Compositional Source Word Representations for Neural Machine Translation

Duygu Ataman
FBK, Trento, Italy
University of Trento, Italy
ataman@fbk.eu

Mattia A. Di Gangi
FBK, Trento, Italy
University of Trento, Italy
digangi@fbk.eu

Marcello Federico
MMT Srl, Trento, Italy
FBK, Trento, Italy
federico@fbk.eu

Abstract

The requirement for neural machine translation (NMT) models to use fixed-size input and output vocabularies plays an important role for their accuracy and generalization capability. The conventional approach to cope with this limitation is performing translation based on a vocabulary of sub-word units that are predicted using statistical word segmentation methods. However, these methods have recently shown to be prone to morphological errors, which lead to inaccurate translations. In this paper, we extend the source-language embedding layer of the NMT model with a bi-directional recurrent neural network that generates compositional representations of the source words from embeddings of character n-grams. Our model consistently outperforms conventional NMT with sub-word units on four translation directions with varying degrees of morphological complexity and data sparseness on the source side.

1 Introduction

Neural machine translation (NMT) has improved the state-of-the-art performance in machine translation of many languages (Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016). However, current NMT systems still suffer from poor performance in translating infrequent or unseen words, limiting their deployment for translating low-resource and morphologically-rich languages. This problem is mainly caused by the fundamental design of the model, which requires observing many examples of a word until its input representation (*i.e.* embedding) becomes effective. Moreover, the convention of limiting the input and output vocabularies to few tens of thousands of words to control the computational complexity of the model leads to coverage issues. In fact, a word can be translated only if an exact match of it is found in the vocabulary.

To cope with this well-known problem, several studies have suggested to redefine a new model vocabulary in terms of the interior orthographic units compounding the words, such as character n-grams (Costa-Jussa and Fonollosa, 2016; Lee et al., 2016; Luong and Manning, 2016) or statistically-learned sub-word units (Sennrich et al., 2016; Wu et al., 2016; Ataman et al., 2017). In spite of providing an ideal open vocabulary solution, the former set of approaches mostly failed to achieve competitive results. This might be related to the semantic ambiguity caused by solely relying on embeddings of character n-grams which are generally learned by disregarding any lexical context, hence, morphology. In fact, building a vocabulary of sub-word units for training the NMT model and performing translation based on sub-word embeddings has now become the prominent approach. However, many studies have shown that statistical word segmentation methods can break the morphological structure of words, leading to loss of semantic and syntactic information in the sentence and, consequently, inaccurate translations (Niehues et al., 2016; Ataman et al., 2017; Pinnis et al., 2017; Huck et al., 2017; Tamchyna et al., 2017). Principally, these solutions are unsupervised methods and can never reach the accuracy of morphological analyzers, which, on the other hand, are not available in every language and can-

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 31–40
Alacant, Spain, May 2018.

not provide sufficiently compact vocabularies for the large training sets typically used in NMT.

In order to increase the accuracy in translating rare and unseen words with NMT, in this paper, we propose to learn information about the *source language* morphology directly from the bilingual lexical context and use this information to compose word representations from a minimal set of input symbols. In addition to improving the quality of input word representations, our approach also aims at eliminating the necessity of using a separate and sub-optimal word segmentation step on the source language. The approach of learning word embeddings compositionally has recently been applied in language modeling and has found to be promising (Vania and Lopez, 2017). In this study, which extends (Ataman and Federico, 2018b)[1], we present and evaluate an approach for improving the source language input representations in NMT by augmenting the *embedding layer* with a *bi-directional recurrent neural network* (bi-RNN), which can learn compositional input word representations from embeddings of character n-grams. We compare our approach against conventional embedding-based representations of sub-word units learned from statistical word segmentation methods in official evaluation benchmarks, under low to medium resource conditions, by pairing English with four languages: Czech, German, Italian and Turkish, where each language represents a distinct morphological typology. The experimental findings show that our compositional input representations provide significantly and consistently better translation quality for rare and unknown words than the prominent sub-word embedding based NMT approaches in all language directions.

