Electronic Preprint for Journal of Information Processing Vol.26

Regular Paper

Exploiting Multilingual Corpora Simply and Efficiently in Neural Machine Translation

Raj Dabre[1,a)] Fabien Cromieres[2,b)] Sadao Kurohashi[1,c)]

Received: August 24, 2017, Accepted: January 15, 2018

Abstract: In this paper, we explore a simple approach for "Multi-Source Neural Machine Translation" (MSNMT) which only relies on preprocessing a N-way multilingual corpus without modifying the Neural Machine Translation (NMT) architecture or training procedure. We simply concatenate the source sentences to form a single, long multi-source input sentence while keeping the target side sentence as it is and train an NMT system using this preprocessed corpus. We evaluate our method in resource poor as well as resource rich settings and show its effectiveness (up to 4 BLEU using 2 source languages and up to 6 BLEU using 5 source languages) and compare them against existing approaches. We also provide some insights on how the NMT system leverages multilingual information in such a scenario by visualizing attention. We then show that this multi-source approach can be used for transfer learning to improve the translation quality for single-source systems without using any additional corpora thereby highlighting the importance of multilingual-multiway corpora in low resource scenarios. We also extract and evaluate a multilingual dictionary by a method that utilizes the multi-source attention and show that it works fairly well despite its simplicity.

Keywords: Neural Machine Translation (NMT), multi-source NMT, empirical comparison, transfer learning, deep learning, dictionary extraction

1. Introduction

Even though Machine Translation is often only considered in the context of the translation between two languages, there are many contexts where it is relevant to consider more than two languages. This is because we can have a sentence in two different languages and want to translate it into a third language (multi-source translation). It can also be the case that the training corpora we have are naturally multilingual which is an aspect that can be leveraged.

A well known example of this situation is the European Parliament Proceedings. These Proceedings are themselves multilingual corpora written in 21 European languages, made available in the often used EuroParl corpus [10]. Furthermore, because they are produced by translating successively the source language in 20 other languages, an MT system could leverage the translations of the first few languages to produce better translations of the other languages.

Neural Machine Translation (NMT) [1], [3], [17] is a recent approach to Machine Translation based on the use of Deep Learning for producing end-to-end trainable MT systems. Some work has already been done to use NMT in a multi-source context [19].

As opposed to these works that design a specific model for multi-source, we explore a simple method (originally proposed to use pre-translations as additional sources [13]) that can train any single-source NMT with a multilingual corpus to produce a multi-source MT system. We show that this system works at least as well as the ones using specifically designed NMT models.

In addition, we propose a method for exploiting a multilingual corpus to improve single-source translation quality. We think this method could have significance on the way resources for low-resource languages are developed, and therefore we focus on low-resource scenarios for a part of our evaluation.

The main contributions of this paper are as follows:

- Exploiting a simple preprocessing step that allows for multi-source NMT (MSNMT) without any change to the NMT architecture [*1].
- A method for improving single-source translation with a multilingual corpus via transfer learning of the multi-source model

We evaluate our approaches in a resource poor as well as a resource rich setting and compare it with two existing methods [6], [19] for MSNMT. We also perform additional analysis by visualizing attention vectors and evaluating a dictionary extracted using the multisource attention.

2. Related Work

One of the first studies on multi-source MT [14] explored how word-based SMT systems would benefit from multiple source languages. The work on multi-encoder multi-source NMT [19] is the first multi-source NMT approach which focused on utilizing French and German as source languages to translate into English.

[1] Graduate School of Informatics, Kyoto University, Kyoto 612–8316, Japan
[2] Japan Science and Technology Agency, Kawaguchi, Saitama 332–0012, Japan
[a)] raj@nlp.ist.i.kyoto-u.ac.jp
[b)] fabien@pa.jst.jp
[c)] kuro@i.kyoto-u.ac.jp

[*1] One additional benefit of our approach is that any NMT architecture can be used, be it attention based or hierarchical NMT.

However their method led to models with substantially larger parameter spaces and they did not experiment with many languages. Multi-source ensembling using a multilingual multi-way NMT model [6] is an end-to-end approach but requires training a very large and complex NMT model. The work on multi-source ensembling which uses separately trained single-source models [7] is comparatively simpler in the sense that one does not need to train additional NMT models but the approach is not truly end-to-end since it needs an ensemble function to be learned. This method also helps eliminates the need for N-way corpora which allows one to exploit bilingual corpora which are larger in size. In all cases one ends up with either one large model or many small models for which an ensemble function needs to be learned.

Concatenating multiple source sentences for multi-source NMT [13] was used for exploiting pre-translations generated by PBSMT systems as additional sources but not for situations where multiple source languages (like French, German and Italian) are available.

Other related works include Transfer Learning [20] and Zero Shot NMT [9] which help improve NMT performance for low resource languages. Finally it is important to note works that involve the creation of N-way corpora. Some examples of N-way corpora (ordered from largest to smallest according to number of lines of corpora) are: United Nations [18], Europarl [10], Ted Talks [2], ILCI [8] and Bible [4] corpora.

3. Neural Machine Translation

NMT is an end-to-end approach for translating from one language to another, that relies on deep learning, to train a translation model [1], [3], [17]. We use an encoder-decoder model with an attention mechanism [1] for all our experiments. For our main approach we do not modify the architecture at all but we do make the necessary modifications to the model for comparison against existing approaches. This model is also known as **"rnnsearch"**. **Figure 1** describes the rnnsearch model [1], which takes in an input sentence and its translation and updates its parameters by minimizing the loss on the predicted translation. The model consists of 3 main parts, namely, the encoder, decoder and attention model.

The encoder consists of an embedding mechanism to obtain continuous space representations of the input words. These embeddings by themselves do not contain information about relationships between words and their positions in the sentence. Using a long short-term memory (LSTM) which is a RNN layer [3], the word relationships and position information can be obtained. A RNN maintains a memory (also called a state or history) which allows it to generate a continuous space representation for a word given all past words that have been seen. There are 2 LSTM layers: forward and backward information to model relationships for words given past as well as future words. By using both forward and backward recurrent information one obtains a continuous space representation for a word given all words before as well as after it.

