Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder

Thanh-Le Ha, Jan Niehues, Alex Waibel

Institute for Anthropomatics and Robotics
KIT - Karlsruhe Institute of Technology, Germany
firstname.lastname@kit.edu

Abstract

In this paper, we present our first attempts in building a multilingual Neural Machine Translation framework under a unified approach in which the information shared among languages can be helpful in the translation of individual language pairs. We are then able to employ attention-based Neural Machine Translation for many-to-many multilingual translation tasks. Our approach does not require any special treatment on the network architecture and it allows us to learn minimal number of free parameters in a standard way of training. Our approach has shown its effectiveness in an under-resourced translation scenario with considerable improvements up to 2.6 BLEU points. In addition, we point out a novel way to make use of monolingual data with Neural Machine Translation using the same approach with a 3.15-BLEU-score gain in IWSLT'16 English→German translation task.

1. Introduction

Neural Machine Translation (NMT) has shown its effectiveness in translation tasks when NMT systems perform best in recent machine translation campaigns [1, 2]. Compared to phrase-based Statistical Machine Translation (SMT) which is basically an ensemble of different features trained and tuned separately, NMT directly modeling the translation relationship between source and target sentences. Unlike SMT, NMT does not require large monolingual data to achieve good performances.

An NMT system consists of an encoder which recursively reads and represents the whole source sentence into a context vector and a recurrent decoder which takes the context vector and its previous state to predict the next target word. It is then trained in an end-to-end fashion to learn parameters which maximizes the likelihood between the outputs and the references. Recently, attention-based NMT has been featured in most state-of-the-art systems. First introduced by [3], attention mechanism is integrated in decoder side as feedforward layers. It allows the NMT to decide which source words should take part in the prediction process of the next target words. It helps to improve NMTs significantly. Nevertheless, since the attention mechanism is specific to a particular source sentence and the considering target word, it is also specific to particular language pairs.

Some recent work has focused on extending the NMT framework to multilingual scenarios. By training such network using parallel corpora in number of different languages, NMT could benefit from additional information embedded in a common semantic space across languages. Basically, the proposed NMT are required to employ multiple encoders or multiple decoders to deal with multilinguality. Furthermore, in order to avoid the tight dependency of the attention mechanism to specific language pairs, they also need to modify their architecture to combine either the encoders or the attention layers. These modifications are specific to the purpose of the tasks as well. Thus, those multilingual NMTs are more complicated, much more free parameters to learn and more difficult to perform standard trainings compared to the original NMT.

In this paper, we introduce a unified approach to seamlessly extend the original NMT to multilingual settings. Our approach allows us to integrate any language in any side of the encoder-decoder architecture with only one encoder and one decoder for all the languages involved. Moreover, it is not necessary to do any network modification to enable attention mechanism in our multilingual NMT systems. We then apply our proposed framework in a demanding scenarios: (simulated) under-resourced translation. The results show that bringing multilinguality to NMT helps to improve individual translations. With some insightful analyses of the results, we set our goal toward a fully multilingual NMT framework.

The paper starts with a detailed introduction to attention-based NMT. In Section 3.1, related work about multi-task NMT is reviewed. Section 3.2 describes our proposed approach and thorough comparisons to the related work. It is followed by a section of evaluating our systems in the aforementioned scenario, in which different strategies have been employed under a unified approach (Section 4). Finally, the paper ends with conclusion and future work.

2. Background

An NMT system consists of an encoder which automatically learns the characteristics of a source sentence into fix-length context vectors and a decoder that recursively combines the produced context vectors with the previous target word to generate the most probable word from a target vocabulary.