2 Neural Machine Translation

The NMT model we use in this paper (Sutskever et al., 2014) is based on the idea of predicting the conditional probability of translating a source sentence $x = (x_1, x_2, \dots x_m)$ of length m, into a target sentence $y = (y_1, y_2, \dots y_j \dots y_l)$ of length l, using the decomposition

$$p(y|x) = \prod_{j=1}^{l} p(y_j | y_{j-1}, .., y_0, x_m, .., x_1) \quad (1)$$

The model is trained by maximizing the log-likelihood of a training dataset consisting of parallel sentence pairs in two languages using stochastic gradient descent methods (Bottou, 2010) and the backpropagation through time (Werbos, 1990) algorithm .

The inputs of the model are one-hot vectors, which have a single bit set to 1 to identify a given word in the vocabulary. Each word vector is mapped to an embedding, a continuous representation of the word in a lower-dimensional but more dense space. Then, the *encoder*, a stacked bi-RNN, learns a distributed representation of the source sentence x in the form of m dense vectors corresponding to its hidden states. The output states of a stacked RNN encoder with L layers is computed using the following equations:

$$h_i^k = RNN(h_i^{k-1}, h_{i-1}^k) \quad (2)$$

where h_i^0 is the embedding of the input word i ($l = 1..L$ and $i = 1..m$). The output of the encoder is fed to the *decoder*, a unidirectional stacked RNN, in order to predict the target sentence y word by word. Each target word y_j is predicted by sampling from a word distribution computed from the previous target word y_{j-1}, the previous hidden state of the decoder, and the *source-context vector*, which is a linear combination of the encoder hidden states. The weights of each hidden state are dynamically computed by the *attention* model (Luong et al., 2015) on the basis of the current decoder hidden state h_t and the corresponding encoder hidden states $\bar{h}_s$. During the generation of each target word y_j, its probability is normalized via a softmax function.

The number of parameters used by the model are mainly defined by the sizes of the source and target vocabularies, which requires to use fixed-size vocabularies in order to control the computational complexity. However, this limitation creates an important bottleneck when translating from and to low-resource and morphologically-rich languages, due to the sparseness of the lexical distribution.

3 Related Work

In order to improve the translation accuracy of rare words in NMT, previous studies have pro-

[1] This paper extends (Ataman and Federico, 2018b) in four ways: with a new and more efficient implementation of the model, with experiments with deeper and wider NMT networks, with results on new translation directions and under significantly larger training data conditions, and by reporting results on sentences containing rare words.

posed several approaches which share the representations of word pieces among different words. These approaches include either engineering new NMT models that efficiently work at the character level, or performing a pre-processing step where words are segmented into smaller units using supervised or statistical tools before computing the NMT vocabulary.

3.1 Character-level NMT

The first set of statistical approaches that attempted to overcome the fixed-size vocabulary problem in NMT is based on the idea of constructing the translation model directly at the level of characters. Most of these approaches are based on the character-level language model of Kim et al. (2016), which uses convolutional and highway networks for transforming character embeddings into feature representations of sentence segments. Costa-Jussa and Fonollosa (2016) applied this approach to NMT for learning the source language input representations with a convolutional neural network while still maintaining the translation model as the same bi-RNN based encoder-decoder network (2016). Lee et al. (2016) further extended this approach to achieve fully character-level NMT by changing the decoder with a character-based one (Chung et al., 2016). Another approach that also implements fully character-level NMT based on convolutional neural networks is ByteNet (Kalchbrenner et al., 2016), which performs translation in linear time steps with respect to the source sentence length.

The main problem with these approaches is that they generally disregard lexical boundaries while learning distributed representations of the input units. Nevertheless, it is controversial whether semantics, and therefore morphology, can be modeled without maintaining a context defined at the lexical level. An additional drawback related to these methods resides in the increased sequence lengths caused by processing the sentences as sequences of characters, which also augments the computational cost despite the reduced complexity in the softmax layer. Moreover, using solely convolution cannot capture information about the relative position of each interior unit inside the word, which could provide important cues about their morphological roles. An earlier approach to character-level NMT was developed by Ling et al. (2015), which instead learns compositional input representations of words using two additional layers of bi-LSTMs in the source and target sides of the NMT model. The decoding is implemented using a softmax over the character vocabulary in the target language. Although this approach allows to maintain NMT at the lexical level, the overall computational complexity of the resulting model becomes too high to be deployed in practical tasks.

