggressive strategies ("fill-in and replace" and "fill-in and add"), while there are few differences in {English, Vietnamese, Indonesian}. One possible reason is that the quality of pseudo-translations by one-to-one NMT in Indonesian and Vietnamese was better than the other languages; in other words, the BLEU from one-to-one NMT in Table VIII was sufficiently good without multi-encoder NMT. Thus the translation performance for Croatian, Serbian, Slovak and Czech could not improve in the experiments here due to *noisy* augmented sentences in those languages. Contrary, the BLEU from "fill-in and add" was the highest when the target language was Indonesian and Vietnamese. In the case the target language is Indonesian, we hypothesize that this is due to much smaller fraction of the missing parts in Indonesian corpus as shown in Table VII, so there should be little room for improvement if we fill in only the missing parts even if the accuracy of the augmented translations is relatively high. On the other hand, in the case the target language is Vietnamese, we hypothesize that the improvement by use of multi-encoder NMT against one-to-one NMT in the baseline was smaller than the other language sets as Table VIII shown, so there is not much difference between augmented machine translations using multi-encoder NMT and one-to-one NMT.

*5) Iterative Augmentation:* It can be noted that if we have a better multi-source NMT system, it can be used to produce better pseudo-translations. This leads to a natural iterative training procedure where we alternatively update the multi-source NMT systems into the two target languages.

Table X shows the results of all languages sets. We can see that this iterative method demonstrate larger gains in BLEU than that at first step in all language sets except when the target language was Vietnamese, though we didn't get monotonically increases in BLEU at each iterative step. We hypothesize the reason is that automatic translations at first step were created from multi-encoder NMT which was trained with the corpus whose missing sentences are filled up with __NULL__, and we used automatic translations created from multi-encoder NMT trained with the corpus filled with automatic translations after the first step. In other words, BLEU score improved if we filled up missing sentences with better quality automatic translations. On the other hand, BLEU score decreased compared to the first step in the case where the target language is Vietnamese. The reason may be that multi-encoder NMT can perform similarly to one-to-one NMT when Vietnamese was the target.

*6) Non-parallelism:* One problem in the use of multilingual corpora is non-parallelism, even in the allegedly manually created and verified translations. In the case of TED multilingual captions, they are translated from English transcripts independently by many volunteers, which may cause some differences in details of the translation in the various target languages. For example in {English, Croatian, Serbian}, Croatian and Serbian translations may not be completely parallel. Table XI shows such an example where the Croatian translation does not have a phrase corresponding to "To be sure," This kind of non-parallelism may be resolved by overriding such translations with pseudo-translations using our proposed strategies; "fill-in and replace" and "fill-in and add". Here, the Croatian pseudo-translation includes the corresponding phrase "Da bi bila sigurna" and can be used to compensate for the missing information. This would be one possible reason of the improvements by "fill-in and replace" or "fill-in and add".

## VI. Test-time Data Augmentation Experiments

Finally, we perform an experiment to confirm the effectiveness of the proposed test-time data augmentation strategy.

### A. Data

We used the same data set and language sets as the experiment of augmentation with pseudo-translations in Section V-E, but here we used a different test set with missing source language sentences, while all source language sentences were complete in the test set in the previous experiment. Table XII shows the number of test sentences for each language set.

TABLE X: BLEU (and BLEU gains compared to step 1) in each step of iterative augmentation.

| Pair | Trg | step 1 | step 2 | step 3 | step 4 |
|---|---|---|---|---|---|
| en-hr/sr | hr | 30.21 (+0.00) | 30.69 (+0.48) | 31.34 (+1.13) | 31.05 (+0.84) |
| | sr | 25.16 (+0.00) | 26.27 (+1.11) | 26.24 (+1.08) | 26.55 (+1.39) |
| en-sk/cs | sk | 22.03 (+0.00) | 22.87 (+0.84) | 22.51 (+0.48) | 22.93 (+0.90) |
| | cs | 23.21 (+0.00) | 23.81 (+0.60) | 24.48 (+1.27) | 24.10 (+0.89) |
| en-vi/id | vi | 25.33 (+0.00) | 25.30 (-0.03) | 25.16 (-0.17) | 25.10 (-0.23) |
| | id | 25.55 (+0.00) | 25.93 (+0.38) | 25.84 (+0.29) | 25.86 (+0.31) |

