Can Monolingual Embeddings Improve Neural Machine Translation?

Mattia A. Di Gangi
University of Trento, Trento, Italy
Fondazione Bruno Kessler
via Sommarive, 18, Trento, Italy
digangi@fbk.eu

Federico Marcello
Fondazione Bruno Kessler
via Sommarive, 18, Trento, Italy
federico@fbk.eu

Abstract

English. Neural machine translation (NMT) recently redefined the state of the art in machine translation, by introducing deep learning architecture that can be trained end-to-end. One limitation of NMT is the difficulty to learn representations of rare words. The most common solution is to segment words into subwords, in order to allow for shared representations of infrequent words. In this paper we present ways to directly feed a NMT network with external word embeddings trained on monolingual source data, thus enabling a virtually infinite source vocabulary. Our preliminary results show that while our methods do not seem effective under large-data training conditions (WMT En-De), they instead show great potential for the typical low-resourced data scenario (IWSLT En-Fr). By leveraging external embeddings learned on Web crawled English texts, we were able to improve a word-level En-Fr baseline trained on 200,000 sentence pairs by up to 4 BLEU points.

Italiano. *La traduzione automatica con reti neurali (neural machine translation, NMT) ha ridefinito recentemente lo stato dell'arte nella traduzione automatica introducendo un'architettura di deep learning che può essere addestrata interamente, dall'input all'output. Una limitazione della NMT è comunque la difficoltà di apprendere rappresentazioni di parole poco frequenti. La soluzione più adottata consiste nel segmentare le parole in sotto-parole, in modo da consentire rappresentazioni condivise per parole poco frequenti. In questo lavoro presentiamo dei metodi per fornire ad una rete word embedding esterni addestrati su testi nella lingua sorgente, consentendo quindi un vocabolario virtualmente illimitato sulla lingua di input. I nostri risultati preliminari mostrano che i nostri metodi, pur non sembrando efficaci sotto condizioni di addestramento con molti dati (WMT En-De), risultano invece promettenti per scenari di addestramento con poche risorse (IWSLT En-Fr). Sfruttando* word embedding *appresi da testi inglesi estratti dal Web, siamo riusciti a migliorare un sistema NMT basato a parole e addestrato su 200.000 coppie di frasi fino a 4 punti BLEU.*

1 Introduction

The latest developments of machine translation have been led by the neural approach (Sutskever et al., 2014; Bahdanau et al., 2014), a deep-learning based technique that has shown to outperform the previous methods in all the recent evaluation campaigns (Bojar et al., 2016; Cettolo et al., 2016).
NMT mainly relies on parallel data, which are expensive to produce as they involve human translation. Recently, *back-translation* (Sennrich et al., 2015a) has been proposed to leverage target language data. This consists in enriching the training data with synthetic translations produced with a reverse MT system (Bertoldi and Federico, 2009). Unfortunately, this method introduces noise and seems really effective only when the synthetic parallel sentences are only a fraction of the true ones. Hence, this approach does not allow to leverage huge quantities of monolingual data.
One consequence of the scarcity of parallel data is the occurrence of out-of-vocabulary (OOV) and rare words. In fact, being NMT a statistical approach, it cannot learn meaningful representations

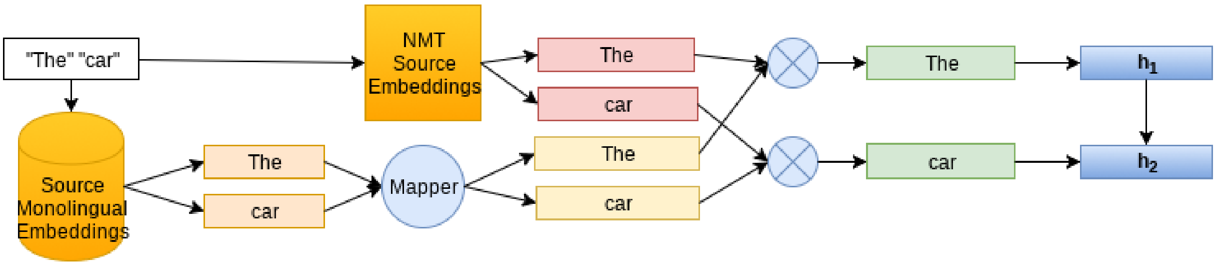

Figure 1: Merging external embeddings with the normal NMT embeddings in the encoder side. The tokens ”The” and ”car” are used to extract the two kinds of embeddings that are merged before being used as input for the encoder RNN.

