Compositional Representation of Morphologically-Rich Input for Neural Machine Translation

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

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

Abstract

Neural machine translation (NMT) models are typically trained with fixed-size input and output vocabularies, which creates an important bottleneck on their accuracy and generalization capability. As a solution, various studies proposed segmenting words into sub-word units and performing translation at the sub-lexical level. However, statistical word segmentation methods have recently shown to be prone to morphological errors, which can lead to inaccurate translations. In this paper, we propose to overcome this problem by replacing the source-language embedding layer of NMT with a bi-directional recurrent neural network that generates compositional representations of the input at any desired level of granularity. We test our approach in a low-resource setting with five languages from different morphological typologies, and under different composition assumptions. By training NMT to compose word representations from character trigrams, our approach consistently outperforms (from 1.71 to 2.48 BLEU points) NMT learning embeddings of statistically generated sub-word units.

1 Introduction

An important problem in neural machine translation (NMT) is translating infrequent or unseen words. The reasons are twofold: the necessity of observing many examples of a word until its input representation (embedding) becomes reliable, and the computational requirement of limiting the input and output vocabularies to few tens of thousands of words. These requirements eventually lead to coverage issues when dealing with low-resource and/or morphologically-rich languages, due to their high lexical sparseness. To cope with this well-known problem, several approaches have been proposed redefining the model vocabulary in terms of interior orthographic units compounding the words, ranging from character n-grams (Ling et al., 2015b; Costa-jussà and Fonollosa, 2016; Lee et al., 2017; Luong and Manning, 2016) to statistically-learned sub-word units (Sennrich et al., 2016; Wu et al., 2016; Ataman et al., 2017). While the former provide an ideal open vocabulary solution, they mostly failed to achieve competitive results. This might be related to the semantic ambiguity caused by solely relying on input representations based on character n-grams which are generally learned by disregarding any morphological information. In fact, the second approach is now prominent and has established a pre-processing step for constructing a vocabulary of sub-word units before training the NMT model. However, several studies have shown that segmenting words into sub-word units without preserving morpheme boundaries can lead to loss of semantic and syntactic information and, thus, inaccurate translations (Niehues et al., 2016; Ataman et al., 2017; Pinnis et al., 2017; Huck et al., 2017; Tamchyna et al., 2017).

In this paper, we propose to improve the quality of input (source language) representations of rare words in NMT by augmenting its *embedding layer* with a *bi-directional recurrent neural network* (bi-RNN), which can learn compositional input representations at different levels of granularity. Compositional word embeddings have recently been applied in language modeling and obtained successful results (Vania and Lopez, 2017). The apparent advantage of our approach is that by feeding NMT with simple character n-grams, our bi-RNN can potentially learn the morphology necessary to create word-level representations of the in-

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 305–311
Melbourne, Australia, July 15 - 20, 2018.

put language directly at training time, thus, avoiding the burden of a separate and sub-optimal word segmentation step. We compare our approach against conventional embedding-based representations learned from statistical word segmentation in a public evaluation benchmark, which provides low-resource training conditions by pairing English with five morphologically-rich languages: Arabic, Czech, German, Italian and Turkish, where each language represents a distinct morphological typology and language family. The experimental results show that our compositional input representations lead to significantly and consistently better translation quality in all language directions.

2 Neural Machine Translation

In this paper, we use the NMT model of Bahdanau et al. (2014). The model essentially estimates the conditional probability of translating a source sequence $x = (x_1, x_2, \dots x_m)$ into a target sequence $y = (y_1, y_2, \dots y_l)$, using the decomposition

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

The model is trained by maximizing the log-likelihood of a parallel training set via stochastic gradient descent (Bottou, 2010) and the backpropagation through time (Werbos, 1990) algorithms.

The inputs of the network are *one-hot* vectors, which are binary vectors with a single bit set to 1 to identify a specific word in the vocabulary. Each one-hot vector is then mapped to an *embedding*, a distributed representation of the word in a lower dimension but a more dense continuous space. From this input, a representation of the whole input sequence is learned using a bi-RNN, the *encoder*, which maps *x* into *m* dense sentence vectors corresponding to its hidden states. Next, another RNN, the *decoder*, predicts each target token y_i by sampling from a distribution computed from the previous target token y_{i-1}, the previous decoder hidden state, and the *context vector*. The latter is a linear combination of the encoder hidden states, whose weights are dynamically computed by a feed-forward neural network called *attention model* (Bahdanau et al., 2014). The probability of generating each target word y_j is normalized via a softmax function.