3.2 Unsupervised Word Segmentation

A more straight-forward and faster method to cope with the high computational complexity in NMT is to apply a statistical word segmentation method as a data pre-processing step before training the model. This step reduces the size of the corpus vocabulary to a maximum number of sub-word units. Although the original NMT model was designed to translate sequences of words, it is now common to perform NMT at the sub-lexical level based on input representations learned from a vocabulary of sub-word units. Indeed, learning embeddings of sub-word units which are more frequently observed in different lexical contexts allows to reduce the data sparseness and improve the quality of input representations (Ataman and Federico, 2018a). In this paper, we discuss two of such approaches: Byte-Pair Encoding (BPE) (Sennrich et al., 2016) and Linguistically-Motivated Vocabulary Reduction (LMVR) (Ataman et al., 2017).

Byte-Pair Encoding is originally a data compression algorithm which aims to minimize the length of a sequence of bytes by finding the most frequent consecutive byte pairs and encoding them using the unused byte values (Gage, 1994). This algorithm was adapted to NMT by Sennrich et al. (2016) for achieving open vocabulary translation. In the modified algorithm, the most frequent character sequences are iteratively merged for a predetermined number of times in order to generate a fixed-size vocabulary of sub-word units. This purely statistical method is based on the hypothesis that many types of words can be translated when segmented into smaller units, such as named entities and loanwords. However, by solely relying on corpus frequency, one cannot provide a sufficiently compact vocabulary that can generalize among the inflected surface forms commonly observed in morphologically-rich languages (Ataman et al., 2017; Huck et al., 2017; Tamchyna et al., 2017). Moreover, many studies have showed that splitting words into sub-word units at positions that disregard the morpheme boundaries can lead to semantically ambiguous sub-word units, and consequently, inaccurate translations (Niehues et al., 2016; Ataman et al., 2017; Pinnis et al., 2017).

Linguistically-Motivated Vocabulary Reduction also constitutes a pre-processing step to NMT where an unsupervised morphology learning algorithm learns the optimal way of segmenting words into morphs and later uses the lexicon of morphs to build a sub-word vocabulary for the translation engine. The method is an extension of *Morfessor FlatCat* (Grönroos et al., 2014), where a Hidden Markov Model (HMM) models the composition of a word based on the transitions between different morphs and their morphological categories (*i.e.* prefix, stem or suffix). The category-based HMM is essential for a linguistically motivated segmentation, as words are only split considering the possible categories of the morphs and not at positions which may break the morphological structure or generate semantically ambiguous sub-word units. Ataman et al. (2017) have modified this method in order to optimize the morphology model with a constraint on the output vocabulary size, allowing it to be adopted as a vocabulary reduction method for NMT. By manipulating regularities in morphological transformations of the concatenating nature, LMVR aids to improve the NMT of languages with agglutinative or templatic morphology. However, it does not yield significant improvements in fusional languages where the boundaries of morphemes inside the words are not transparent (Ataman and Federico, 2018a).