More specifically, a bidirectional recurrent encoder reads every words x_i of a source sentence $\boldsymbol{x} = \{x_1, ..., x_n\}$ and encodes a representation $\boldsymbol{s}$ of the sentence into a fixed-length vector $\boldsymbol{h}_i$ concatenated from those of the forward and backward directions:

$$\begin{aligned} \boldsymbol{h}_i &= [\overrightarrow{\boldsymbol{h}}_i, \overleftarrow{\boldsymbol{h}}_i] \\ \overrightarrow{\boldsymbol{h}}_i &= f(\overrightarrow{\boldsymbol{h}}_{i-1}, \boldsymbol{s}) \\ \overleftarrow{\boldsymbol{h}}_i &= f(\overleftarrow{\boldsymbol{h}}_{i+1}, \boldsymbol{s}) \\ \boldsymbol{s} &= \boldsymbol{E}_s \bullet \boldsymbol{x}_i \end{aligned}$$

Here $\boldsymbol{x}_i$ is the one-hot vector of the word x_i and $\boldsymbol{E}_s$ is the word embedding matrix which is shared across the source words. f is the recurrent unit computing the current hidden state of the encoder based on the previous hidden state. $\boldsymbol{h}_i$ is then called an *annotation vector*, which encodes the source sentence up to the time i from both forward and backward directions. Recurrent units in NMT can be a simple recurrent neural network unit (RNN), a Long Short-Term Memory unit (LSTM) [4] or a Gated Recurrent Unit (GRU) [5]

Similar to the encoder, the recurrent decoder generates one target word y_j to form a translated target sentence $\boldsymbol{y} = \{y_1, ..., y_m\}$ in the end. At the time j, it takes the previous hidden state of the decoder $\boldsymbol{z}_{j-1}$, the previous embedded word representation $\boldsymbol{t}_{j-1}$ and a time-specific context vector $\boldsymbol{c}_j$ as inputs to calculate the current hidden state $\boldsymbol{z}_j$:

$$\begin{aligned} \boldsymbol{z}_j &= g(\boldsymbol{z}_{j-1}, \boldsymbol{t}_{j-1}, \boldsymbol{c}_j) \\ \boldsymbol{t}_{j-1} &= \boldsymbol{E}_t \bullet \boldsymbol{y}_{j-1} \end{aligned}$$

Again, g is the recurrent activation function of the decoder and $\boldsymbol{E}_t$ is the shared word embedding matrix of the target sentences. The context vector $\boldsymbol{c}_j$ is calculated based on the annotation vectors from the encoder. Before feeding the annotation vectors into the decoder, an *attention mechanism* is set up in between, in order to choose which annotation vectors should contribute to the predicting decision of the next target word. Intuitively, a relevance between the previous target word and the annotation vectors can be used to form some attention scenario. There exists several ways to calculate the relevance as shown in [6], but what we describe here follows the proposed method of [3]

$$rel(\boldsymbol{z}_{j-1}, \boldsymbol{h}_i) = \boldsymbol{v}_a \bullet \tanh(\boldsymbol{W}_a \bullet \boldsymbol{z}_{j-1} + \boldsymbol{U}_a \bullet \boldsymbol{h}_i)$$

$$\alpha_{ij} = \frac{\exp(rel(\boldsymbol{z}_{j-1}, \boldsymbol{h}_i))}{\sum_{i'} \exp(rel(\boldsymbol{z}_{j-1}, \boldsymbol{h}_{i'}))}, \ \boldsymbol{c}_j = \sum_i \alpha_{ij} \boldsymbol{h}_i$$

In [3], this attention mechanism, originally called *alignment model*, has been employed as a simple feedforward network with the first layer is a learnable layer via adaptation factors $\boldsymbol{v}_a$,$\boldsymbol{W}_a$ and $\boldsymbol{U}_a$. The relevance scores rel are then normalized into attention weights α_{ij} and the context vector $\boldsymbol{c}_j$ is calculated as the weighted sum of all annotation vectors $\boldsymbol{h}_i$. Depending on how much attention the target word at time j put on the source states $\boldsymbol{h}_i$, a soft alignment is learned. By being employed this way, word alignment is not a latent variable but a parametrized function, making the alignment model differentiable. Thus, it could be trained together with the whole architecture using backpropagation.

One of the most severe problems of NMT is handling of the rare words, which are not in the short lists of the vocabularies, i.e. out-of-vocabulary (OOV) words, or do not appear in the training set at all. In [7], the rare target words are copied from their aligned source words after the translation. This heuristic works well with OOV words and named entities but unable to translate unseen words. In [8], their proposed NMT models on the preprocessed data using Byte-Pair Encoding have been shown to not only be effective on reducing vocabulary sizes but also have the ability to generate unseen words. This is achieved by segmenting the rare words into subword units using the robust, unsupervised compressing method Byte-Pair Encoding in the preprocessing phase. Then one could use a normal NMT system to translating those subword units. The state-of-the-art translation systems essentially employ subword NMT [8].