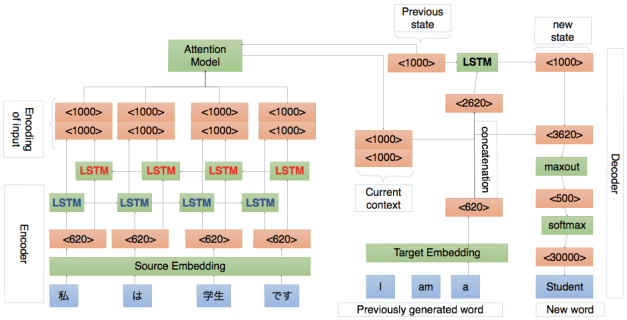

Fig. 1 The architecture of an attention based NMT model, as described in Ref. [1]. The notation "<1000>" means a vector of size 1,000. The vector sizes shown here are the same as in the original paper.

The decoder is conceptually a RNNLM with its own embedding mechanism, a LSTM layer to remember previously generated words and a deep softmax layer to predict a target word. The encoder and decoder are coupled using an attention mechanism which computes a weighted average of the recurrent representations generated by the encoder. The attention mechanism thereby acts as a soft alignment mechanism. This weighted averaged vector, also known as the context or attention vector, is fed to the decoder LSTM along with the embedding of the previously predicted word to produce a representation that is passed to the deep softmax layer [*2] to predict the next word.

The parameters of this NMT model are updated using a variety of optimization algorithms to minimize loss, or namely Stochastic Gradient Descent (SGD), Adaptive Gradient (AdaGrad) and ADAM. We use ADAM for our experiments since it helps accelerate the rate of learning.

4. Our Approaches

We will first describe our method for training a standard (single-source) NMT model using a Multilingual Corpus to produce a multi-source NMT model. We then propose in Section 4.2 an extension of this method that also leads to better single-source translation. Finally, we describe an additional extention which is a simple method in Section 4.3 to extract a multilingual dictionary with a significantly larger number of entries than the subword vocabulary size.

4.1 Multi-Source Models

Here we describe our method for training a single-source NMT model using a multilingual corpus to produce a multi-source NMT model. Simply put we convert the multilingual multiway corpus into a bilingual corpus. To do this, for each target sentence we concatenate the corresponding source sentences leading to a parallel corpus where the source sentence is a very long sentence that conveys the same meaning in multiple languages. An example line in such a corpus would be: source: "I am a boy Je suis un garcon" and target: "Watashiwa otokonoko desu" [*3]. The 2 source languages (this is just an example but in reality this is applicable for N source languages) here are English and French whereas the target language is Japanese. In this example each source sentence is a word conveying "I am a boy" in different

[*2] The deep softmax layer contains a maxout layer which is a feedforward layer with max pooling. It takes in the attention vector, the embedding of the previous word and the recurrent representation generated by the decoder LSTM and computes a final representation, which is fed to a simple softmax layer.

[*3] We romanized the Japanese sentence for readability.

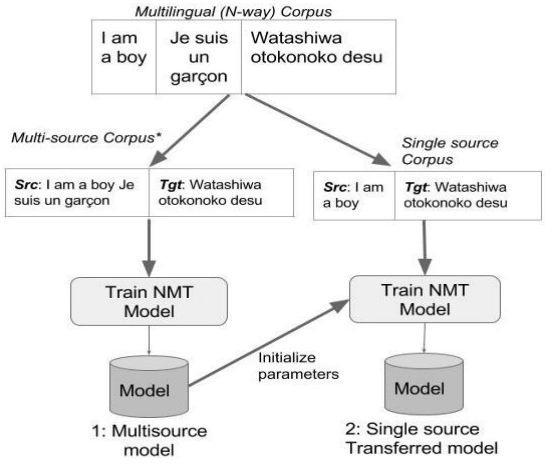

Fig. 2 Our multi-source NMT approach and applying it to Transfer Learning. The left hand side represents the flow for training a single-source NMT model using a Multilingual Corpus to produce a multi-source NMT model (See Section 4.1). This model can then be used as a parent model for transfer learning to improve the single-source translation quality (See Section 4.2).

languages [*4]. One can now use this bilingual corpus to learn an NMT model using any off the shelf NMT toolkit. Refer to the left hand side of **Fig. 2** for a visual representation.

This NMT model can then be used for multi-source translation by simply concatenating the input sentences in the same order as when the training corpus was created. We expect that the NMT model will be clever enough to utilize the information contained in all the input sentences and as can be seen in Sections 6.1.1 and 6.1.5 this is indeed the case.

4.2 Using Multi-Source Models for Transfer Learning

We expect that the method above will give good improvements in translation quality when the input sentences are available in multiple languages. However we find that it is possible to leverage a multiway corpus to improve single-source translation quality. This can be achieved by a form of transfer learning of the multi-source model. To perform transfer learning we use the approach proposed by Ref. [20]. We simply initialize the parameters of a single-source model with those learned for the multi-source model [*5]. These multi-source models are known as the parent models where as the transferred models are known as the child models. Refer to the right hand side of Fig. 2 for a visual representation of the flow.

4.3 Using Multi-Source Models for Dictionary Extraction

One major limitation of NMT for extracting dictionaries is that they work with a limited vocabulary size by considering only the most frequent words which leads to tiny dictionaries. Subword units using BPE segmentation [16] allow for infinite vocabulary sizes which we exploit for extracting our dictionaries. We simply rely on gluing the subword units and updating the attention values of subword units they are aligned to. This approach works surprisingly well and is able to successfully reconstruct and align words that were split into subwords in many cases. We first force align the multi-source corpus with the target language corpus in order to obtain the attention probabilities. We dump all the attention information into a single file with the following format:

- Line i: source sentence (concatenated multi-source sentence)
- Line i+1: target sentence
- Lines j = i+2 to i+k+2 (where k is the number of subwords in the target sentence): <target-sentence-subword-j><tab>list(<attention-value>:<source-sentence-subword-x>for source-sentence-subword-x in multi-source-sentence)

An example of what a subword looks like is: "Po__ ta__ to" for the word "Potato" where "__" is the delimiter that indicates that the current subword is not the end of the surface word. We also assume that each (multi) source sentence subword is tagged with a token that indicates the language corresponding to the source sentence that contains it. Algorithm 1 [*6] contains the detailed steps for extracting the dictionary [*7].

5. Experimental Settings

All of our experiments were performed using an encoder-decoder NMT system with attention for the various baselines and multi-source experiments. In order to enable infinite vocabulary and reduce data sparsity we use the Byte Pair Encoding (BPE) based word segmentation approach [16]. We evaluate our models using the standard BLEU [15] metric [*8] on the translations of the test set. Baseline models are single-source models.