for rare words and no representation at all for OOV words. The solution up to this moment is to segment words into sub-words (Sennrich et al., 2015b; Wu et al., 2016) in order to have a better representation of rare and OOV words, as parts of their representation will be ideally shared with other words. The drawback of this approach is that it generates longer input sequences, thus exacerbates the handling of long-term dependencies (Bentivogli et al., 2016). In this paper, we propose to keep the source input at a word level while alleviating the problem of rare and OOV words. We do it by integrating the usual word indexes with word embeddings that have been pre-trained on huge monolingual data. The intuition is that the network should learn to use the provided representations, which should be possibly more reliable for the rare words. This should be true particularly for the low-resource settings, where parameter transfer has shown to be an effective approach (Zoph et al., 2016). Because of the softmax layer, the same idea cannot be applied straightforwardly to the target side, hence we continue to use sub-words there. We show that the network is capable to learn how to translate from the input embeddings while replacing the source embedding layer with a much smaller feed-forward layer. Our results show that this method seems effective in a small training data setting, while it does not seem to help under large training data conditions. In the following section we briefly describe the state-of-the-art NMT architecture. Then, we introduce our modification to enable the use of external word embeddings. In Section 4, we introduce the experimental setup and show our results, while in Section 5 we discuss our solution. Finally, in Section 6 we presents our conclusions and the future work.

2 State of the art

Neural machine translation is based on the encoder-decoder-attention architecture (Bahdanau et al., 2014) which jointly learns the translation and alignment models with a sequence-to-sequence process. A sequence of source words $f_1, f_2, \ldots, f_m$ is mapped to sequence of embedding vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}$, via a look-up table $X \in R^{|V| \times d}$, where $|V|$ is the vocabulary size and d is the dimensionality of the embedding vectors. Hence, the memory occupied by the vocabulary is linear in both the vocabulary size and the embeddings size.

The embedding sequence is then processed by a bi-directional RNN (Schuster and Paliwal, 1997):

$$\overrightarrow{\mathbf{h}}_j = g(\mathbf{x}_j, \overrightarrow{\mathbf{h}}_{j-1}), \quad j = 1, ..m$$

$$\overleftarrow{\mathbf{h}}_j = g(\mathbf{x}_j, \overleftarrow{\mathbf{h}}_{j+1}), \quad j = m, .., 1$$

where g is the LSTM (Hochreiter and Schmidhuber, 1997) or the GRU (Cho et al., 2014) function, and the two directions are merged with functions like the vector concatenation or the point-wise sum. The sequence of vectors produced by the bidirectional RNN is the encoded representation of the source sentence.

The decoder takes as input the encoder outputs (or states) and produces a sequence of target words $e_1, e_2, \ldots, e_l$. The decoder works by progressively predicting the probability of the next target word e_i given the previously generated target words and the source context vector $\mathbf{c_i}$. At each step, the decoder computes a word embeddings $\mathbf{y}_{i-1}$ of the previous target word, applies one or more recurrent layers, an attention model function and a softmax layer. The recurrent layers produce an hidden state $\mathbf{s}_i$

$$\mathbf{s}_i = g(\mathbf{y}_{i-1}, \mathbf{s}_{i-1})$$

where, g can be computed with one or more LSTM or GRU layers. The output of the RNN is then used by the attention model (Luong et al., 2015a) to weight the source vectors according to their similarity with it.

$$\alpha_{ij} = \frac{\exp(score(\mathbf{s}_i, \mathbf{h}_j))}{\sum_{k=1}^{m} \exp(score(\mathbf{s}_i, \mathbf{h}_k))}$$