3. Multilingual Neural Machine Translation

While the majority of previous research has focused on improving the performance of NMT on individual language pairs with individual NMT systems, recent work has started investigating potential ways to conduct the translation involved in multiple languages using a single NMT system. The possible reason explaining these efforts lies on the unique architecture of NMT. Unlike SMT, NMT consists of separated neural networks for the source and target sides, or the encoder and decoder, respectively. This allows these components to map a sentence in any language to a representation in an embedding space which is believed to share common semantic among the source languages involved[1]. From that shared space, the decoder, with some implicit or explicit relevant constraints, could transform the representation into a concrete sentence in any desired language. In this section, we review some related work on this matter. We then describe a unified approach toward a universal attention-based NMT scheme. Our approach does not require any architecture modification and it can be trained to learn a minimal number of parameters compared to the other work.

3.1. Related Work

By extending the solution of sequence-to-sequence modeling using encoder-decoder architectures to multi-task learning,

[1] But not necessarily syntactic since the embeddings are learned from parallel sentences which essentially share the same meaning although they might be very different in word order

[9] managed to achieve better performance on some *many-to-many* tasks such as translation, parsing and image captioning compared to individual tasks. Specifically in translation, the work utilizes multiple encoders to translate from multiple languages, and multiple decoders to translate to multiple languages. In this view of multilingual translation, each language in source or target side is modeled by one encoder or decoder, depending on the side of the translation. Due to the natural diversity between two tasks in that multi-task learning scenario, e.g. translation and parsing, it could not feature the attention mechanism although it has proven its effectiveness in NMT.

There exists two directions which are proposed the ways to leverage attention mechanism for multilingual translation scenarios. The first one is indicated in the work from [10], where it introduce an *one-to-many* multilingual NMT system to translates from one source language into multiple target languages. Having one source language, the attention mechanism is then handed over to the corresponding decoder. The objective function is changed to adapt to multilingual settings. In testing time, the parameters specific to a desired language pair are used to perform the translation.

[11] proposed another approach which genuinely delivers attention-based NMT to multilingual translation. As in [9], their approach utilizes one encoder per source language and one decoder per target language for *many-to-many* translation tasks. Instead of a quadratic number of independent attention layers, however, one single attention mechanism is integrated into their NMT, performing an affine transformation between the hidden layer of m source languages and that one of n target languages. It is required to change their architecture to accommodate such a complicated shared attention mechanism.

In a separate effort to achieve multilingual NMT, the work of [12] leverages available parallel data from other language pairs to help reducing possible ambiguities in the translation process into a single target language[2]. They employed the multi-source attention-based NMT in a way that only one attention mechanism is required despite having multiple encoders. To achieve this, the outputs of the encoders were combined before feeding to the attention layer. They implemented two types of encoder combination; One is adding a non-linear layer on the concatenation of the encoders' hidden states. The other is using a variant of LSTM taking the respective gate values from the individual LSTM units of the encoders. As a result, the combined hidden states contain information from both encoders , thus encode the common semantic of the two source languages.

3.2. Universal Encoder and Decoder

Inspired by the multi-source NMT as additional parallel data in several languages are expected to benefit single translations, we aim to develop a NMT-based approach toward a universal framework to perform multilingual translation. Our solution is to perform some coding on the same word (meaning) in different languages to differentiate them as different words in the language-mixed vocabularies. The concrete idea, referred to as Language-specific Coding, is described in the following.