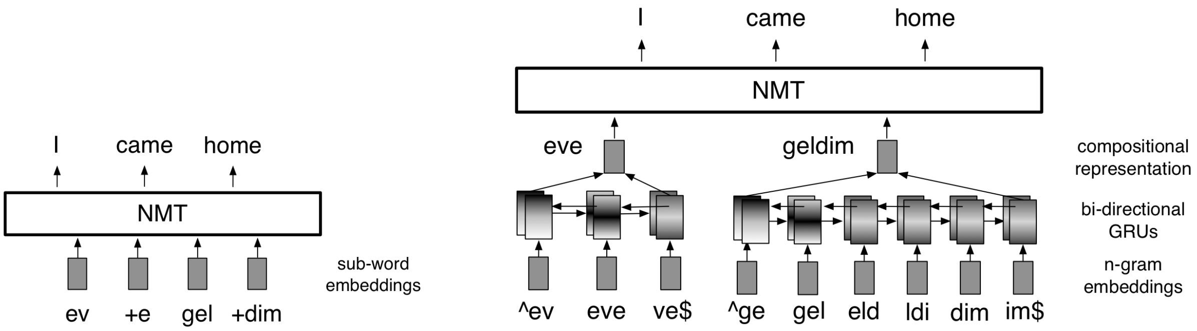
3.3 Morphological Analysis

In contrast to statistical approaches, few studies have opted to use supervised morphological analysis tools in order to reduce data sparseness in NMT. For instance, Sanchez and Toral (2016) have used a supervised morphological segmentation tool for English–Finnish NMT in order to separate words into root and inflection boundaries, whereas Huck and colleagues (2017) suggested to perform NMT based on a vocabulary of morphological features predicted by a morphological analyzer. While such methods aid in predicting a more compact NMT vocabulary in terms of root and affixes, they cannot reduce the vocabulary of a given text to fit any vocabulary size, which obliges one to further reduce the vocabulary using an unsupervised word segmentation method. Moreover, morphological analyzers are language-specific tools and as such they cannot provide general solutions to machine translation.

4 Learning Compositional Input Representations via bi-RNNs

One drawback of using statistical word segmentation methods for vocabulary prediction in NMT is that these methods constitute a pre-processing step to NMT, and hence they are not optimized for the translation task. Moreover, as given in Figure 1a, transforming sentences into sequences of sub-words leads to distributing the probability of a source word among multiple tokens, thus, increases the complexity of the alignment task performed by the attention model. In order to improve the accuracy in translating rare words in NMT, instead, we propose to perform NMT using word representations learned compositionally from smaller orthographic symbols inside the words, such as character n-grams, that can easily fit in the model vocabulary. This composition is essentially a function which can establish a mapping between combinations of orthographic units and lexical meaning, that is learned using the bilingual context, so that it can produce representations that are optimized for machine translation.

In our model (Figure 1b), the one-hot vectors retrieve the corresponding source embeddings for every word and feed them to an additional *composition layer*, which computes the final representations that are input to the encoder. For learning the mapping between the sublexical units and the lexical context, we employ a bi-RNN. Hence, by encoding the context of each interior unit inside the word, we believe that the network be able to capture important cues about their functional role, *i.e.* semantic or syntactic contribution to the word meaning. We implement the network using GRUs (Cho et al., 2014), which have shown comparable performance to LSTM units (Hochreiter and Schmidhuber, 1997) while performing faster computation. As a minimal set of input symbols required to cope with contextual ambiguities, and at the same time optimize the size of the NMT vocabulary, we opt to use intersecting sequences of character trigrams, as recently suggested by Vania and Lopez (2017). Our preliminary experiments (Ataman and Federico, 2018b) also confirmed the stand-alone sufficiency of character tri-



(a) Input: sub-word embeddings.

(b) Input: word representations built from char trigrams.

Figure 1: NMT of the Turkish sentence *Eve geldim* (*I came home*) using different input representations.

grams as fundamental units in the compositional NMT model.

Given a bi-RNN with a forward (f) and backward (b) layer, the input representation $\mathbf{w}$ of a token of t characters is computed from the hidden states $\mathbf{h}_t^f$ and $\mathbf{h}_b^0$, *i.e.* the final outputs of the forward and backward RNNs, as follows:

$$\mathbf{w} = \mathbf{W}_f \mathbf{h}_f^t + \mathbf{W}_b \mathbf{h}_b^0 + \mathbf{b} \qquad (3)$$

where $\mathbf{W}_f$ and $\mathbf{W}_b$ are weight matrices and $\mathbf{b}$ is a bias vector (Ling et al., 2015). These parameters are jointly learned together with the internal parameters of the GRUs and the input token embedding matrix to minimize the cost of the overall network while training the NMT model. For an input of m tokens, the computational complexity of the network is increased by $O(Kt_{\max}m)$, where K is the average cost of one bi-RNN layer and $t_{\max}$ is the maximum number of symbols per word.

5 Experiments

In order to evaluate our approach in NMT, we set up an evaluation benchmark which models NMT from four languages: Czech (*CS*), German (*DE*), Italian (*IT*) and Turkish (*TR*) into English (*EN*), where each input language represents a different lexical distribution reflected by its morphological characteristics, simulating conditions ranging from the low-resource and high sparseness (Turkish) to the high-resource and low sparseness (Italian) cases.