TABLE XI: Example of the Croatian pseudo-translation. This pseudo-translation is the output of {English, Serbian}-to-Croatian translation.

| Type | Sentence |
|---|---|
| Original (En) | **To be sure,** governments have different resources to bring to the table. |
| Original (Hr) | Vlade imaju razne resurse kojima raspolau. |
| Pseudo (Hr) | **Da bi bila sigurna,** vlada ima razliite resurse koje treba za stol. |

TABLE XII: "Available test sentences" shows the number of available test sentences for each target language.

| Pair | Trg | Available test sentences |
|---|---|---|
| en-hr/sr | hr | 1,145 |
| | sr | 896 |
| en-sk/cs | sk | 602 |
| | cs | 1,966 |
| en-vi/id | vi | 1,405 |
| | id | 333 |

### *B. Baseline methods*

We compared the proposed method with the following two baseline methods:

**One-to-one NMT:** whose source language is English.

**Multi-encoder NMT with 1-best translations:** where an incomplete corpus's missing parts were filled with 1-best translations using beam search (beam width=5) from one-to-one NMT.

We didn't use the mixture of NMT experts as a multi-source NMT in this experiment because the result of it is worse than that of multi-encoder NMT as shown in Chapter V.

### *C. NMT settings*

We conducted the experiment with the same NMT settings as the experiment "augmentation with pseudo-translations" in the section V-E. The multi-encoder NMT model itself was trained with the "fill-in" strategy. We took 5 as the size of the n-best in our proposed method.

### *D. Results*

Table XIII shows the results in BLEU. "proposed" in the results shows the best BLEU with the appropriate hyperparameter $\lambda$. We chose this $\lambda$ by performing a grid search from 0 to 1 with increments of 0.05, and chose the value where BLEU is the highest on the development data. In this table, "1-best" means multi-encoder NMT with an incomplete corpus whose missing parts were filled in with 1-best translations using beam search (beam width=5) from one-to-one NMT.

First of all, we can see that our proposed method achieved larger gains in BLEU than almost all of the multi-encoder NMT baseline methods with 1-best automatic translations from one-to-one NMT. These results show that our proposed method can effectively consider which translations in the n-best are better to fill in. However, our proposed method was worse than the "one-to-one NMT" baseline in English-to-Croatian, English-to-Serbian and English-to-Czech. In the case where the target language is Czech, the amount of training data is larger than that the case where the target language is Slovak, as shown in Table VII. Because of this, Czech-English can more easily get good accuracy than Slovak-English translation, and thus it is less likely that using Slovak as an additional source will be helpful. Thus it is reasonable that the BLEU of our proposed method in Czech was lower than that of one-to-one NMT, while BLEU of our proposed method in Slovak is about 2 points larger than that of one-to-one NMT. In the case where the target language is Croatian and Serbian, the difference of the amount of training data between Croatian and Serbian is small, and we can assume that our proposed method is less effective if the difference of the amount of training data is small.

In conclusion, the proposed method is effective for using an incomplete corpus at test time, although this is dependent on the amount of data available in each of the languages.

TABLE XIII: BLEU results with incomplete test sets for English-Croatian/Serbian (en-hr/sr), English-Slovak/Czech (en-sk/cs), and English-Vietnamese/Indonesian (en-vi/id). In this table, "1-best" means multi-encoder NMT with an incomplete corpus whose missing parts were filled with 1-best translations with the beam width of 5.

| Pair | Trg | baseline method | | proposed method |
|---|---|---|---|---|
| | | one-to-one (En-to-Trg) | 1-best | |
| en-hr/sr | hr | **22.58** | 22.55 | 22.43 |
| | sr | **16.38** | 15.71 | 16.07 |
| en-sk/cs | sk | 14.16 | 16.57 | **16.59** |
| | cs | **15.13** | 13.63 | 13.85 |
| en-vi/id | vi | 22.62 | 22.96 | **23.39** |
| | id | 26.41 | 26.23 | **26.96** |

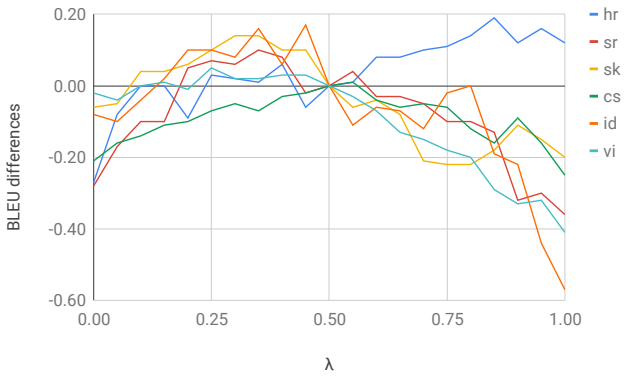



Fig. 8: BLEU differences for each target language by changing $\lambda$. In this figure, BLEU when $\lambda$ is equal to 0.5 is used as a baseline, and the BLEU difference at each $\lambda$ is compared to this baseline.