Language-specific Coding. When the encoder of a NMT system considers words across languages as different words, with a well-chosen architecture, it is expected to be able to learn a good representation of the source words in an embedding space in which words carrying similar meaning would have a closer distance to each others than those are semantically different. This should hold true when the words have the same or similar surface form, such as (@***de***@*Obama*; @***en***@*Obama*) or (@***de***@*Projektion*; @***en***@*projection*)[3]. This should also hold true when the words have the same or similar meaning across languages, such as (@**en**@*car*; @***en***@*automobile*) or (@***de***@*Flussufer*; @***en***@*bank*). In this way, the same words in different languages are treated as synonyms in one language or as the words having the same meaning across languages. Our encoder then acts similarly to the one of multi-source approach[12], collecting additional information from other sources for better translations, but with a much simpler embedding function. Unlike them, we need only one encoder, so we could reduce the number of parameters to learn. Furthermore, we neither need to change the network architecture nor depend on which recurrent unit (GRU, LSTM or simple RNN) is currently using in the encoder.

We could apply the same trick to the target sentences and thus enable *many-to-many* translation capability of our NMT system. Similar to the multi-target translation[10], we exploit further the correlation in semantics of those target sentences across different languages. The main difference between our approach and the work of [10] is that we need only one decoder for all target languages. Given one encoder for multiple source languages and one decoder for multiple target languages, it is trivial to incorporate the attention mechanism as in the case of a regular NMT for single language translation. In training, the attention layers were directed to learn relevant alignments between words in specific language pair and forward the produced context vector to the decoder. Now we rely totally on the network to learn good alignments between source and target sides.

In comparison to other research that could perform complete multi-task learning, e.g. the work from [9] or the approach proposed by [11], our method is able to accommodate the attention layers seamlessly. It also draws a clear distinction from those works in term of the complexity of the whole network: considerably less parameters to learn, thus reduces overfitting, with a conventional attention mechanism and a

[2] An example taken from the paper is when we want to translate the English word *bank* into French, it might be easier if we have an additional German sentence containing the word *Flussufer* (*river bank*).

[3] @***lang_code***@*a_word* is a simple way that transforms the word *a_word* into a different surface form associated with its language ***lang_code***. For example, @***de***@*Projektion* is referred to the word *Projektion* appearing in a German (***de***) sentence.

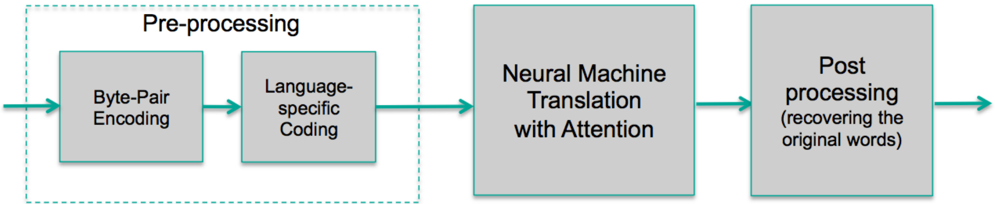

Pre-processing
Byte-Pair Encoding
Language-specific Coding
Neural Machine Translation with Attention
Post processing (recovering the original words)

Figure 1: *Preprocessing steps to employ a multilingual attention-based NMT system*

standard training procedure.

Figure 1 illustrates the essence of our approach. With the deployment of language-specific coding in the preprocessing phase, we are able to employ multilingual attention-based NMT without any special treatment in training such a standard architecture. Other preprocessing steps, for example, applying Byte-Pair Encoding (BPE) to address rare-word problem, are done as in usual NMT systems. Our encoder and attention-enable decoder can be seen as a shared encoder and decoder across languages, or a *universal* encoder and decoder. The flexibility of our approach allow us to integrate any language into source or target side and choose which kind of translation units (words, subwords or characters) to be used. As we will see in Section 4, it has proven to be extremely helpful not only in low-resourced scenarios but also in translation of well-resourced language pairs as it provides a novel way to make use of large monolingual corpora.

4. Evaluation

In this section, we describe the evaluation of our proposed approach in comparisons with the strong baselines in a simulated under-resourced scenario. We believe that in an under-resourced situation, the benefit of having more data in different languages can be observed more clearly, hence emphasizes the effectiveness of our approach. Nevertheless, later we show that our approach can also help to bring noticeable improvements in well-resourced translation.

