Character-based Neural Machine Translation

Marta R. Costa-jussà and **José A. R. Fonollosa**
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
{marta.ruiz,jose.fonollosa}@upc.edu

Abstract

Neural Machine Translation (MT) has reached state-of-the-art results. However, one of the main challenges that neural MT still faces is dealing with very large vocabularies and morphologically rich languages.

In this paper, we propose a neural MT system using character-based embeddings in combination with convolutional and highway layers to replace the standard lookup-based word representations. The resulting unlimited-vocabulary and affix-aware source word embeddings are tested in a state-of-the-art neural MT based on an attention-based bidirectional recurrent neural network. The proposed MT scheme provides improved results even when the source language is not morphologically rich. Improvements up to 3 BLEU points are obtained in the German-English WMT task.

1 Introduction

Machine Translation (MT) is the set of algorithms that aim at transforming a source language into a target language. For the last 20 years, one of the most popular approaches has been statistical phrase-based MT, which uses a combination of features to maximise the probability of the target sentence given the source sentence (Koehn et al., 2003). Just recently, the neural MT approach has appeared (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015) and obtained state-of-the-art results.

Among its different strengths neural MT does not need to pre-design feature functions beforehand; optimizes the entire system at once because

it provides a fully trainable model; uses word embeddings (Sutskever et al., 2014) so that words (or minimal units) are not independent anymore; and is easily extendable to multimodal sources of information (Elliott et al., 2015). As for weaknesses, neural MT has a strong limitation in vocabulary due to its architecture and it is difficult and computationally expensive to tune all parameters in the deep learning structure.

In this paper, we use the neural MT baseline system from (Bahdanau et al., 2015), which follows an encoder-decoder architecture with attention, and introduce elements from the character-based neural language model (Kim et al., 2016). The translation unit continues to be the word, and we continue using word embeddings related to each word as an input vector to the bidirectional recurrent neural network (attention-based mechanism). The difference is that now the embeddings of each word are no longer an independent vector, but are computed from the characters of the corresponding word. The system architecture has changed in that we are using a convolutional neural network (CNN) and a highway network over characters before the attention-based mechanism of the encoder. This is a significant difference from previous work (Sennrich et al., 2015) which uses the neural MT architecture from (Bahdanau et al., 2015) without modification to deal with subword units (but not including unigram characters).

Subword-based representations have already been explored in Natural Language Processing (NLP), e.g. for POS tagging (Santos and Zadrozny, 2014), name entity recognition (Santos and aes, 2015), parsing (Ballesteros et al., 2015), normalization (Chrupala, 2014) or learning word representations (Botha and Blunsom, 2014; Chen et al., 2015). These previous works show different advantages of using character-level information. In our case, with the new character-

based neural MT architecture, we take advantage of intra-word information, which is proven to be extremely useful in other NLP applications (Santos and Zadrozny, 2014; Ling et al., 2015a), especially when dealing with morphologically rich languages. When using the character-based source word embeddings in MT, there ceases to be unknown words in the source input, while the size of the target vocabulary remains unchanged. Although the target vocabulary continues with the same limitation as in the standard neural MT system, the fact that there are no unknown words in the source helps to reduce the number of unknowns in the target. Moreover, the remaining unknown target words can now be more successfully replaced with the corresponding source-aligned words. As a consequence, we obtain a significant improvement in terms of translation quality (up to 3 BLEU points).

The rest of the paper is organized as follows. Section 2 briefly explains the architecture of the neural MT that we are using as a baseline system. Section 3 describes the changes introduced in the baseline architecture in order to use character-based embeddings instead of the standard lookup-based word representations. Section 4 reports the experimental framework and the results obtained in the German-English WMT task. Finally, section 5 concludes with the contributions of the paper and further work.

2 Neural Machine Translation

Neural MT uses a neural network approach to compute the conditional probability of the target sentence given the source sentence (Cho et al., 2014; Bahdanau et al., 2015). The approach used in this work (Bahdanau et al., 2015) follows the encoder-decoder architecture.First, the encoder reads the source sentence $s = (s_1, ..s_I)$ and encodes it into a sequence of hidden states $h = (h_1, ..h_I)$. Then, the decoder generates a corresponding translation $t = t_1, ..., t_J$ based on the encoded sequence of hidden states h. Both encoder and decoder are jointly trained to maximize the conditional log-probability of the correct translation.

This baseline autoencoder architecture is improved with a attention-based mechanism (Bahdanau et al., 2015), in which the encoder uses a bi-directional gated recurrent unit (GRU). This GRU allows for a better performance with long sentences. The decoder also becomes a GRU and each word t_j is predicted based on a recurrent hidden state, the previously predicted word t_{j-1}, and a context vector. This context vector is obtained from the weighted sum of the annotations h_k, which in turn, is computed through an alignment model α_{jk} (a feedforward neural network). This neural MT approach has achieved competitive results against the standard phrase-based system in the WMT 2015 evaluation (Jean et al., 2015).