For training the Czech–English and German–English NMT models, we use the available data sets from the WMT[2] (Bojar et al., 2017) shared task on machine translation of news, which consist of Europarl (Koehn, 2005), Commoncrawl and News Commentary (Tiedemann, 2009). For achieving a comparable size of training data, we reduce the training set in German–English using the Invitation Model (Cuong and Simaan, 2014). We evaluate these models on the official test sets from 2016. Due to the lack of sufficient amount of news domain data, for the Italian–English and Turkish–English directions, we build generic NMT systems using data collected from TED Talks (Cettolo et al., 2012), EU Bookshop (Skadins et al., 2014), Global Voices, Gnome, Tatoeba, Ubuntu (Tiedemann, 2012), KDE4 (Tiedemann, 2009), Open Subtitles (Lison and Tiedemann, 2016) and SETIMES (Tyers and Alperen, 2010), and reduce the size of the training data for having comparable numbers of tokens (Italian) and types (Turkish) with the other languages. These models are evaluated on the official test sets from the evaluation campaign of IWSLT[3] (Cettolo et al., 2017). The morphological characteristics of the languages used in our study are presented in Table 1, while the statistics of the data sets used in our experiments can be seen in Tables 2 and 3.

We perform NMT by keeping the segmentation

Language	Morphological Typology	Morphological Complexity
Italian	*Fusional*	*Low*
German	*Fusional*	*Medium*
Czech	*Fusional,* *Agglutinative*	*High*
Turkish	*Agglutinative*	*High*

Table 1: The evaluated languages in our study along with their morphological characteristics.

[2] The First Conference on Machine Translation

[3] The International Workshop on Spoken Language Translation with shared tasks organized between 2003-2017.

Language	# sentences (K)	# tokens (M)	# types (K)
IT-EN	785	21(IT) - 22(EN)	152(IT) - 106(EN)
DE-EN	992	19(DE) - 18(EN)	501(DE) - 261(EN)
CS-EN	965	22(CS) - 25(EN)	385(CS) - 204(EN)
TR-EN	434	6(TR) - 8(EN)	373(TR) - 135(EN)

Table 2: Training sets. (*M*: Million, *K*: Thousand.)

Language	Data sets		# sentences (K)	# tokens (K)
IT-EN	Dev	dev2010 & test2010	3,5	74(IT) - 79(EN)
	Test	test2011 & test2012	3,2	55(IT) - 60(EN)
DE-EN	Dev	test2015	2,2	44(DE) - 46(EN)
	Test	test2016	3,0	62(DE) - 65(EN)
CS-EN	Dev	test2015	2,7	46(CS) - 54(EN)
	Test	test2016	3,0	57(CS) - 65(EN)
TR-EN	Dev	dev2010 & test2010	2,4	34(TR) - 47(EN)
	Test	test2011 & test2012	2,7	39(TR) - 53(EN)

Table 3: Development and Testing Sets. All data set are official evaluation sets from WMT (Czech and German) and IWSLT (Italian and Turkish). (*M*: Million, *K*: Thousand.)

on the English side constant and applying different open vocabulary NMT approaches to the input languages. We segment the English side with LMVR as it provides a segmentation that is more consistent with the morpheme boundaries (Ataman and Federico, 2018b).

The compositional bi-RNN is implemented in PyTorch (Paszke et al., 2017) and integrated into the OpenNMT-py toolkit (Klein et al., 2017). The *simple* NMT model constitutes the baseline in our study and performs translation directly at the level of sub-word units, using a two-layer encoder based on Stacked GRUs, a two-layer GRU decoder, input feeding and the general global attention mechanism (Luong et al., 2015). For segmenting the words in the source side, we chose to use BPE for the fusional languages (Czech, German and Italian), whereas in Turkish we use LMVR, as suggested in (Ataman and Federico, 2018a). The *compositional* model, on the other hand, performs NMT with input representations composed from a vocabulary of character trigrams. All the models use an embedding and GRUs with size 512. In order to achieve a fair comparison, we use a one-layer encoder for the compositional model, which allows the two models to have comparable number of parameters, whereas we use the same settings for the remaining network properties and hyper-parameters. All models are trained using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.0002 and default values for the other hyper-parameters. We clip the gradient norm at 1.0 (Pascanu et al., 2013) and set the dropout at 0.1 after hyper-parameter tuning. All models are trained with a model vocabulary of 30,000 units. The compositional model uses a trigram vocabulary of the same size whereas the segmentation methods (BPE and LMVR) are trained to fit in this exact vocabulary limit. We evaluate the accuracy of each model output using the (case-sensitive) BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and chrF (Popovic, 2015) metrics. Significance tests are computed only for BLEU with Multeval (Clark et al., 2011).