4.1. Experimental Settings

Training Data. We choose WIT3's[4] TED corpus[13] as the basis of our experiments since it might be the only high-quality parallel data of many low-resourced language pairs. TED is also multilingual in a sense that it includes numbers of talks which are commonly translated into many languages. In addition, we use a much larger corpus provided freely by WMT organizers[5] when we evaluate the impact of our approach in a real machine translation campaign. It includes the parallel corpus extracted from the digital corpus of European Parliament (EPPS), the News Commentary (NC) and the web-crawled parallel data (CommonCrawl). While the number of sentences in popular TED corpora varies from 16 thousands to 20 thousands, the total number of sentences in those larger corpora is approximately 3 million sentences.

Neural Machine Translation Setup. All experiments here have been conducted using the NMT framework `Nematus`[6]. Following the work of [8], subword segmentation is handled in the preprocessing phase using Byte-Pair Encoding (BPE). Excepts stated clearly in some experiments, we set the number of BPE merging operations at 39500 on the joint of source and target data. When training all NMT systems, we take out the sentence pairs exceeding 50-word length and shuffle them inside every minibatch. Our shortlist vocabularies contain 40,000 most frequent words while the others are considered as rare words and applied the subword translation. We use an 1024-cell GRU layer and 1000-dimensional embeddings with dropout at every layer with the probability of 0.2 in the embedding and hidden layers and 0.1 in the input and output layers. We trained our systems using gradient descent optimization with Adadelta[14] on minibatches of size 80 and the gradient is rescaled whenever its norm exceed 1.0. All the trainings last approximately seven days if the early-stopping condition could not be reached. At a certain time, an external evaluation script on BLEU[15] is conducted on a development set to decide the early-stopping condition. This evaluation script has also being used to choose the model achieving the best BLEU on the development set instead of the maximal loglikelihood between the translations and target sentences while training. In translation, the framework produces n-best candidates and we then use a beam search with the beam size of 12 to get the best translation.

Language-specific coding and post processing. As mentioned in the Section 3.2, we performed language-specific coding in the preprocessing phase, after other preprocessing steps. It does not affect to which kind of translation units to be used since we apply it on the already-segmented texts. In our case, we trained the BPE first, and performed the language-specific coding on the corpus and the test sets after applying BPE on them. So here the multilinguality is featured on the subword level. In the post processing phase, we did the steps in the reversed order: first remove the language-specific codes and then remove the BPE tags to recover the word-based translated sentence in the desired language.

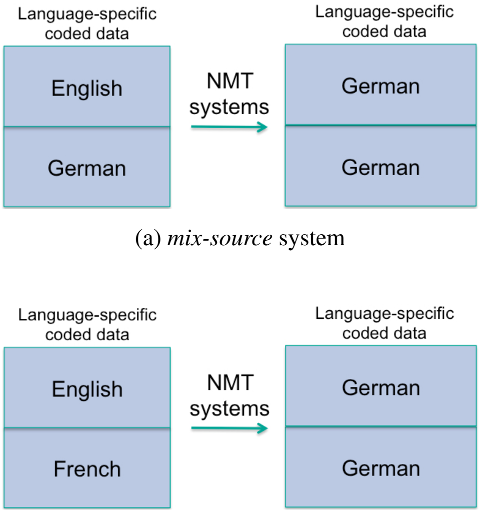
[4] `https://wit3.fbk.eu/`

[5] `http://www.statmt.org/wmt15/`

[6] `https://github.com/rsennrich/nematus`

4.2. Under-resourced Translation

First, we consider the translation for an under-resourced pair of languages. Here *a small portion* of the available, large parallel corpus for English-German is used as *a simulation* for the scenario where we do not have much parallel data. When we assume there is no large parallel corpus by *using only* the TED corpus, we can *simulate* the translation task from English to German as *an under-resourced scenario*. The reason that we chose a simulated under-resourced language pair but not a real one is to have the comparable improvements with the well-resourced scenario using the same proposed approach. We perform language-specific coding in both source and target sides. By accommodating the German monolingual data (the target language) as an additional input (German→German), which we called the *mix-source* approach, we could enrich the training data in a simple, natural way. Given this under-resourced situation, it could help our NMT obtains a better representation and more information of the source side, hence, able to learn the translation relationship better. Including monolingual data in this way might also improve the translation of some rare-word types such as named entities. Furthermore, as the ultimate goal of our work, we would like to investigate the advantages of multilinguality in NMT. We incorporate a similar portion of French-German parallel corpus, which includes also TED talks into the English-German one. As discussed in Section 3.2, it is expected to help reducing the ambiguity in translation between one language pair since it utilizes the semantic context provided by the other source language. We name this *multi-source*. Figure 2 shows those two strategies.