3 Character-based Machine Translation

Word embeddings have been shown to boost the performance in many NLP tasks, including machine translation. However, the standard lookup-based embeddings are limited to a finite-size vocabulary for both computational and sparsity reasons. Moreover, the orthographic representation of the words is completely ignored. The standard learning process is blind to the presence of stems, prefixes, suffixes and any other kind of affixes in words.

As a solution to those drawbacks, new alternative character-based word embeddings have been recently proposed for tasks such as language modeling (Kim et al., 2016; Ling et al., 2015a), parsing (Ballesteros et al., 2015) or POS tagging (Ling et al., 2015a; Santos and Zadrozny, 2014). Even in MT (Ling et al., 2015b), where authors use the character transformation presented in (Ballesteros et al., 2015; Ling et al., 2015a) both in the source and target. However, they do not seem to get clear improvements. Recently, (Luong and Manning, 2016) propose a combination of word and characters in neural MT.

For our experiments in neural MT, we selected the best character-based embedding architecture proposed by Kim et al. (Kim et al., 2016) for language modeling. As the Figure 1 shows, the computation of the representation of each word starts with a character-based embedding layer that associates each word (sequence of characters) with a sequence of vectors. This sequence of vectors is then processed with a set of 1D convolution filters of different lengths (from 1 to 7 characters) followed with a max pooling layer. For each convolutional filter, we keep only the output with the maximum value. The concatenation of these max values already provides us with a representation of each word as a vector with a fixed length equal to the total number of convolutional ker-

nels. However, the addition of two highway layers was shown to improve the quality of the language model in (Kim et al., 2016) so we also kept these additional layers in our case. The output of the second Highway layer will give us the final vector representation of each source word, replacing the standard source word embedding in the neural machine translation system.
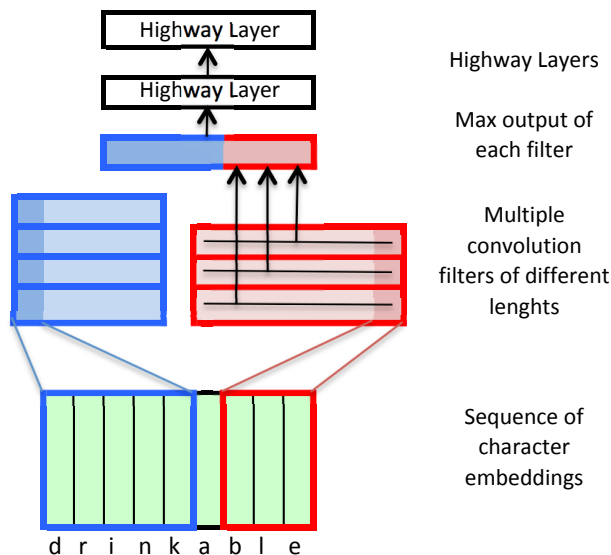

Figure 1: Character-based word embedding

In the target size we are still limited in vocabulary by the softmax layer at the output of the network and we kept the standard target word embeddings in our experiments. However, the results seem to show that the affix-aware representation of the source words has a positive influence on all the components of the network. The global optimization of the integrated model forces the translation model and the internal vector representation of the target words to follow the affix-aware codification of the source words.

4 Experimental framework

This section reports the data used, its preprocessing, baseline details and results with the enhanced character-based neural MT system.

4.1 Data

We used the German-English WMT data[1] including the EPPS, NEWS and Commoncrawl. Preprocessing consisted of tokenizing, truecasing, normalizing punctuation and filtering sentences with more than 5% of their words in a language

other than German or English. Statistics are shown in Table 1.

L	Set	S	W	V	OOV
De	Train	3.5M	77.7M	1.6M	-
	Dev	3k	63.1k	13.6k	1.7k
	Test	2.2k	44.1k	9.8k	1.3k
En	Train	3.5M	81.2M	0.8M	-
	Dev	3k	67.6k	10.1k	0.8k
	Test	2.2k	46.8k	7.8k	0.6k

Table 1: Corpus details. Number of sentences (S), words (W), vocabulary (V) and out-of-vocabulary-words (OOV) per set and language (L). M standing for millions, k standing for thousands.

4.2 Baseline systems

The phrase-based system was built using Moses (Koehn et al., 2007), with standard parameters such as grow-final-diag for alignment, Good-Turing smoothing of the relative frequencies, 5-gram language modeling using Kneser-Ney discounting, and lexicalized reordering, among others. The neural-based system was built using the software from DL4MT[2] available in github. We generally used settings from previous work (Jean et al., 2015): networks have an embedding of 620 and a dimension of 1024, a batch size of 32, and no dropout. We used a vocabulary size of 90 thousand words in German-English. Also, as proposed in (Jean et al., 2015) we replaced unknown words (UNKs) with the corresponding source word using the alignment information.