Both the source and target vocabulary sizes play an important role in terms of defining the complexity of the model. In a standard architecture, like ours, the source and target embedding matrices actually account for the vast majority of the network parameters. The vocabulary size also plays an important role when translating from and to low-resource and morphologically-rich languages, due to the sparseness of the lexical distribution. Therefore, a conventional approach has now become to compose both the source and target vocabularies of sub-word units generated through statistical segmentation methods (Sennrich et al., 2016; Wu et al., 2016; Ataman et al., 2017), and performing NMT by directly learning embeddings of sub-word units. A popular one of these is the Byte-Pair Encoding (BPE) method (Gage, 1994; Sennrich et al., 2016), which finds the optimal description of a corpus vocabulary by iteratively merging the most frequent character sequences. A more recent approach is the Linguistically-Motivated Vocabulary Reduction (LMVR) method (Ataman et al., 2017), which similarly generates a new vocabulary by segmenting words into sub-lexical units based on their likeliness of being morphemes and their morphological categories. A drawback of these methods is that, as pre-processing steps to NMT, they are not optimized for the translation task. Moreover, they can suffer from morphological errors at different levels, which can lead to loss of semantic or syntactic information.

3 Learning Compositional Input Representations via bi-RNNs

In this paper, we propose to perform NMT from input representations learned by composing smaller symbols, such as character n-grams (Ling et al., 2015a), that can easily fit in the model vocabulary. This composition is essentially a function which can establish a mapping between combinations of ortographic 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 1), the one-hot vectors, after being fed into the embedding layer, are processed by an additional *composition layer*, which computes the final input representations passed to the encoder to generate translations. For learning the composition function, we employ a bi-RNN. Hence, by encoding each interior unit inside the word, we hope to capture important cues about their functional role, *i.e.* semantic or syn-

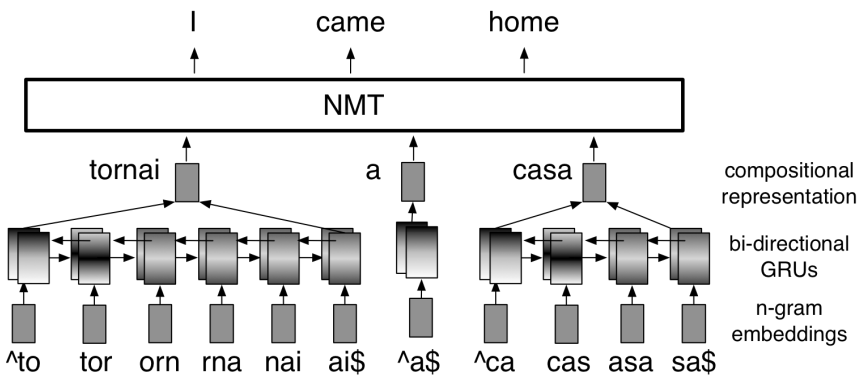

Figure 1: Translation of the Italian sentence *tornai a casa* (*I came home*) with a word-level representation composed from character trigrams.

tactic contribution to the word. We implement the network using gated recurrent units (GRUs) (Cho et al., 2014), which have shown comparable performance to long-short-term-memory units (Hochreiter and Schmidhuber, 1997), whereas they provide much faster computation. As a minimal set of input symbols required to cope with contextual ambiguities, we opt to use intersecting sequences of character trigrams, as recently suggested by Vania and Lopez (2017).

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:

$$\mathbf{w} = \mathbf{W}_f \mathbf{h}_f^t + \mathbf{W}_b \mathbf{h}_b^0 + \mathbf{b} \quad (2)$$

where $\mathbf{W}_f$ and $\mathbf{W}_b$ are weight matrices associated to each RNN and $\mathbf{b}$ is a bias vector (Ling et al., 2015a). These parameters are jointly learned together with the internal parameters of the GRUs and the input token embedding matrix while training the NMT model. For an input of m tokens, our implementation increases the computational complexity of the network by $O(Kt_{\max}m)$, where K is the bi-RNN cost and $t_{\max}$ is the maximum number of symbols per word. However, since computation of each input representation is independent, a parallelised implementation could cut the overhead down to $O(Kt_{\max})$.

4 Experiments

We test our approach along with statistical word segmentation based open vocabulary NMT methods in an evaluation benchmark simulating a low-resource translation setting pairing English (*En*) with five languages from different language families and morphological typologies: Arabic (*Ar*), Czech (*Cs*), German (*De*), Italian (*It*) and Turkish (*TR*). The characteristics of each language are given in Table 1, whereas Table 2 presents the statistical properties of the training data. We train our NMT models using the TED Talks corpora (Cettolo et al., 2012) and test them on the official data sets of IWSLT[1] (Mauro et al., 2017).