5.1 Languages and Corpora Settings

All of our experiments were performed using the publicly available ILCI [*9] [8], United Nations [*10] [18] and Europarl [*11] [10]. We use the UN corpus for a resource rich setting whereas the others are used for a resource poor setting. We tried to use as many datasets as possible to indicate that our work is not dataset specific.

The ILCI corpus is a 6-way multilingual corpus spanning the languages Hindi, English, Tamil, Telugu, Marathi and Bengali was provided as a part of the task. The target language is Hindi and thus there are 5 source languages. The training, development and test sets contain 45,600, 1,000 and 2,400 6-lingual sentences respectively.

From the IWSLT corpus we extract a trilingual French, German and English training set of 191,381 lines, a development set of 880 lines (called dev2010) and two test sets of 1,060 (tst2010) and 886 (tst2013) lines. English is the target language. We experimented with 4-lingual and 5-lingual scenarios comprising of two additional languages, Arabic and Czech but we omit the results for brevity.

The UN corpus spans 6 languages: French, Spanish, Arabic,

[*4] Note that there are no delimiters between the individual source sentences.
[*5] It is important to note that the target language vocabularies for both the models should be the same, which they are in our setting.
[*6] The algorithm assumes that each source and target sentence is delimited by an end of sentence delimiter such as a full-stop which will ensure that all words before the delimiter will be included in the dictionary.
[*7] the pseudo-code is similar to the python coding style
[*8] This is computed by the multi-bleu.pl script, which can be downloaded from the public implementation of Moses [11].
[*9] This was used for the Indian Languages MT task in ICON 2014 and 2015.
[*10] https://conferences.unite.un.org/uncorpus
[*11] http://www.statmt.org/europarl

```
Result: finaldict (The final multilingual dictionary)
finaldict = hashmap() (The keys are surface words and the values are
surface word - fractional count pairs);
for line in force-aligned-file do
    if line is either "source sentence" or "target sentence" then
        worddict = hashmap() (This contains the source language
        surface word and the cumulative attention value which acts as
        a fractional count);
        previous-target-subword = "";
    else
        (the current line contains target subwords (line[0]) and the
        source subwords with attention values (line[1:]))
        current-target-subword = line[0];
        if "__" is not the ending of previous-target-subword then
            sort worddict by value and add top N entries to finaldict
            using previous-target-subword as the key (We
            experimented with N=5 to minimize the number of noisy
            entries);
            previous-target-subword = current-target-subword;
            worddict = hashmap();
        else
            if "__" is the ending of previous-target-subword then
                previous-target-subword =
                previous-target-subword[:-2] +
                current-target-subword (since the last 2 characters are
                the delimiters);
                previous-source-subword, previous-attention-value =
                line[1].split(":");
                for subword-attention-pair in line[2:] do
                    current-source-subword, current-attention-value
                    = subword-attention-pair.split(":");
                    if "__" is not the ending of
                    previous-source-subword then
                        worddict[previous-source-subword] +=
                        previous-attention-value;
                        previous-source-subword =
                        current-source-subword;
                        previous-attention-value =
                        current-attention-value;
                    else
                        if "__" is the ending of
                        previous-source-subword then
                            previous-source-subword =
                            previous-source-subword[:-2] +
                            current-target-subword;
                            previous-attention-value +=
                            current-attention-value;
                        end
                    end
                end
            end
        end
    end
end
```

Algorithm 1: Algorithm for dictionary extraction that uses the multilingual attention obtained from a multi-source model that uses the concatenation approach.

Chinese, Russian and English. Although there are 11 million 6-lingual sentences we use only 2 million for training since our purpose was not to train the best system but to show that our method works in a resource rich situation as well. The development and test sets provided contain 4,000 lines each and are also available as 6-lingual sentences. We chose English to be the target language and focused on Spanish, French, Arabic and Russian as source languages. Due to lack of computation time constraints we only worked with the following source language combinations: French and Spanish, French and Russian, French and Arabic and Russian and Arabic.

The Europarl corpus spans over 20 languages but is not multilingual multi-way. For our experiments, we simulate a low resource scenario by using a 200,000 line, 5-lingual training subset of the full corpus spanning French, German, Spanish, Italian and English. We use 5-lingual, dev and test sets of 4,000 lines each which are disjoint from the training set. We performed the transfer learning and dictionary extraction and evaluation experiments on the Europarl corpus only.

5.2 NMT Model Settings

For training various NMT systems, we used the open source KyotoNMT toolkit [*12] [5]. KyotoNMT implements an Attention-based Encoder-Decoder [1] with slight modifications to the training procedure. We modify the NMT implementation in KyotoNMT to enable multi-encoder multi-source NMT [19]. In the case of multiple encoders, one for each language, each encoder has its own separate vocabulary and attention mechanism. Since the NMT model architecture used in Ref. [19] is slightly different from the one in KyotoNMT, the multi-encoder implementation is not identical (but is equivalent) to the one in the original work. The model and training details are as below. Unless mentioned otherwise these settings remain the same throughout the paper.

- BPE vocabulary size: 8k (separate models for source and target) for ILCI and IWSLT corpus setting and 16k for the UN corpus setting. When training the BPE model for the source languages we learn a single shared BPE model. In case of languages that use the same script this allows for cognate sharing thereby reducing the overall vocabulary size requirement. In the case of multiple encoders, one for each language, each encoder has its own separate vocabulary.
- Model architecture: Same as that in Ref. [1] except that we use LSTMs instead of GRUs and we use 500 node hidden layer for attention.
- Maximum sentence length threshold during training: For all settings we set this to 100 for all single-source models and N*100 for multisource models that use the concatenation approach. In the ILCI setting the maximum sentence length in the training corpus is less than 100 and thus for a 5 source model a maximum sentence length threshold of 500 ensures that the complete training data is used.
- Training steps: 10k [*13] for 1 source, 15k for 2 source and 40k for 5 source settings when using the IWSLT and ILCI corpora. 200k for 1 source and 400k for 2 source for the UN corpus setting to ensure that in both cases the models get saturated with respect to heir learning capacity. The increased number of iterations for the multi-source models is to compensate for the smaller batch sizes that we used.
- Batch size: 64 for single-source, 16 for 2 sources and 8 for 3 sources and above for ILCI corpus setting. 32 for single-source and 16 for 2 sources for the UN corpus set-

[*12] https://github.com/fabiencro/knmt
[*13] We observed that the models start overfitting around 7k-8k iterations.

ting. Because, longer sequences require more GPU memory for training we used smaller batch sizes to compensate. It might seem unfair that different models (single-source versus multi-source) use different batch sizes for training but based on preliminary experiments, smaller batch sizes only affected the time taken to reach optimal performance and not the final BLEU scores [*14].