Language-specific coded data
English
German
NMT systems
Language-specific coded data
German
German
Language-specific coded data
English
French
NMT systems
Language-specific coded data
German
German

(a) *mix-source* system

(b) *multi-source* system

Figure 2: Different strategies of multi-source NMT

Table 1 summarizes the performance of our systems measured in BLEU[7] on two test sets, *tst2013* and *tst2014*. Compared to the baseline NMT system which is solely trained on TED English-German data, our *mix-source* system achieves a considerable improvement of 2.6 BLEU points on *tst2013* and 2.1 BLEU points on and *tst2014*. Adding French data to the source side and their corresponding German data to the target side in our *multi-source* system also help to gain 2.2 and 1.6 BLEU points more on *tst2013* and *tst2014*, respectively. We observe a better improvement from our *mix-source* system compared to our *multi-source* system. Due to the mismatch between the number of sentences in our English-German and French-German corpora (196k sentence pairs versus 165k sentence pairs), one reason of the less effective *multi-source* versus the *mix-source* might be because we have a little bigger corpus in case of the *mix-source*. We verified this hypothesis by training *mix-source* with the same size of the *multi-source* corpus when we took the German side of the French-German as the additional input instead of the German side of the English-German. It achieved even better BLEU scores compared to the original *mix-source* approach (*mix-source 2*, table 1). So it is not because of the size of the training corpus.

We speculate the reason that in the *mix-source*, the encoder utilizes the same information shared in two languages, while the *multi-source* receives and processes similar information in the other language but not necessarily the same, because the French-German corpus shares some common TED talks with the English-German corpus but not all. We might verify this hypothesis by comparing two systems trained on a common English-German-French corpus of TED or even more languages involved. We put it in our future work's plan.

4.3. Using large monolingual data

A standard NMT system employs parallel data only. While good parallel corpora are limited in number, getting monolingual data of an arbitrary language is trivial. To make use of German monolingual data in an English→German NMT system, [16] built a separate German→English NMT using the same parallel corpus, then they used that system to translate the German monolingual data back to English, forming a synthesis parallel data. [17] trained another RNN-based language model to score the monolingual corpus and integrate it to the NMT system through shallow or deep fusion. Both methods requires to train separate systems with possibly different hyperparameters for each. Conversely, by applying *mix-source* method to the big monolingual data, we need to train only one network. In this scenario, we use all the parallel data we have been provided by the organizer of the IWSLT16 machine translation campaign[8], which is around 200 times larger than the TED corpus for English-German. The baseline system is the strong NMT trained with all the parallel data. We then mix the large parallel corpus with the larger monolingual corpus using *mix-source* strategy.

[7] We used the script `mteval-v13a.pl` of the Moses framework (`http://statmt.org/moses/`) as the official way to calculate BLEU scores in main machine translation campaigns.

[8] `http://workshop2016.iwslt.org/`

System	**tst2013**		**tst2014**	
	BLEU	ΔBLEU	BLEU	ΔBLEU
Baseline (En→De)	24.35	–	20.62	–
Mix-source (En,De→De,De)	26.99	+2.64	22.71	+2.09
Multi-source (En,Fr→De,De)	26.64	+2.21	22.21	+1.59
Mix-source 2 (En,De→De,De)	27.18	+2.83	23.74	+3.12

Table 1: *Results of the English→German systems in a simulated under-resourced scenario.*

System	**tst2013**		**tst2014**	
	BLEU	ΔBLEU	BLEU	ΔBLEU
Baseline (En→De)	25.74	–	22.54	–
(1) Mix-source additional monolingual (En,De→De,De)	27.74	+2.00	24.39	+1.85
(2) Mix-source target monolingual part (En,De→De,De)	28.89	+3.15	24.86	+2.32