The weights are used to compute a weighted average of the encoder outputs, which represents the source context

$$\mathbf{c}_i = \sum_{j=1}^{m} \alpha_{ij} \mathbf{h}_j$$

The source context vector is then combined with the output of the last RNN layer in a new vector $\mathbf{z}_i$ that is passed as input to the softmax layer to compute the probability for each word in the vocabulary to be the next word, such that:

$$p(e \mid e_{i-1}, c_i) \propto \exp(\mathbf{o}^\top \mathbf{z}_i)$$

where $\mathbf{z_i}$ is a column of $\mathbf{Z}$, a matrix with the same size of the target-side embedding matrix. Let Θ be the set of all the network parameters, then the objective of the training is to find parameter values maximizing the likelihood of the training set S, i.e.:

$$\sum_{(\mathbf{f},\mathbf{e}) \in S} \sum_{i=1}^{|e|} \log p(e_i | e_{<i}, \mathbf{c}_i; \Theta)$$

In order to achieve open-vocabulary translation with a limited vocabulary size, the words are segmented into sub-words, and the words with shared sub-words share part of their representation. The most common segmenting approach was introduced by Sennrich et al. (2015b) and exploits only statistical information, but there are promising research lines trying to use linguistically motivated segmentations (Ataman et al., 2017)

3 Using external word embeddings

The method we propose is based on the training of word embeddings from source-language monolingual data. We use these embeddings as an input to the network, and we remove the source-side embedding matrix. As the external embeddings have been learned for a task that is not machine translation, we introduce a feed-forward layer to map the embeddings into a new space that is more useful for the translation task:

$$\tilde{\mathbf{x}}_j = \tanh(\bar{\mathbf{x}}_j^\top \mathbf{W} + \mathbf{b}) \text{ for j = } 1, \ldots, m$$

where $\bar{\mathbf{x}}_j$ is the external embedding for the word j and the vectors $\tilde{\mathbf{x}}_i$ are used merged with the internal embeddings.
In this work we experimented three different settings: (1) *only external*, (2) *mix sum*, (3) *mix gate*. While *only external* is the setting we have just described above, the other two settings combine the external embeddings with the internal NMT embeddings. The *mix sum* setting inserts a vector sum between the embeddings and the RNN which simply sums the internal embedding for the word f_j and the mapped external embedding for the same word:

$$\hat{\mathbf{x}}_j = \mathbf{x}_j + \tilde{\mathbf{x}}_j$$

In the *mix gate* setting, we let the network learn parameters to combine the internal and the external embeddings. A gate is a function that produces a vector of the same dimensionality of the input, with all elements between 0 and 1 to represent the proportion of the corresponding input element that is propagated to the following layer:

$$\mathbf{z}_j = \sigma([\mathbf{x}_j; \tilde{\mathbf{x}}_j]^\top \mathbf{W_z} + \mathbf{b_z})$$

where $\mathbf{z}_j$ is the output of the gate and σ is the sigmoid function. The new vector is produced by combining linear transformations of the inputs with the gate $\mathbf{z}_j$:

$$\hat{\mathbf{x}}_j = \tanh(\mathbf{z}_j \odot ff_1(\mathbf{x}_j) + (1 - \mathbf{z}_j) \odot ff_2(\tilde{\mathbf{x}}_j))$$

Where ff is a feed-forward layer. In this setting the network has more parameters to learn for combining the internal and external embeddings in an effective way.

4 Experimental setup

Model	TED-14
Baseline	25.37
Only External Crawl	26.13
Mix Sum Crawl	**29.45**
Mix Gate Crawl	27.10

Table 1: Small data condition: BLEU score on IWSLT TED Talk Task En-Fr.