- Optimization algorithms: Adam with a default initial learning rate of 0.01
- Gradient clipping threshold: 1.0 for all settings. This value is used in most existing works for NMT.
- Choosing the best model: Evaluate the model on the development set and select the one with the best BLEU [15] after reversing the BPE segmentation on the output of the NMT model. This is also called early stopping.
- Beam size for decoding: 16 for all settings. We performed evaluation using beam sizes 4, 8, 12 and 16 but found that the differences in BLEU between beam sizes 12 and 16 are small and gains in BLEU for beam sizes beyond 16 are insignificant.
- Number of steps in decoding: 1.5 times the source sentence length for all settings. For the multi-source models that use the concatenation approach 1.5 times seems overkill but our decoder automatically stops generating new tokens when the "end of sentence (EOS)" token is generated.

5.3 NMT Models

5.3.1 Multi-Source Models

We train and evaluate one source to one target (baselines) and N-source to one target models using the following 3 methods: Ours, Multi-Encoder [19] [*15] and Ensembling [6] [*16]. The latter two methods are for comparison. For the 2-source models in the ILCI corpus setting we considered all possible source language pairs. In the IWSLT corpus setting there is only one possibility: French+German to English model. However, in the UN corpus setting we only tried the following one source one target models: French-English, Russian-English, Spanish-English and Arabic-English. The two source combinations we tried were: French+Spanish, French+Arabic, French+Russian, Russian+Arabic. The target language is English.

5.3.2 Single-Source Transferred Models

Our transfer learning experiments are performed using the Europarl corpus. The BPE vocabulary size is 12,000 for both source and target languages, irrespective of single or multi-source models. The embedding, LSTM and attention hidden layer sizes are 512 each. We use a batch size of 32 for single-source models and 8 for the multi-source models.

We train the following 4 source models: French+Spanish+Italian+German to English, French+Spanish+Italian+English to Spanish and French+Spanish+Italian+English to German. For each of these 4 source models we also train corresponding single-source models as baselines. For instance we train French-English, Spanish-English, Italian-English and German-English corresponding to French+Spanish+Italian+German to English. We train 3 additional models (corresponding to each of the multi-source models) using corpora obtained by merging all the corpora of the 4 individual language pairs. These multilingual models are essentially the same as the ones in Zero Shot NMT [9] except that there is only one target language and thus we do not use any tokens to indicate the target language. We call these models 4S1T models which can only translate single-source sentences. We use both the 4 source and 4S1T models to initialize the single-source models for transfer learning. Unlike the original work [20] we do not perform any regularization by freezing parts of the model while training.

We used the French+Spanish+Italian+German to English model to extract multilingual dictionaries using our algorithm proposed in Section 4.3. We extracted dictionaries for the Europarl corpus by force decoding (to obtain attention values) the training set multisource sentences using the target reference sentences. The multisource sentences comprised of concatenated Spanish, French, Italian and German sentences and the target sentences are English sentences. We manually evaluated the dictionaries obtained for the 100 most frequent English words in the Europarl corpus. Our reason for choosing these languages is that these are the easiest to evaluate manually given the number of resources available online. We leave the evaluation of dictionaries for other language pairs as future work.

6. Results

We divide our results into two subsections: Section 6.1 for the evaluation of our multi-source method and Section 6.2 for the evaluation of our work on transfer learning using the multi-source models.

6.1 Evaluation of Multi-Source Models

For the ILCI corpus setting, **Table 1** contains the BLEU scores for all the multi-source models and the lexical similarity scores for all combinations of source languages, two at a time. The last row of Table 1 contains the BLEU score for all the multi-source settings which uses all 5 source languages. The caption contains a complete description of the table. Refer to **Table 3** for the results of the UN corpus setting, and to **Table 2** for the IWSLT corpus setting.

6.1.1 Main Findings

From Tables 1, 2 and Table 3 it is clear that our simple source sentence concatenation based approach (under columns labeled "our") is able to leverage multiple languages leading to significant improvements compared to the BLEU scores obtained using any of the individual source languages. The ensembling (under columns labeled "ens") and the multi-encoder (under columns la-

[*14] Recent research seems to indicate that models trained with larger batch sizes are better than those trained with smaller batch sizes. By this logic our multi-source models that use smaller batch sizes are already at a natural disadvantage. Despite this it will be seen that the multi-source models beat the single-source models.

[*15] To be specific we implemented the technique where attentions are computed for both source languages and concatenated before feeding then to the decoder to predict a target word.

[*16] We use the multi-source ensembling approach that late averages N one source to one target models. Late averaging implies averaging the logits of multiple decoders before computing softmax to predict the target word. In the original work a single multilingual multiway NMT model was trained and ensembled but we train separate NMT models for each source language.

Table 1 **ILCI corpus results for multi-source models**: BLEU scores for two source to one target setting for all language combinations and for five source to one target using the ILCI corpus. The languages are Bengali (Bn), English (En), Marathi (Mr), Tamil (Ta), Telugu (Te) and Hindi (Hi). Each language is accompanied by the BLEU score for translating to Hindi from that language and its lexical similarity with Hindi. Each cell in the upper right triangle contains the BLEU scores using a. Our proposed approach (our), b. Multi-Source ensembling approach (ens), c. Multi-Encoder Multi-Source approach (me) and d. The lexical similarity (sim; in tiny font size). The best BLEU score is in bold. The train, dev, test split sizes are 45,600, 1,000 and 2,400 lines respectively.