6 Results and Discussion

The performance of NMT models in translating each language using different types of encoder input representations can be seen in Table 4. The results show that the compositional model achieves the best translation accuracy in translation of all morphologically-rich languages. The overall improvements obtained with this model over the best performing simple model are **0.77** BLEU points in German, **0.74** BLEU points in Czech and **0.11** BLEU points in Turkish to English translation directions. The improvements are more evident for Turkish in terms of other evaluation metrics, where the compositional model improves the translation accuracy by **0.016** TER and **0.009** chrF points. In

Language Direction	Model	BLEU	TER	chrF
IT-EN	Simple (BPE)	**29.02**	**0.501**	**0.5328**
	Compositional	28.66	0.506	0.5293
DE-EN	Simple (BPE)	20.46	0.591	**0.4544**
	Compositional	**21.23**	**0.585**	0.4537
CS-EN	Simple (BPE)	19.59	0.615	0.4724
	Compositional	**20.33**	**0.614**	**0.4780**
TR-EN	Simple (LMVR)	23.02	0.585	0.4613
	Compositional	**23.13**	**0.569**	**0.4703**

Table 4: Experiment Results. Best scores for each translation direction are in bold font. All improvements over the baseline are statistically significant (p-value < 0.01).

Italian to English translation direction, the performance of the simple model is higher than the compositional model by **0.36** BLEU, **0.005** TER and **0.0035** chrF points.

The better performance of the compositional model in translating German, Czech and Turkish suggests that our approach is beneficial in eliminating the morphological errors caused by segmentation in languages with different morphological typologies. The improvements are highest for Czech and German, both of which have a fusional morphology of medium to high complexity, and the source language vocabulary of the training data ranges from around 400,000 to 500,000 types of words, indicating a high level of lexical sparseness. At a comparable vocabulary size, the improvements are generally lower in Turkish to English translation direction, where the input language has an agglutinative morphology with a much higher level of data sparseness. This might be due to the efficient performance of LMVR in generating morphologically-consistent sub-word units in the low-resource setting of agglutinative languages. Nevertheless, the results suggest that our compositional model can learn a higher level of morphological knowledge than LMVR, which was previously found to provide comparable performance to morphological analyzers in Turkish–English NMT using the embedding-based input representations (Ataman et al., 2017). Moreover, it can also generalize over different types of morphology in both low and high resource settings.

In the Italian to English translation direction, despite the comparable size of training data with Czech and German in the high-resource setting, the source word vocabulary is around 150,000 words, which represents the low level of sparseness. The higher overall performance of the NMT model which uses BPE for vocabulary reduction compared to the compositional model suggests that the embedding based sub-word representations are sufficient in reducing this vocabulary to fit into a space of 30,000 units. Nevertheless, in order to observe the actual accuracy in translating rare words, we carry out a focused analysis where we sample from the test sets only the sentences that contain singletons (*i.e.* words that are observed once in the training corpus) in the source side and evaluate the translation accuracy obtained with each NMT model on these sentences. This sampling results in 190 sentences in Italian, 470 sentences in Turkish, 562 sentences in German and 611 sentences in Czech to English directions. The results of this analysis, which can be found in Table 5, show that the compositional model translates sentences containing rare words more accurately than the simple model in all languages, with improvements ranging from **0.53** to **2.72** BLEU points. The improvement obtained also in the Italian to English translation direction shows that although in overall sub-word segmentation achieves higher output accuracy, it is still not as efficient as our approach in translating the small portion of rare words in the Italian corpus.