4.3 Results

Table 3 shows the BLEU results for the baseline systems (including phrase and neural-based, NN) and the character-based neural MT (CHAR). We also include the results for the CHAR and NN systems with post-processing of unknown words, which consists in replacing the UNKs with the corresponding source word (+Src), as suggested in (Jean et al., 2015). BLEU results improve by almost 1.5 points in German-to-English and by more than 3 points in English-to-German. The reduction in the number of unknown words (after postprocessing) goes from 1491 (NN) to 1260 (CHAR) in the direction from German-to-English and from 3148 to 2640 in the opposite direction. Note the

---

[1] http://www.statmt.org/wmt15/translation-task.html

[2] http://dl4mt.computing.dcu.ie/

1	SRC	Berichten zufolge hofft Indien darber hinaus auf einen Vertrag zur **Verteidigungszusammenarbeit** zwischen den beiden Nationen .
	Phrase	reportedly hopes India , in addition to a contract for the defence cooperation between the two nations .
	NN	according to reports , India also hopes to establish a contract for the UNK between the two nations .
	CHAR	according to reports , India hopes to see a Treaty of **Defence Cooperation** between the two nations .
	REF	India is also reportedly hoping for a deal on **defence collaboration** between the two nations .
2	SRC	der durchtrainierte Mainzer sagt von sich , dass er ein " ambitionierter **Rennradler** " ist .
	Phrase	the will of Mainz says that he a more ambitious .
	NN	the UNK Mainz says that he is a " ambitious , . "
	CHAR	the UNK in Mainz says that he is a ' **ambitious racer** ' .
	REF	the well-conditioned man from Mainz said he was an " **ambitious racing cyclist** . "
3	SRC	die GDL habe jedoch nicht gesagt , wo sie **streiken** wolle , so dass es schwer sei , die Folgen konkret vorherzusehen .
	Phrase	the GDL have , however , not to say , where they strike , so that it is difficult to predict the consequences of concrete .
	NN	however , the UNK did not tell which they wanted to UNK , so it is difficult to predict the consequences .
	CHAR	however , the UNK did not say where they wanted to **strike** , so it is difficult to predict the consequences .
	REF	the GDL have not said , however , where they will **strike** , making it difficult to predict exactly what the consequences will be .
4	SRC	die Premierminister Indiens und Japans trafen sich in Tokio .
	Phrase	the Prime Minister of India and Japan in Tokyo .
	NN	the Prime Minister of India and Japan met in Tokyo
	CHAR	the Prime **Ministers** of India and Japan met in Tokyo
	REF	India and Japan prime **ministers** meet in Tokyo
5	SRC	wo die Beamten es aus den Augen verloren .
	Phrase	where the officials lost sight of
	NN	where the officials lost it out of the eyes
	CHAR	where officials **lose sight of it**
	REF	causing the officers to **lose sight of it**

Table 2: Translation examples.

	De->En	En->De
Phrase	20.99	17.04
NN	18.83	16.47
NN+Src	20.64	17.15
CHAR	21.40	19.53
CHAR+Src	**22.10**	**20.22**

Table 3: De-En BLEU results.

number of out-of-vocabulary words of the test set is shown in Table 1.

The character-based embedding has an impact in learning a better translation model at various levels, which seems to include better alignment, reordering, morphological generation and disambiguation. Table 2 shows some examples of the kind of improvements that the character-based neural MT system is capable of achieving compared to baseline systems. Examples 1 and 2 show how the reduction of source unknowns improves the adequacy of the translation. Examples 3 and 4 show how the character-based approach is able to handle morphological variations. Finally, example 5 shows an appropriate semantic disambiguation.

5 Conclusions

Neural MT offers a new perspective in the way MT is managed. Its main advantages when compared with previous approaches, e.g. statistical phrase-based, are that the translation is faced with trainable features and optimized in an end-to-end scheme. However, there still remain many challenges left to solve, such as dealing with the limi-

tation in vocabulary size.

In this paper we have proposed a modification to the standard encoder/decoder neural MT architecture to use unlimited-vocabulary character-based source word embeddings. The improvement in BLEU is about 1.5 points in German-to-English and more than 3 points in English-to-German.

As further work, we are currently studying different alternatives (Chung et al., 2016) to extend the character-based approach to the target side of the neural MT system.

Acknowledgements

References

Dimitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359, Lisbon, Portugal, September. Association for Computational Linguistics.

Jan A. Botha and Phil Blunsom. 2014. Compositional Morphology for Word Representations and Language Modelling. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, Beijing, China, jun. *Award for best application paper*.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In Qiang Yang and Michael Wooldridge, editors, *IJCAI*, pages 1236–1242. AAAI Press.

Kyunghyun Cho, Bart van van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proc. of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha.

Grzegorz Chrupala. 2014. Normalizing tweets with edit scripts and recurrent neural embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 680–686.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. *CoRR*, abs/1603.06147.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR*, abs/1510.04709.

Sebastien Jean, Orhan Firat, Kyunghun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for wmt15. In *Proc. of the 10th Workshop on Statistical Machine Translation*, Lisbon.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, Seattle.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicolas Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

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*, pages 1520–1530, Lisbon, Portugal, September. Association for Computational Linguistics.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. 2015b. Character-based neural machine translation. *CoRR*, abs/1511.04586.

Min-Thang Luong and Cristopher D. Manning. 2016. Character-based neural machine translation. *CoRR*, abs/1511.04586.

Cicero D. Santos and Victor Guimar aes. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China, July. Association for Computational Linguistics.

Cicero D. Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.