Source Language 1	**Source Language 2** [XX-Hi BLEU] *XX-Hi sim*															
	En [11.08] *0.20*				**Mr** [24.60] *0.51*				**Ta** [10.37] *0.30*				**Te** [16.55] *0.42*			
	our	ens	me	*sim*	our	ens	me	*sim*	our	ens	me	*sim*	our	ens	me	*sim*
Bn [19.14] *0.52*	**20.70**	19.45	19.10	*0.18*	29.02	**30.10**	27.33	*0.46*	19.85	**20.79**	18.26	*0.30*	22.73	**24.83**	22.14	*0.39*
En [11.08] *0.20*	-				25.56	23.06	**26.01**	*0.20*	14.03	**15.05**	13.30	*0.18*	18.91	**19.68**	17.53	*0.20*
Mr [24.60] *0.51*	-				-				**25.64**	24.70	23.79	*0.33*	27.62	**28.00**	26.63	*0.43*
Ta [10.37] *0.30*	-				-				-				18.14	**19.11**	17.34	*0.38*
All	our: **31.56**				ens: 30.29				me: 28.31							

Table 2 **IWSLT corpus results for multi-source models**: BLEU scores for the single-source and N source settings using the IWSLT corpus. The languages are French (Fr), German (De) and English (En). We give the BLEU scores for two test sets tst2010 and tst2013. The best BLEU score is in bold. The train, dev, test split sizes are 191,381, 880 and 1,060/886 (tst2010/tst2013) lines respectively.

Language Pair	**BLEU tst2010**			**BLEU tst2013**		
Fr-En	19.72			22.05		
De-En	16.19			16.13		
Fr+De-En	our	ens	me	our	ens	me
	22.56	18.64	22.03	**24.02**	18.45	23.92

Table 3 **UN corpus results for multi-source models**: BLEU scores for the single-source and 2 source settings using the UN corpus. The languages are Spanish (Es), French (Fr), Russian (Ru), Arabic (Ar) and English (En). We give the BLEU scores for the test set. The highest score is the one in bold. All BLEU score improvements are statistically significant ($p < 0.001$) compared to those obtained using either of the source languages independently. The train, dev, test split sizes are 2M, 4,000 and 4,000 lines respectively.

Language Pair	**BLEU**
Es-En	49.20
Fr-En	40.52
Ar-En	40.58
Ru-En	38.94

Source Combination	**BLEU**		
	our	ens	me
Es+Er	**49.93***	46.65	47.39
Fr+Ru	**43.99**	40.63	42.12
Fr+Ar	43.85	41.13	**44.06**
Ar+Ru	41.66	43.12	**43.69**

beled "me") approaches also leads to improvements in BLEU. Note that in every single case, gains in BLEU are statistically significant regardless of the methods used. It should be noted that in a resource poor scenario ensembling generally outperforms all other approaches but in a resource rich scenario our method as well as the multi-encoder method are much better. However, the comparison with the ensembling method is unfair to our method since the former uses N times more parameters than the latter. However, one important aspect of our approach is that the model size for the multi-source systems is the same as that of the single-source systems since the vocabulary sizes are exactly the same. The multi-encoder systems involve more parameters whereas the ensembling approach does not allow for the source languages to truly interact with each other.

6.1.2 Correlation between Linguistic Similarity and Gains Using Multiple Sources

We calculated the ***lexical similarity*** [*17] between the languages involved in using the Indic NLP Library [*18]. The objective behind this is to determine whether or not lexical similarity, which is also one of the indicators of linguistic similarity and hence translation quality [12], is also an indicator of how well two source languages work together.

In the case of the ILCI corpus setting, Table 1, it is clear that no matter which source languages are combined, the BLEU scores are higher than those given by the single-source systems. Marathi and Bengali are the closest to Hindi (linguistically speaking) compared to the other languages and thus when used together they help obtain an improvement of 4.39 BLEU points compared to when Marathi is used as the only source language (24.63). However it can be seen that combining any of Marathi, Bengali and Telugu with either English or Tamil leads to smaller gains. There is a strong correlation between the gains in BLEU and the lexical similarity. Bengali and English which have the least lexical similarity (0.18) give only a 1.56 BLEU improvement whereas Bengali and Marathi which have the highest lexical similarity (0.46) give a BLEU improvement of 4.42 using our multi-source method. This seems to indicate that although multiple source languages do help, source languages that are linguistically closer to each other are responsible for maximum gains (as evidenced by the correlation between lexical similarity and gains in BLEU). Finally, the last row of Table 1 shows that using additional languages leads to further gains leading to a BLEU score of 31.3 which is 6.5 points above when only Marathi is used as the only source language and 2.11 points above when Marathi and Bengali are used as the source languages. As future work it will be worthwhile to investigate the diminishing returns in BLEU improvement obtained per additional language.

6.1.3 Performance in Resource Rich Settings

In the UN corpus setting, Table 3, where we used approximately 2 million training sentences, we also obtained improvements in BLEU. In the case of the single-source systems we observed that the BLEU score for Spanish-English was around 9 BLEU points higher than for French-English which is consistent with the observations in the original work concerning the construction of the UN corpus [18]. Furthermore, combining using French and Spanish together leads to a small (0.7) improvement in BLEU (over Spanish-English) that is statistically significant ($p < 0.001$) which is to be expected since the BLEU

[*17] https://en.wikipedia.org/wiki/Lexical_similarity

[*18] http://anoopkunchukuttan.github.io/indic_nlp_library

for Spanish-English is already much better than the BLEU for French-English. Since the BLEU scores for French, Arabic and Russian to English are closer to each other we can see that the BLEU scores for French+Arabic, French+Russian and Arabic+Russian to English are around 3 BLEU points higher than those of their respective single-source counterparts.

6.1.4 Regarding Sequence Lengths and Vocabulary Size Limits

In Section 5.2 we mentioned that we learn a shared subword vocabulary for all source languages. A subword vocabulary leads to a slight increase in the length of sentences but eliminates the problem of unknown words. There are two related important aspects that must be considered: combinations of languages and maximum number of source languages that can be combined in order to obtain maximal improvements in translation quality. In theory it is possible to combine any number of source languages but from a practical point of view two to three is sufficient.

In a setting where the source languages use the same script, the sizes (in terms of number of characters per subword) of subword units that can be learned is significantly larger than in the case of languages that use completely different scripts. Shorter subwords leads to vocabularies that approach characters and the traditional NMT approach is known to perform poorly when using character sequences. Moreover, increasing the number of source languages also causes the subword vocabulary to approach a character level vocabulary. In Table 1 it can be seen that using more languages does leads to an increase in translation quality but such a case is not practical. This leads to a situation where unnecessarily longer sequences are used for little gain. As such, it is better to use two to three source languages that are linguistically closer because they also increase the chances of script sharing and cognate sharing. Cognate and script sharing also leads to larger subword units for rarer words. Whether this is a good thing or not should be verified experimentally, something we leave for future work.