We extend our analysis in order to also evaluate the performance of different approaches in translating out-of-vocabulary (OOV) words. Similarly, we sample from the test sets only the sentences which contain OOVs, resulting in relatively larger test sets of 443 Italian, 1096 Turkish, 1396 Czech and 1449 German sentences. The evaluation of each NMT model on these sets, results of which are also given in Table 5, show that the compositional model again outperforms the simple NMT

Language Direction	**Model**	**BLEU** (Singletons)	**BLEU** (OOVs)
IT-EN	Simple (BPE)	23.54	23.23
	Compositional	**24.07**	**24.98**
DE-EN	Simple (BPE)	14.19	14.30
	Compositional	**16.91**	**16.76**
CS-EN	Simple (BPE)	16.33	16.83
	Compositional	**16.60**	**17.73**
TR-EN	Simple (LMVR)	19.69	20.31
	Compositional	**20.91**	**21.50**

Table 5: Translation accuracy of NMT models evaluated only on sentences containing singletons and OOVs. Best scores for each translation direction are in bold font. All improvements over the baseline are statistically significant (p-value < 0.01).

model in all languages, where the improvements range from **0.90** to **2.46** BLEU points. These findings suggest that our compositional NMT approach provides a higher generalization capability compared to conventional approaches to open vocabulary NMT.

7 Conclusion

In this paper, we have addressed the problem of translating rare words in NMT and proposed to solve it by replacing the conventional sub-word embeddings with input representations compositionally learned from character n-grams using a bi-RNN. Our approach showed significant and consistent improvements over a variety of languages with different morphological typologies, making it a competitive approach for NMT of low-resource and morphologically-rich languages. In the future, we plan to extend our approach in order to improve also the target side representations used by the NMT decoder and to evaluate it under similar morphological and data sparseness conditions on the target side. Finally, our benchmark and implementation are available for public use.

Acknowledgments

The authors would like to thank NVIDIA for their donation of Titan Xp GPUS that allowed to conduct this research.

References

Ataman, Duygu, Matteo Negri, Marco Turchi and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics* 108.1 (2017): 331-342.

Ataman, Duygu and Marcello Federico. 2018a. An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation. *Proceedings of the The 13th Conference of The Association for Machine Translation in the Americas*, Boston, USA. 97-110.

Ataman, Duygu and Marcello Federico. 2018b. Compositional Representation of Morphologically-Rich Input for Neural Machine Translation arXiv preprint arXiv:2251036.

Barone, Antonio Valerio Miceli, Jindrich Helcl, Rico Sennrich, Barry Haddow and Alexandra Birch. 2017. Deep architectures for Neural Machine Translation. *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. 99–107.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, USA. 257–267.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann and Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, and others. 2017. Findings of the 2017 Conference on Machine Translation (WMT) *Proceedings of the Second Conference on Machine Translation.* , Copenhagen, Denmark. 169–214.

Bottou, Léon 2010. Large-Scale Machine Learning with Stochastic Gradient Descent *Proceedings of 19th International Conference on Computational Statistics (COMPSTAT)*, Paris, France. Springer. 177–186.

Cettolo, Mauro, Christian Girardi and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, Trento, Italy.

Cettolo, Mauro, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuitho Sudoh, Koichiro Yoshino, Christian Federmann 2017 Overview of the IWSLT 2017 Evaluation Campaign. *International Workshop on Spoken Language Translation* 2–14.

Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches *Syntax, Semantics and Structure in Statistical Translation (2014): 103.*

Chung, Junyoung, Kyunghyun Cho and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany. (Volume 1: Long Papers). 1693–1703.

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL).* 176–181.

Costa-jussà, Marta R. and José A. R. Fonollosa. 2016. Character-based Neural Machine Translation *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL).* 357–361.

Cuong, Hoang and Khalil Simaan. 2014. Latent domain translation models in mix-of-domains haystack *Proceedings of Proceedings of the 25th International Conference on Computational Linguistics (COLING).* 1928–1939.

Gage, Philip 1994. A New Algorithm for Data Compression *The C Users Journal.* 12(2):23–38.

Grönroos, Stig-Arne, Sami Virpioj, Peter Smit and Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology. *Proceedings of the 25th International Conference on Computational Linguistics (COLING).* 1177–1185.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory *Neural computation.* MIT Press 1735–1780.

Huck, Matthias, Simon Riess and Alexander Fraser. 2017. Target-Side Word Segmentation Strategies for Neural Machine Translation *Proceedings of the 2nd Conference on Machine Translation (WMT)*, Copenhagen, Denmark. 56–67.

Junczys-Dowmunt, Marcin, Tomasz Dwojak and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*

Kalchbrenner, Nal, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves and Koray Kavukcuoglu. 2016 Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099.*

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *3rd International Conference for Learning Representations*, San Diego, USA.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart and Alexander M. Rush. 2017. *OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, System Demonstrations, 67-72.