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

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

Language Pair	# tokens		# types	
	Src	Tgt	Src	Tgt
Tr - En	2,7M	2,0M	171K	53K
Ar - En	3,9M	4,9M	220K	120K
Cs - En	2,0M	2,3M	118K	50K
De - En	4,0M	4,3M	144K	69K
It - En	3,5M	3,8M	95K	63K

Table 2: Sizes of the training sets and vocabularies in the TED Talks benchmark. Development and test sets are on average 50K to 100K tokens. (*M*: Million, *K*: Thousand.)

The *simple* NMT model constitutes the baseline in our study and performs translation directly at the level of sub-word units, which can be of four different types: characters, character trigrams, BPE sub-word units, and LMVR sub-word units.

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

The *compositional* model, on the other hand, performs NMT with input representations composed from sub-lexical vocabulary units. In our study, we evaluate representations composed from character trigrams, BPE, and LMVR units. In order to choose the segmentation method to apply on the English side (the output of NMT decoder), we compare BPE and LMVR sub-word units by carrying out an evaluation on the official data sets of Morpho Challenge 2010[2](Kurimo et al., 2010). The results of this evaluation, as given in Table 3, suggest that LMVR seems to provide a segmentation that is more consistent with morpheme boundaries, which motivates us to use sub-word tokens generated by LMVR for the target side. This choice aids us in evaluating the morphological knowledge contained in input representations in terms of the translation accuracy in NMT.

The compositional bi-RNN layer is implemented in Theano (Team et al., 2016) and integrated into the Nematus NMT toolkit (Sennrich et al., 2017). In our experiments, we use a compositional bi-RNN with 256 hidden units, an NMT model with a one-layer bi-directional GRU encoder and one-layer GRU decoder of 512 hidden units, and an embedding dimension of 256 for both models. We use a highly restricted dictionary size of 30,000 for both source and target languages, and train the segmentation models (BPE and LMVR) to generate sub-word vocabularies of the same size. We train the NMT models using the Adagrad (Duchi et al., 2011) optimizer with a mini-batch size of 50, a learning rate of 0.01, and a dropout rate of 0.1 (in all layers and embeddings). In order to prevent over-fitting, we stop training if the perplexity on the validation does not decrease for 5 epochs, and use the best model to translate the test set. The model outputs are evaluated using the (case-sensitive) BLEU (Papineni et al., 2002) metric and the Multeval (Clark et al., 2011) significance test.

[2]Shared Task on Unsupervised Morphological Analysis, http://morpho.aalto.fi/events/morphochallenge.

Method	Precision	Recall	F_1 Score
BPE	52.87	24.44	33.43
LMVR	**70.22**	**55.66**	**62.10**

Table 3: The performance of different segmentation models trained on the English portion of our benchmark in the Morpho Challenge shared task.

5 Results

The performance of NMT models in translating each language using different vocabulary units and encoder input representations can be seen in Table 4. With the simple model, LMVR based units achieve the best accuracy in translating all languages, with improvements over BPE by **0.85** to **1.09** BLEU points in languages with high morphological complexity (Arabic, Czech and Turkish) and **0.32** to **0.53** BLEU points in languages with low to medium complexity (Italian and German). This confirms our previous results in (Ataman and Federico, 2018). Moreover, simple models using character trigrams as vocabulary units reach much higher translation accuracy compared to models using characters, indicating their superior performance in handling contextual ambiguity. In the Italian to English translation direction, the performance of simple models using character trigrams and BPE sub-word units as input representations are almost comparable, showing that character trigrams can even be sufficient as the standalone vocabulary units in languages with low lexical sparseness. These findings suggest that each type of sub-word unit used in the simple model is specifically convenient for a given morphological typology.

Using our compositional model improves the quality of input representations for each type of vocabulary unit, nevertheless, the best performance is obtained by using character trigrams as input symbols and words as input representations. The higher quality of these input representations compared to those obtained from sub-word units generated with LMVR suggest that our compositional model can learn morphology better than LMVR, which was found to provide comparable performance to morphological analyzers in Turkish to English NMT (Ataman et al., 2017). Moreover, sample outputs from both models show that the compositional model is also able to better capture syntactic information of input sentences. Figure 5 illustrates two example translations from Italian and Turkish. In Italian, the simple model fails to understand the common subject of different verbs in the sentence due to the repetition of the same inflective suffix after segmentation. In Turkish, the genitive case ”yerlerin fotoğraflarının” (*the photographs of places*) and the complex predicate ”birleştirilmesiyle meydana geldi” (*is composed of*) are both incorrectly