In the case of languages like Chinese and Japanese in which the number of basic characters (which form the initial subword vocabulary) is extremely high, the sequences tend to be much longer if a smaller vocabulary size is specified. Using Chinese and Japanese as sources together is much better than using either of them with other languages like English or Hindi. The reason for this is that Chinese and Japanese scripts contain a large number of similar characters and this increases the possibility of cognate sharing and thereby larger subwords for rarer words. But if Chinese or Japanese is combined with English then half the subword vocabulary quota will be allotted to English which means that the Chinese subwords will be mostly characters. This also increases the effective lengths of input sequences which makes training NMT models more difficult.

As a rule of thumb it would be better to consider using more source languages if they are linguistically closer and share scripts and cognates. In other situations it would be better to use two or three source languages and avoid the problem of subwords vocabularies that approach character level vocabularies which also leads to extremely long sequences.

6.1.5 Studying Multi-Source Attention

To study multi-source attention, we obtained visualizations for the attention vectors for a few sentences from the test set. Refer to **Fig. 3** for an example. Note that, in the figure, we use a horizontal line to separate the languages but the NMT system receives a single, long multi-source sentence. The words of the target sentence in Hindi are arranged from left to right along the columns whereas the words of the multi-source sentence are arranged from top to bottom across the rows. Note that the source languages (and lexical similarity scores with Hindi) are in the following order: Bengali (0.52), English (0.20), Marathi (0.51), Tamil (0.30), Telugu (0.42).

The most interesting thing that can be seen is that the attention mechanism focuses on each language but with varying degrees of focus. Bengali, Marathi and Telugu are the three languages that receive most of the attention (highest lexical similarity scores with Hindi) whereas English and Tamil (lowest lexical similarity scores with Hindi) barely receive any. Building on this observation we believe that the gains we obtained by using all 5 source languages were mostly due to Bengali, Telugu and Marathi whereas the NMT system learns to practically ignore Tamil and English. However there does not seem to be any detrimental effect from using English and Tamil.

From **Fig. 4** it can be seen that this observation also holds in the UN corpus setting for French+Spanish to English where the attention mechanism gives a higher weight to Spanish words compared to French words since the Spanish-English translation quality is about 9 BLEU points higher than the French-English translation quality. It is also interesting to note that the attention can potentially be used to extract a multilingual dictionary simply by learning a N-source NMT system and then generating a dictio-

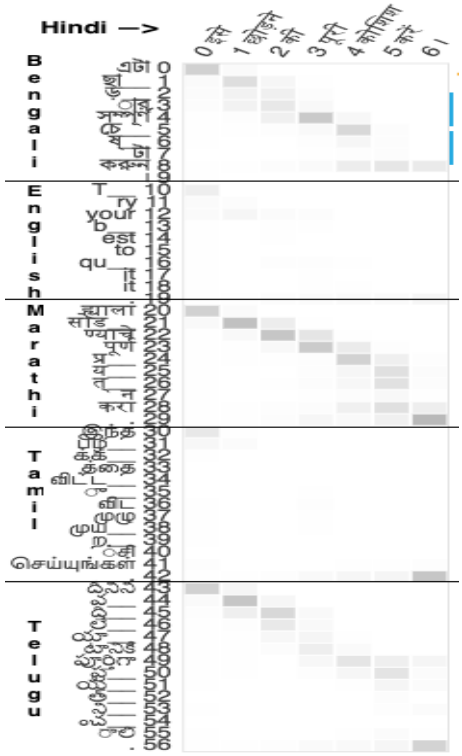

Fig. 3 Attention Visualization for ILCI corpus setting for Bengali, English, Marathi, Tamil and Telugu to Hindi. A horizontal black line is used to separate the source languages but the NMT system receives a single, long multi-source sentence.

Table 4 **Europarl corpus results for Transfer Learning using multi-source Models:** BLEU scores for the 5-lingual corpus spanning French (Fr), Spanish (Es), German (De), Italian (It), English (En). For each language pair, we give BLEU scores for the test set translated using the a. Baseline model, b. 4S1T model, c. Transferred Model using the multi-source model as the parent model and d. Transferred Model using the 4S1T model as the parent model. The scores in bold are statistically significant ($p < 0.001$) compared to the baseline scores. Transfer model scores obtained using multi-source models as parents are marked with an asterisk (*) when they are statistically significant ($p < 0.001$) compared to scores obtained using 4S1T models as parents. The train, dev, test split sizes are 200k, 4k and 4k respectively.

Model Type	Language Pair											
	Fr-En	Es-En	It-En	De-En	Fr-Es	En-Es	It-Es	De-Es	Fr-De	Es-De	It-De	En-De
Baseline	29.06	32.17	26.88	26.40	30.08	32.94	28.94	21.79	15.79	15.9	14.53	18.36
4S1T	28.45	30.99	26.89	23.99	29.37	31.35	28.57	22.48	**16.32**	**17.12**	**15.47**	18.01
Transfer Using 4 source model	**30.46***	**33.54***	**28.30**	26.09	**32.52***	**35.45***	**31.42***	**24.37***	**17.64**	**17.79**	**16.43**	**19.82**
Transfer Using 4S1T model	**29.61**	**33.10**	**28.70***	25.81	**31.29**	**33.75**	**30.32**	**24.52**	**17.57**	**17.42**	**16.23**	**19.66**

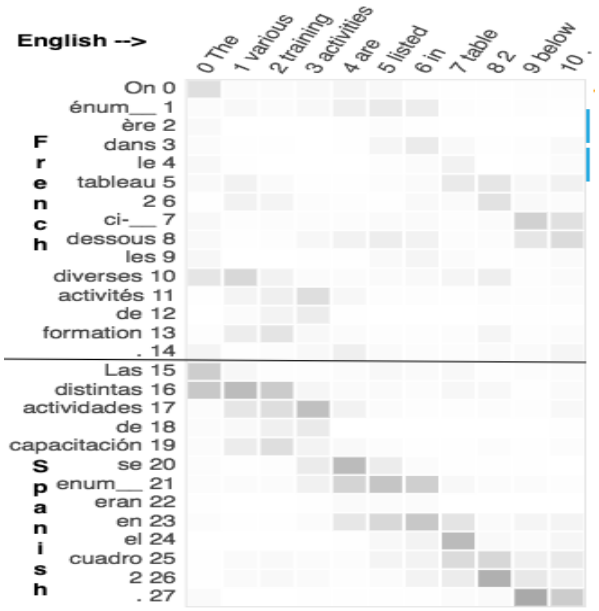

Fig. 4 Attention Visualization for UN corpus setting for French and Spanish to English. A horizontal black line is used to separate the source languages but the NMT system receives a single, long multi-source sentence.

nary by extracting the words from the source sentence that receive the highest attention for each target word generated.