Table 2: *Results of the English→German system using large monolingual data.*

The first result (1) is encouraging when the *mix-source* using additional monolingual data achieves the improvements on *tst2013 tst2014* nearly as good as in the under-resourced situation (2.00 and 1.85 BLEU scores, respectively). Not using the same information in the source side, as we discussed in case of *multi-source* strategy, could explain why less improvements are observed in performance of such a system. As a quick attempt, we train the *mix-source* model on only the target-side monolingual part of the parallel data (2), which refers to the same information in the source side. Although using a smaller corpus, it brings a big improvement of 3.15 BLEU scores on *tst2013* and 2.32 BLEU on *tst2014* (Table 2). It is included as one of the systems in our English-German NMT combination submitted to the IWSLT'16 evaluation campaign[18].

5. Conclusion and Future Work

In this paper, we present our first attempts in building a multilingual Neural Machine Translation framework. By treating words in different languages as different words and leveraging the disambiguation abitily of such a setting, we are able to employ attention-enable NMT toward a multilingual translation system. Our proposed approach alleviates the need of complicated architecture re-designing when accommodating attention mechanism. In addition, the number of free parameters to learn in our network does not go beyond that magnitude of a single NMT system. With its universality, our approach has shown its effectiveness in an under-resourced scenario with considerable improvements. In addition, the approach has achieved great results when applied to the monolingual data in a well-resourced task.

Nevertheless, there are issues that we can continue working on in future work. We could conduct more experiments in a real under-resourced scenario instead of a simulated one. Our approach could be applied in a zero-resourced scenario where we do not have any direct parallel corpus. We could perform detailed analyses of the various strategies under the framework to show the advantages of our approach compared to other work on multilingual NMT.

6. Acknowledgements

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 645452. This work was supported by the Carl-Zeiss-Stiftung. The research by Thanh-Le Ha was supported by Ministry of Science, Research and Arts, Baden-Württemberg.

7. References

[1] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico, "The IWSLT 2015 Evaluation Campaign," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, Danang, Vietnam, 2015.

[2] O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, *et al.*, "Findings of the 2016 Conference on Machine Translation (WMT16)," in *Proceedings of the First Conference on Machine Translation (WMT16)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 12–58.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[5] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Repre-

sentations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8*. Baltimore, ML, USA: Association for Computational Linguistics, Jule 2014.

[6] M.-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 (EMNLP 15*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1412–1421. [Online]. Available: http://aclweb.org/anthology/D15-1166

[7] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," in *Proceedings of ACL-IJNLP 2015*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 11–19. [Online]. Available: http://www.aclweb.org/anthology/P15-1002

[8] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Association for Computational Linguistics (ACL 2016)*, Berlin, Germany, August 2016.

[9] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," in *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, May 2016.

[10] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-Task Learning for Multiple Language Translation," in *Proceedings of ACL-IJNLP 2015*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 1723–1732. [Online]. Available: http://www.aclweb.org/anthology/P15-1166

[11] O. Firat, K. Cho, and Y. Bengio, "Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism," *CoRR*, vol. abs/1601.01073, 2016. [Online]. Available: http://arxiv.org/abs/1601.01073

[12] B. Zoph and K. Knight, "Multi-Source Neural Translation," in *The North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, San Diego, CA, USA, June 2016.

[13] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[14] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *CoRR*, vol. abs/1212.5701, 2012. [Online]. Available: http://arxiv.org/abs/1212.5701

[15] 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 on Association for Computational Linguistics (ACL 2002)*. Association for Computational Linguistics, 2002, pp. 311–318.

[16] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Association for Computational Linguistics (ACL 2016)*, Berlin, Germany, August 2016.

[17] Ç. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On Using Monolingual Corpora in Neural Machine Translation," *CoRR*, vol. abs/1503.03535, 2015. [Online]. Available: http://arxiv.org/abs/1503.03535

[18] E. Cho, J. Niehues, T.-L. Ha, M. Sperber, M. Mediani, and A. Waibel, "Adaptation and Combination of NMT Systems: The KIT Translation Systems for IWSLT 2016," in *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016) - To be appeared.*