Kim, Yoon, et al. 2016. Character-Aware Neural Language Models. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, USA. 2741-2749.

Koehn, Philipp 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit (MT Summit)* 79–86.

Lee, Jason, Kyunghyun Cho and Thomas Hofmann. 2015. *Fully Character-Level Neural Machine Translation without Explicit Segmentation.* Transactions of the Association for Computational Linguistics (TACL). 5: 365–378

Ling, Wang, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W. Black and Isabel Trancoso. 2015. *Finding function in form: Compositional character models for open vocabulary word representation.* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 1520–1530.

Ling, Wang, Isabel Trancoso, Chris Dyer and Alan W. Black. 2015. *Character-based neural machine translation.* arXiv preprint arXiv:1511.04586.

Lison, Pierre and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation.* 923–929.

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

Luong, Thang, and Christopher D. Manning. 2016. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL).* 1054–1063.

Niehues, Jan, Eunah Cho, Thanh-Le Ha and Alex Waibel. 2016. *Pre-Translation for Neural Machine Translation.* Proceedings of The 26th International Conference on Computational Linguistics (COLING). 1828–1836.

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

Pascanu, Razvan, Tomas Mikolov and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, USA. 1310–1318.

Pascanu, Razvan, Caglar Gulcehre, Kyunghyun Cho and Yoshua Bengio. 2014. How to construct deep recurrent neural networks. *Proceedings of the 2nd International Conference on Learning Representations*, Banff, Canada.

Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga and Adam Lerer. 2017. Automatic differentiation in PyTorch. *NIPS 2017 Autodiff Workshop*, Long Beach, USA.

Pinnis, Mārcis, Rihards Krišlauks, Daiga Deksne and Toms Miks. 2017. Neural Machine Translation for Morphologically Rich Languages with Improved Subword Units and Synthetic Data *Proceedings of the International Conference on Text, Speech, and Dialogue (TSD)* 237–245.

Popovic, Maja. 2015. chrF: Character n-gram F-score for Automatic MT Evaluation. *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT)*, Lisbon, Portugal. 392–395.

Sánchez-Cartagena, Vıctor M. and Antonio Toral. 2016. Abu-MaTran at WMT 2016 Translation Task: Deep Learning, Morphological Segmentation and Tuning on Character Sequences *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. 362-370.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany. 1715–1725.

Snover, Matthew and Dorr, Bonnie and Schwartz, Richard and Micciulla, Linnea and Makhoul, John 2006. A study of translation edit rate with targeted human annotation. *Proceedings of association for machine translation in the Americas*. Vol. 200. No. 6. 223–231.

Skadiņš, Raivis, Jörg Tiedemann, Roberts Rozis and Daiga Deksne. 2014. Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus. *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland. 1850–1855.

Sutskever, Ilya, Oriol Vinyals and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems.* 3104–3112.

Tamchyna, Aleš, Marion Weller-Di Marco and Alexander Fraser. 2017. Modeling Target-Side Inflection in Neural Machine Translation *Proceedings of the 2nd Conference on Machine Translation (WMT)*, Copenhagen, Denmark. 32–42.

Tiedemann, Jörg 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *In Recent Advances in Natural Language Processing* Amsterdam, Philadelphia. Vol. 5. 237–248.

Tiedemann, Jörg 2012. Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the eighth international conference on Language Resources and Evaluation* Vol. 2012. 2214–2218.

Tyers, Francis M. and Murat Serdar Alperen. 2010. South-east European Eimes: A parallel corpus of balkan languages. *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages* 49–53.

Vania, Clara and Adam Lopez. 2009. From Characters to Words to in Between: Do We Capture Morphology? *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)* 2016–2027.

Werbos, Paul J 1990. Backpropagation Through Time: What it does and How to do it *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)* 78:1550–1560.

Wu, Yonghui and Schuster, Mike and Chen, Zhifeng and Le, Quoc V and Norouzi, Mohammad and Macherey, Wolfgang and Krikun, Maxim and Cao, Yuan and Gao, Qin and Macherey, Klaus and others 2016. Googles Neural Machine Translation System: Bridging the Gap between Human and Machine Translation *arXiv preprint arXiv:1609.08144*