6.2 Evaluation of Transfer Learning Using Multi-Source Models

Table 4 contains the results for the transfer learning experiments on the Europarl corpus. Regardless of the target language, there is a statistically significant improvement in BLEU using both the multi-source as well as the 4S1T models as parent languages. In a number of cases the multi-source model acts as a better parent than the 4S1T model.

German-English is the only language pair that fails to improve via transfer learning. We believe that this happens since German is different from the other source languages it was grouped with because French, Italian and Spanish are romance languages and German is not. In the future we plan to conduct experiments with various language families and verify whether grouping languages according to language families is beneficial to transfer learning or not.

It must be noted that we do not use any regularization by freezing parts of the model, as in Ref. [20], while training and hence the transferred model learns and overfits quickly. By using proper regularization methods we believe that we can obtain further improvements in the translation quality as a result of transfer learning.

In order to investigate why such transfer learning works well we investigated the learning curves of our various models. Consider the following lowest achieved per word development set losses:

- French+Spanish+Italian+German to English: 1.76
- 4S1T model (for French, Spanish, Italian, German to English): 2.17
- French-English: 2.31
- Spanish-English: 2.25
- Italian-English: 2.44
- German-English: 2.32

Moreover, we noticed that the single-source baseline exhibited a certain amount of overfitting which happens in low resource scenarios. However, the multi-source and 4S1T models did not overfit at all and could achieve significantly lower losses. This indicates that multiple input sentences and languages act as regularizers. Since a situation with low loss is an indicator that the decoder is able to predict target words much better than a situation with high loss, we feel that using the multi-source model, which has the least loss, helps in improving translation quality.

This shows that while large bilingual corpora can be used for transfer learning, there is a substantial amount of untapped potential in multilingual, multiway corpora. Large bilingual corpora with English as the target language might be abundant but large bilingual corpora with Hindi or Marathi as target languages are not as abundant and thus such multilingual, multiway corpora can be beneficial. We feel that our results could have some implications in the way one would develop corpora for low resource languages: Adding an additional language to a N-lingual corpus not only provides N additional bilingual corpora but also enables one to improve the translation quality of single-source translations for all languages.

6.3 Evaluation of Multilingual Dictionaries Extracted Using Multi-Source Models

6.3.1 Evaluation Procedure

We manually evaluated the bilingual and multilingual dictio-

naries generated for the 100 most frequent English words. The reference translations for these English words are obtained from Google translate which is completely reliable for single word translations for European languages. In order to make sure that our evaluation is as accurate as possible, we only considered references which were marked as "checked by the translate community" [*19]. We report the 1-best, 2-best and 5-best accuracies for the same. A multilingual dictionary is a collection of N-tuples and an N-tuple counts towards the top 1 accuracy if the topmost entries for each of the bilingual dictionaries is correct. A valid example of a 5 tuple is (Mr, Señor (Spanish), Monsieur (French), Herr (German), Signor (Italian)). A 5 tuple counts towards the top 5 accuracy (but not the top 1 accuracy) if the valid translation of the English word in any of the languages is fifth highest entry (according to frequency) in the respective bilingual dictionary.

6.3.2 Observations

Table 5 contains the top 1, top 2 and top 5 accuracies for English-XX bilingual dictionary (where XX is one of Italian, French, Spanish and German). We also give the same accuracies for a 5 tuple dictionary (a multilingual dictionary entry) for the 5 languages involved.

One interesting point to note is that although the BPE subword vocabulary size we chose for English is 12,000 the total number of dictionary entries we obtained was around 55,000 which means that our method is able to successfully reconstruct surface words from subwords. As can be seen in Table 5, despite the simple approach, the quality of the bilingual dictionaries extracted for the 100 most frequent words is reasonably high (all 85% and above for top 1 accuracy and above 90% for the top 2 accuracy). Moreover the top 1 accuracy for the 5 lingual (multilingual) dictionary is 74%.

Following are some examples of multilingual dictionary entries not in the list of 100 entries we evaluated:

- ignorance (English), ignorancia (Spanish), ignorare (italian), unkenntnis [*20] (German), ignorance (French)
- college (English), colegio (Spanish), college (French), kollegium (German), collegio (Italian)
- Moreira (English), Moreira (Italian), Moreira (French), Moreira (German)

The most interesting thing we noticed was that although all the words above were segmented into 2 to 3 subword units after BPE segmentation our method managed to correctly generate and align the surface forms. For example: Moreira is split as "Mor__ ei__ ra" and appears only 7 times in the corpus of 200,000 lines. Similarly, unkenntnis is split as "unk__ enn__ tnis" and occurs only 14 times. Our method manages to correctly align proper names in most cases we investigated despite their infrequent occurrences. This leads us to believe that our approach will definitely allow for high quality dictionary entries for rare words as well.

We believe that further modifications to our algorithm and appropriate post processing techniques will leads to even higher accuracies. The next step will be the evaluation of dictionaries for rare words which we leave as future work but we expect reasonably high quality dictionaries.

Table 5 The results of the evaluation of a dictionary extracted using the method in Section 4.3. We give the top 1, 2 and 5 accuracies for the bilingual English-XX and 5 lingual dictionaries extracted for the 100 most freuqent words in the Europarl corpus. The languages involved are English (En), French (Fr), German (De), Italian (It) and Spanish (Es).

Language Pair	Accuracy		
	Top 1	Top 2	Top 5
It-En	85	93	93
Es-En	86	91	94
Fr-En	96	99	100
De-En	95	98	99
5 lingual	74	82	87

[*19] https://translate.google.co.in/#en/es/eat

[*20] This was the Top 2 entry. The Top 1 entry was Geschichtliche which is wrong

7. Conclusion and Future Work

In this paper, we explore a simple approach for "Multi-Source Neural Machine Translation" that can be used with any single-source NMT system seen as a black-box. We evaluate it in a resource poor as well as a resource rich setting using the ILCI and UN corpora. We compare our approach with two other previously proposed approaches and show that it gives competitive results with other state of the art methods while using less than half the number of parameters (for 2 source models). It is domain and language independent and the gains are significant. We also observe, by visualizing attention, that NMT focuses on some languages by practically ignoring others indicating that language relatedness is one of the aspects that should be considered in a multilingual MT scenario. Finally, we explore how multilingual, multiway corpora can be leveraged for improving single-source translation quality by using transfer learning. This points to unexpected advantages in developing multiway corpora for low resource languages. We also propose a simple method for the extraction of dictionaries using the multi-source model and evaluated the dictionaries extracted. We show that the dictionaries obtained are of sufficiently high quality despite the limitations of using attention for word alignment purposes.