Model	Vocabulary Units	Input Representations	BLEU				
			Tr-En	Ar-En	Cs-En	De-En	It-En
Simple	Characters	Characters	12.29	8.95	13.42	21.32	22.88
	Char Trigrams	Char Trigrams	16.13	11.91	20.87	25.01	26.68
	Subwords (BPE)	Subwords (BPE)	16.79	11.14	21.99	26.61	27.02
	Subwords (LMVR)	Subwords (LMVR)	17.82	12.23	22.84	27.18	27.34
Compositional	Char Trigrams	Subwords (BPE)	15.40	11.50	21.67	27.05	27.80
	Char Trigrams	Subwords (LMVR)	16.63	13.29	23.07	26.86	26.84
	Char Trigrams	Words	**19.53**	**14.22**	**25.16**	**29.09**	**29.82**
	Subwords (BPE)	Words	12.64	11.51	23.13	27.10	27.96
	Subwords (LMVR)	Words	18.90	13.55	24.31	28.07	28.83

Table 4: Experiment results. Best scores for each translation direction are in bold font. All improvements over the baseline (simple model with BPE) are statistically significant (p-*value* < 0.05).

Input (Simple Model)	e comunque, em@@ ig@@ *riamo* , circol@@ *iamo* e mescol@@ *iamo* così tanto che non esiste più l' isolamento necessario affinché avvenga un' evoluzione .
NMT Output (Simple Model)	and anyway , *we* repair, and *we* mix so much that there 's no longer the isolation that *we* need to happen to make an evolution .
Input (Compositional Model)	e comunque, emigriamo, circoliamo e mescoliamo così tanto che non esiste più l' isolamento necessario affinché avvenga un' evoluzione.
NMT Output (Compositional Model)	and anyway , *we* migrate , circle and mix so much that there 's no longer the isolation necessary to become evolutionary .
Reference	and by the way , *we* immigrate and circulate and intermix so much that you can 't any longer have the isolation that is necessary for evolution to take place .

Input (Simple Model)	ama aslında bu resim tamamen , farklı *yerlerin fotoğraf@@ larının* **birleştir@@ il@@ mesiyle meydana geldi** .
NMT Output (Simple Model)	but in fact , this picture **came up with** a completely different *place of photographs* .
Input (Compositional Model)	ama aslında bu resim tamamen , farklı *yerlerin fotoğraflarının* **birleştirilmesiyle meydana geldi** .
NMT Output (Compositional Model)	but in fact , this picture **came from collecting** *pictures of* different *places* .
Reference	but this image **is** actually entirely **composed of** *photographs from* different *locations* .

Table 5: Example translations with different approaches in *Italian* (above) and *Turkish* (below).

translated by the simple model. On the other hand, the compositional model is able to capture the correct sentence semantics and syntax in either case. These findings suggest that maintaining translation at the lexical level apparently aids the attention mechanism and provides more semantically and syntactically consistent translations. The overall improvements obtained with this model over the best performing simple model are **1.99** BLEU points in Arabic, **2.32** BLEU points in Czech, **1.91** BLEU points in German, **2.48** BLEU points in Italian and **1.71** BLEU points in Turkish to English translation directions. As evident from the significant and consistent improvements across all languages, our approach provides a more promising and generic solution to the data sparseness problem in NMT.

6 Conclusion

In this paper, we addressed the problem of translating infrequent 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, making it a competitive solution for NMT of low-resource and morphologically-rich languages. In the future, we plan to optimize our implementation and to test its scalability on larger data sets.

Acknowledgments

The authors would like to thank NVIDIA for their computational support that aided this research.

References

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

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. In *The Prague Bulletin of Mathematical Linguistics*, volume 108, pages 331–342.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. In *arXiv preprint arXiv:1409.0473*.

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

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, pages 103–111.

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

Marta R Costa-jussà and José AR Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 357–361.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. In *Journal of Machine Learning Research*, volume 12, pages 2121–2159.

Philip Gage. 1994. A New Algorithm for Data Compression. In *The C Users Journal*, volume 12, pages 23–38.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. In *Neural computation*, volume 9, pages 1735–1780. MIT Press.

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

Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. 2010. Morpho challenge competition 2005–2010: evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-Level Neural Machine Translation without Explicit Segmentation. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 5, pages 365–378.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015a. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1520–1530.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015b. Character-based Neural Machine Translation. In *arXiv preprint arXiv:1511.04586*.

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

Cettolo Mauro, Federico Marcello, Bentivogli Luisa, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

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

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

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

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin

Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for Neural Machine Translation. In *Proceedings of the 15th Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 65–68.

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

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

The Theano Development Team, Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, et al. 2016. Theano: A python framework for fast computation of mathematical expressions. In *arXiv preprint arXiv:1605.02688*.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 2016–2027.

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Googles neural machine translation system: Bridging the gap between human and machine translation. In *arXiv preprint arXiv:1609.08144*.