TABLE XIV: $\lambda$ on the development data and the test data when BLEU is the highest for English-Croatian/Serbian (en-hr/sr), English-Slovak/Czech (en-sk/cs), and English-Vietnamese/Indonesian (en-vi/id). The best BLEU on test data with an appropriate $\lambda$ for test data is also shown.

| Pair | Trg | dev. data | test data | |
|---|---|---|---|---|
| | | $\lambda$ | $\lambda$ | BLEU |
| en-hr/sr | hr | 0.50 | 0.85 | 22.62 |
| | sr | 0.50 | 0.35 | 16.17 |
| en-sk/cs | sk | 0.85 | 0.35 | 16.91 |
| | cs | 0.40 | 0.55 | 13.89 |
| en-vi/id | vi | 0.35 | 0.25 | 23.42 |
| | id | 0.35 | 0.45 | 26.97 |

### E. Discussion

We examined which $\lambda$ value is the best by changing $\lambda$ by increments of 0.05. Figure 8 shows the BLEU differences from the one with $\lambda = 0.5$ for different values of $\lambda$. The BLEU differences were negative for a larger value of $\lambda$, except when the target language is Croatian (hr). From the definition in Eq. (6), a larger value of $\lambda$ means the probabilities of multi-encoder NMT are more heavily weighted than one-to-one NMT when filling in the missing parts. The Croatian results show the use of 1-best English-to-Serbian results to fill the missing parts contributed the final performance, as also shown by the results in Table XIII. From the other results, we can see that we get larger gains in BLEU if the probabilities of the multi-encoder NMT outputs are more important. On the other hand, the $\lambda$ that got the best BLEU on the development data is sometimes very different than the $\lambda$ that got the best BLEU on the test data as shown in Table XIV. Thus, it is assumed that $\lambda$ is really affected by which sentences we use.

## VII. Conclusion

In this paper, we examined approaches for multi-source NMT using *incomplete* multilingual corpora . This problem of missing translations in multi-source NMT happens at both training and test time. Thus we proposed three methods, two for training time and one for test time.

For training time, we proposed to fill missing sentences by __NULL__ at first and improved the performance of multi-source NMT using an incomplete part of a multilingual corpus. Then we improved it further by data augmentation with multi-source NMT. For test time, we proposed to consider n-best translations for a missing source language sentence by one-to-one NMT and achieved better BLEU results than just using 1-best translations. This work is the first study on the problem of multi-source NMT using an incomplete multilingual corpus.

One remaining issue with the third proposed method is that it could not get improvements compared to one-to-one NMT in some language sets when some source sentences were not available at test time. Since the third method relies on all of the techniques proposed in this paper, it is not clear that the poor performance is due to training or test time problem. Future work includes further investigation of the relationship between training and test time augmentation to solve the problems jointly. Language combinations are also important for the proposed approach as we found in some mixed results, so further extensive investigation on related factors such as morphological and syntactic similarities.

## Acknowledgment

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

## References

[1] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-Task Learning for Multiple Language Translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 1723–1732.

[2] O. Firat, K. Cho, and Y. Bengio, "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*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 866–875.

[3] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Vigas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," *Transactions of the Association for Computational Linguistics, vol. 5*, pp. 339–351, 2017.

[4] T.-L. Ha, J. Niehues, and A. Waibel, "Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder," in *Proceedings of the 13th International Workshop on Spoken Language Translation*, Seattle, Washington, December 2016.

[5] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Conference Proceedings: the tenth Machine Translation Summit*, AAMT. Phuket, Thailand: AAMT, 2005, pp. 79–86.

[6] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The United Nations Parallel Corpus v1.0," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), May 2016.

[7] M. Cettolo, C. Girardi, and M. Federico, "WIT$^3$: Web Inventory of Transcribed and Translated Talks," in *Proceedings of the 16th EAMT Conference*, May 2012, pp. 261–268.