In the future we plan on exploring the language relatedness phenomenon by considering even more languages. We also plan on exploring approaches to train models that can translate both single and multi-source models as well as multilingual dictionaries for rare words for a variety of languages.

References

[1] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *Proc. 3rd International Conference on Learning Representations* (*ICLR 2015*) (2015).

[2] Cettolo, M., Girardi, C. and Federico, M.: WIT^3: Web Inventory of Transcribed and Translated Talks, *Proc. 16th Conference of the European Association for Machine Translation* (*EAMT*), pp.261–268 (2012).

[3] Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, *Proc. 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.1724–1734, Association for Computational Linguistics (2014) (online), available from ⟨http://www.aclweb.org/anthology/D14-1179⟩.

[4] Christodouloupoulos, C. and Steedman, M.: A massively parallel corpus: The Bible in 100 languages, *Language Resources and Evaluation*, Vol.49, No.2, pp.375–395 (online), DOI: 10.1007/s10579-014-9287-y (2015).

[5] Cromieres, F., Chu, C., Nakazawa, T. and Kurohashi, S.: Kyoto University Participation to WAT 2016, *Proc. 3rd Workshop on Asian*

Translation (*WAT2016*), pp.166–174, The COLING 2016 Organizing Committee (2016) (online), available from ⟨http://aclweb.org/anthology/W16-4616⟩.

[6] Firat, O., Sankaran, B., Al-Onaizan, Y., Yarman-Vural, F.T. and Cho, K.: Zero-Resource Translation with Multi-Lingual Neural Machine Translation, *Proc. 2016 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2016*), pp.268–277 (2016) (online), available from ⟨http://aclweb.org/anthology/D/D16/D16-1026.pdf⟩.

[7] Garmash, E. and Monz, C.: Ensemble Learning for Multi-Source Neural Machine Translation, *Proc. COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp.1409–1418, The COLING 2016 Organizing Committee (2016) (online), available from ⟨http://aclweb.org/anthology/C16-1133⟩.

[8] Jha, G.N.: The TDIL Program and the Indian Langauge Corpora Intitiative (ILCI), *Proc. 7th International Conference on Language Resources and Evaluation* (*LREC '10*), Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M. and Tapias, D. (Eds.), European Language Resources Association (ELRA) (2010).

[9] Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F.B., Wattenberg, M., Corrado, G., Hughes, M. and Dean, J.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, *CoRR*, Vol.abs/1611.04558 (2016) (online), available from ⟨http://arxiv.org/abs/1611.04558⟩.

[10] Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation, *Conference Proceedings: The 10th Machine Translation Summit*, pp.79–86 AAMT, AAMT, (2005) (online), available from ⟨http://mt-archive.info/MTS-2005-Koehn.pdf⟩.

[11] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *ACL*, The Association for Computer Linguistics (2007).

[12] Kunchukuttan, A. and Bhattacharyya, P.: Orthographic Syllable as basic unit for SMT between Related Languages, *Proc. 2016 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2016*), pp.1912–1917 (2016) (online), available from ⟨http://aclweb.org/anthology/D/D16/D16-1196.pdf⟩.

[13] Niehues, J., Cho, E., Ha, T. and Waibel, A.: Pre-Translation for Neural Machine Translation, *COLING 2016, 26th International Conference on Computational Linguistics, Proc. Conference: Technical Papers*, pp.1828–1836 (2016) (online), available from ⟨http://aclweb.org/anthology/C/C16/C16-1172.pdf⟩.

[14] Och, F.J. and Ney, H.: Statistical multi-source translation, *Proc. MT Summit*, Vol.8, pp.253–258 (2001).

[15] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation, *Proc. 40th Annual Meeting on Association for Computational Linguistics* (*ACL '02*), pp.311–318, Association for Computational Linguistics (online), DOI: 10.3115/1073083.1073135 (2002).

[16] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proc. 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp.1715–1725, Association for Computational Linguistics (2016) (online), available from ⟨http://www.aclweb.org/anthology/P16-1162⟩.

[17] Sutskever, I., Vinyals, O. and Le, Q.V.: Sequence to Sequence Learning with Neural Networks, *Proc. 27th International Conference on Neural Information Processing Systems* (*NIPS '14*), pp.3104–3112, MIT Press (2014) (online), available from ⟨http://dl.acm.org/citation.cfm?id=2969033.2969173⟩.

[18] Ziemski, M., Junczys-Dowmunt, M. and Pouliquen, B.: The United Nations Parallel Corpus v1.0, *Proc. 10th International Conference on Language Resources and Evaluation* (*LREC 2016*) (Chair), Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (Eds.), European Language Resources Association (ELRA) (2016).

[19] Zoph, B. and Knight, K.: Multi-Source Neural Translation, *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.30–34, Association for Computational Linguistics (2016) (online), available from ⟨http://www.aclweb.org/anthology/N16-1004⟩.

[20] Zoph, B., Yuret, D., May, J. and Knight, K.: Transfer Learning for Low-Resource Neural Machine Translation, *Proc. 2016 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2016*), pp.1568–1575 (2016) (online), available from ⟨http://aclweb.org/anthology/D/D16/D16-1163.pdf⟩.

Raj Dabre was born in 1989. He is a 3rd year Ph.D. student in Kyoto University. His research interest is Neural Machine Translation and Deep Learning in general. He is a member of ACL.

Fabien Cromieres was born in 1978. He received his Ph.D. from Kyoto University and is currently a researcher in Japan Science and Technology Agency. His research interest is Natural Language Processing, especially Neural Machine Translation. He is a member of ACL.

Sadao Kurohashi was born in 1966. He received his Ph.D. from Kyoto University and is currently a professor in Kyoto University. His interests are in Fundamental Analysis of Japanese language, Knowledge Representation, Question Answering, Semantic Analysis, Discourse and Machine Translation. He is a member of ANLP, IPSJ, JSAI, IEIEC, ACL and ACM.