We performed our experiments on two tasks representing two different training conditions:

Model	NEWS-15	NEWS-16
Baseline	**16.67**	**20.07**
Only External Crawl	12.73	15.58
Mix Sum Crawl	15.59	18.72
Mix Gate Crawl	16.20	19.44
Only External news	13.36	16.40
Mix sum news	16.01	19.15
Mix gate news	16.45	19.35

Table 2: Large data condition: BLEU scores on WMT News Task En-De.

large data and *small data*. The first task is the 2017 WMT News translation task, from English to German, which provides a substantial amount of parallel data. For this experiment, we use all the available training data, about 5 million sentence pairs[1], newstest2013 and 2014 as a validation set and newstest2015 (NEWS-15) and newstest 2016 (NEWS-16) as test sets. The second task in the 2016 IWSLT TED Talk translation task, from English to French, for which we only deployed a small in-domain data set consisting of 200,000 sentence pairs, dev and test sets from 2010 to 2013 as a validation sets and the test set 2014 as test set (TED-14) [2].

We used two sets of pre-trained English word-embeddings. The first is the Common Crawl set available from the GloVe website[3], which contains $1.9M$ word embeddings (dim=300) trained with Glove (Pennington et al., 2014). The second set was instead created by us with *fastText* (Bojanowski et al., 2016) from the newscrawl 2015 and 2016 corpora (also available from the WMT 2017 website), which can be considered in-domain for the wmt task. We selected only words appearing at least 5 times in the corpus, and did not use any character n-gram information. This process produced embedding vectors (dim=500) of about $640K$ words in the news domain.

For all the experiments we used an NMT with 500 dimensions in the embeddings and in the hidden sizes of RNN. With the WMT dataset we used vocabularies of size $40,000$ in both sides. They are words in the source side and sub-words in the target side. For IWSLT we used $80,000$ words vocabularies, which cover more than 99% of the training set vocabulary. For the training we applied Adam (Kingma and Ba, 2014) with initial learning rate 0.0003 until convergence. As a codebase we used Nematus (Sennrich et al., 2017) for all of our experiments. The reported BLEU scores (Papineni et al., 2002) are computed with multi-blue.pl from the Moses suite on detokenized texts. The results are presented in Tables 2 and 1.

Table 3: Out-of-vocabulary words in internal and external vocabularies

	TED	News15	News16
Int	289	556	738
Ext Crawl	1581	4460	6532
Ext News	-	487	394
Both Crawl	176	477	625
Both News	-	352	285

	NEWS15	NEWS16	TED
Baseline	14	15	337
Only Ext. Crawl	10	8	687
Mix Sum Crawl	64	77	672
Mix Gate Crawl	102	337	689
Only Ext. News	10	7	-
Mix Sum News	132	335	-
Mix Gate News	85	117	-

Table 4: Numbers of generated unknown words in the translations.

5 Results and Discussion

Results show that our approach is greatly beneficial in our small data condition (table 1), improving up to 4 bleu scores with the simple strategy of summing the external and internal word embeddings. For the large-data condition (table 2) the picture is instead very different, as none of the settings using external embeddings reaches the results of the baseline.

In order to verify our hypothesis that external embeddings help to extend the vocabulary, we firstly counted the number of OOV words with respect to the internal and external vocabularies for each test set, and also the number of words that are unknown in both of them. The results listed in table 3 show that in the case of TED, the number of OOVs in both vocabularies is 39% smaller than in the internal vocabulary, but at the same time in the external vocabulary it is more than 5 times larger. In all the experiments, the embeddings trained on Gigacrawl have many more OOVs than the inter-

[1] http://www.statmt.org/wmt17/translation-task.html

[2] https://wit3.fbk.eu/mt.php?release=2016-01

[3] https://nlp.stanford.edu/projects/glove/

src	so I was trained to become a gymnast for two years in Hunan , China in the 1970s .
ref	J'ai été entraînée pour devenir gymnaste pendant 2 ans , dans la province d' Hunan en Chine dans les années 1970 .
baseline	J'ai été formée pour devenir gymnaste , pendant deux ans au Texas, en Chine dans les années 1970 .
mix-gate	J'ai donc été formé pour devenir une gymnaste pendant deux ans en UNK, en Chine dans les années 70.
src	Egyptologists have always known the site of Itjtawy was located somewhere near the pyramids of the two kings [...] .
ref	les égyptologues avaient toujours présumé qu' Itjtawy se trouvait quelque part entre les pyramides des deux rois [...] .
baseline	Nous avons toujours connu le site de Londres , situé quelque part prés des pyramides des deux rois [...]
mix-gate	Et on sait toujours que le site de UNK était situé quelque part près des pyramides des deux rois [...].