[8] J. Tiedemann, "News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces," in *Recent Advances in Natural Language Processing*, N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, Eds. Borovets, Bulgaria: John Benjamins, Amsterdam/Philadelphia, 2009, vol. V, pp. 237–248.

[9] F. J. Och and H. Ney, "Statistical Multi-Source Translation," in *Proceedings of the eighth Machine Translation Summit (MT Summit VIII)*, September 2001, pp. 253–258.

[10] B. Zoph and K. Knight, "Multi-Source Neural Translation," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 30–34.

[11] E. Garmash and C. Monz, "Ensemble Learning for Multi-Source Neural Machine Translation," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, December 2016, pp. 1409–1418.

[12] H. Wu and H. Wang, "Pivot language approach for phrase-based statistical machine translation," *Machine Translation*, vol. 21, no. 3, pp. 165–181, 2007. [Online]. Available: https://doi.org/10.1007/s10590-008-9041-6

[13] Y. Nishimura, K. Sudoh, G. Neubig, and S. Nakamura, "Multi-Source Neural Machine Translaion with Missing Data," in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Association for Computational Linguistics, July 2018, pp. 92–99.

[14] ——, "Multi-source neural machine translation with data augmentation," in *15th International Workshop on Spoken Language Translation (IWSLT)*, Brussels, Belgium, October 2018. [Online]. Available: https://arxiv.org/abs/1810.06826

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *Proceedings of the 3rd International Conference on Learning Representations*, May 2015.

[16] T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1412–1421.

[17] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016, pp. 86–96.

[18] K. Imamura, A. Fujita, and E. Sumita, "Enhancement of encoder and attention using target monolingual corpora in neural machine translation," in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Association for Computational Linguistics, 2018, pp. 55–63. [Online]. Available: http://aclweb.org/anthology/W18-2707

[19] O. Firat, B. Sankaran, Y. Al-onaizan, F. T. Yarman Vural, and K. Cho, "Zero-resource translation with multi-lingual neural machine translation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 268–277. [Online]. Available: https://www.aclweb.org/anthology/D16-1026

[20] X. Wang, H. Pham, Z. Dai, and G. Neubig, "Switchout: an efficient data augmentation algorithm for neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[21] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, November 2016, pp. 1317–1327. [Online]. Available: https://aclweb.org/anthology/D16-1139

[22] M. Morishita, J. Suzuki, and M. Nagata, "NTT Neural Machine Translation Systems at WAT 2017," in *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, 2017, pp. 89–94.

[23] G. Neubig, P. Arthur, and K. Duh, "Multi-target machine translation with multi-synchronous context-free grammars," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May–June 2015, pp. 293–302. [Online]. Available: http://www.aclweb.org/anthology/N15-1033

[24] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018, pp. 66–75.

[25] D. P. K. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, May 2015.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 311–318.

[27] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, 2018, pp. 186–191. [Online]. Available: https://aclanthology.info/papers/W18-6319/w18-6319

**Yuta Nishimura** received his B.E. from Tohoku University, Miyagi, Japan in 2017, and his M.E. in information science from Nara Institute of Science and Technology in 2019. From 2019, he has been doing research and development at PLAID, Inc., where he is pursuing research in user analysis with a focus on machine learning approach.

**Katsuhito Sudoh** is an associate professor of Nara Institute of Science and Technology. He received a bachelor's degree in engineering in 2000, and a master's and Ph.D. degree in informatics in 2002 and 2015, respectively, from Kyoto University. He was in NTT Communication Science Laboratories from 2002 to 2017. He currently works on machine translation and natural language processing. He is a member of the Association for Computational Linguistics (ACL), the Association of Natural Language Processing (ANLP), the Information Processing Society of Japan (IPSJ) and the Acoustical Society of Japan (ASJ).

**Graham Neubig** received his B.E. from University of Illinois, Urbana-Champaign in 2005, and his M.S. and Ph.D. in informatics from Kyoto University in 2010 and 2012 respectively. He is currently an assistant professor at Carnegie Mellon University and an visiting associate professor at Nara Institute of Science and Technology. His research interests include speech and natural language processing, with a focus on machine learning approaches for applications such as machine translation, speech recognition, and spoken dialog.

**Satoshi Nakamura** is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011 and IEEE Signal Processing Magazine Editorial Board member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.