Table 5: Example translations with words that are out of one of the two vocabularies. In the first sentence "China" is not in the external vocabulary, but it is still trained properly. In the second sentence "Egyptologists" is not in the internal vocabulary. It cannot be translated at all, but the network finds a way to come around the problem.

nal counterpart, and the difference is particularly large in newstest15 and 16. This can be a reason for degradation of representations, unless the network learns to correct the noise coming from the external side.

To have a glimpse of the degradation, we also counted the number of generated unknown words for each test set. The results are listed in table 4. What we can observe is a slightly reduced number of unknown tokens in newstest when using only the external embeddings, but in a setting where the target side uses subwords. In all the other cases, the number of unknown words during translations increases dramatically. The increase is from 5 to 22 times in WMT and about 2 times in TED. Now we want to understand if this is due to a corrupted representation of words, which mixes good embeddings with the external embedding for the unknown token, or the reason is to find somewhere else. This is particularly true because of the contemporary improvement in BLEU score.

To verify the correction capabilities of the network, we check some translations where one word is missing in one of the two vocabularies. Two example translations are shown in table 5. In the first example, the word "China" exists only in the internal vocabulary, but it's correctly translated also by the mix-gate system. Furthermore, the baseline translates the OOV word "Hunan" with "Texas", while our system translates it with an unknown token. The second behavior is surely one of the main reasons of the increased number of generated unknown words using external embeddings, and it is also preferable as there are methods for replacing the unknown tokens in a postprocessing step. (Luong et al., 2015b).

In the second example, "Egyptologists" is OOV for the internal vocabulary. Lacking the subject, the baseline resorts to the first person plural, and it also adds a subordinate sentence that change s the meaning with respect to the source. Moreover, again an unknown word for a location is translated with another word that is related with the source only because it is another location (in this case the system translates with "Londres", which is the French word for "London"). By contrast, in absence of more information about the subject, the mix-gate uses the impersonal form and the grammar of its translation is better in general.

In the large-data setting, the best system using external embeddings is the mix-gate with data from the news domain. From table 3, we can relate the improvement also to the reduced number of external OOV words, but the improvement is so small that we suppose that using better corpora is not a path to follow. Moreover, our results lower than the baseline are an empirical proof that pre-trained embeddings are not useful when there are large parallel data available.

6 Conclusions

In this paper we propose three methods to extend the input word embeddings to an NMT network in order to leverage a word representation coming from a big monolingual corpus. Our results show that this approach greatly improves over an NMT baseline in a low-resource scenario, while it is not helpful for better-resourced tasks.

Using monolingual data for improving NMT is a problem also in the latter case, thus our future work will focus on how to integrate models larger than word embeddings, and trained on monolingual data, to improve word and sentence representations.

Acknowledgments

This work has been partially supported by the EC-funded projects ModernMT (H2020 grant agreement no. 645487) and QT21 (H2020 grant agreement no. 645452). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs used for this research.

References

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the fourth workshop on statistical machine translation*, pages 182–189. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *Proceedings of the First Conference on Machine Translation (WMT)*, volume 2, pages 131–198.

M. Cettolo, J. Niehues, S. Stker, L. Bentivogli, and M. Federico. 2016. The IWSLT 2016 evaluation campaign. In *Proceedings of the 13th Workshop on Spoken Language Translation, Seattle, pp. 14, WA*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015a. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Minh-thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in neural machine translation. In *In ACL*. Citeseer.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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. *arXiv preprint arXiv:1703.04357*.

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

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

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.