Multilingual Neural Machine Translation for Low Resource Languages

Surafel M. Lakew
University of Trento, Italy
Fondazione Bruno Kessler

Mattia A. Di Gangi
University of Trento, Italy
Fondazione Bruno Kessler

Marcello Federico
Fondazione Bruno Kessler

via Sommarive, 18, Trento, Italy
{lakew,digangi,federico}@fbk.eu

Abstract

Neural Machine Translation (NMT) has been shown to be more effective in translation tasks compared to the Phrase-Based Statistical Machine Translation (PBMT). However, NMT systems are limited in translating low-resource languages (LRL), due to the fact that neural methods require a large amount of parallel data to learn effective mappings between languages. In this work we show how so-called multilingual NMT can help to tackle the challenges associated with LRL translation. Multilingual NMT forces words and subwords representation in a shared semantic space across multiple languages. This allows the model to utilize a positive parameter transfer between different languages, without changing the standard attention-based encoder-decoder architecture and training modality. We run preliminary experiments with three languages (English, Italian, Romanian) covering six translation directions and show that for all available directions the multilingual approach, i.e. just one system covering all directions is comparable or even outperforms the single bilingual systems. Finally, our approach achieve competitive results also for language pairs not seen at training time using a pivoting (x-step) translation.

Italiano. *La traduzione automatica con reti neurali (neural machine translation, NMT) ha dimostrato di essere più efficace in molti compiti di traduzione rispetto a quella basata su frasi (phrase-based machine translation, PBMT). Tuttavia, i sistemi NMT sono limitati nel tradurre lingue con basse risorse (LRL). Questo è dovuto al fatto che i metodi di deep learning richiedono grandi quantit di dati per imparare una mappa efficace tra le due lingue. In questo lavoro mostriamo come un modello NMT multilingua può aiutare ad affrontare i problemi legati alla traduzione di LRL. La NMT multilingua costringe la rappresentrazione delle parole e dei segmenti di parole in uno spazio semantico condiviso tra multiple lingue. Questo consente al modello di usare un trasferimento di parametri positivo tra le lingue coinvolte, senza cambiare l'architettura NMT encoder-decoder basata sull'attention e il modo di addestramento. Abbiamo eseguito esperimenti preliminari con tre lingue (inglese, italiano e rumeno), coprendo sei direzioni di traduzione e mostriamo che per tutte le direzioni disponibili l'approccio multilingua, cioè un solo sistema che copre tutte le direzioni è confrontabile o persino migliore dei singolo sistemi bilingue. Inoltre, il nostro approccio ottiene risultati competitivi anche per coppie di lingue non viste durante il trainig, facendo uso di traduzioni con pivot.*

1 Introduction

Neural machine translation (NMT) has recently shown its effectiveness by delivering the best performance in various evaluation campaigns (IWSLT 2016 (Cettolo et al., 2016), WMT 2016 (Bojar et al., 2016)). Unlike rule-based or phrase-based MT, the end-to-end learning approach of NMT models the mapping from source to target language directly through a posterior probability. The basic component of an NMT system include an encoder, a decoder and an attention mechanism (Bahdanau et al., 2014). Despite the continuous improvement in performance and

translation quality, NMT models are highly dependent on the availability of large parallel data, which in practice can only be acquired for a very limited number of language pairs. For this reason, building effective NMT systems for low-resourced languages becomes a primary challenge (Koehn and Knowles, 2017). Recently, (Zoph et al., 2016) showed how a standard string-to-tree statistical MT system (Galley et al., 2006) can effectively outperform NMT methods for low-resource languages, such as Hausa, Uzbek, and Urdu. In this work, we focus on a so-called multilingual NMT (Johnson et al., 2016; Ha et al., 2016), which considers the use of NMT to target many-to-many language translation. Our motivation is that intensive cross-lingual transfer (Terence, 1989) via parameter sharing should ideally help in the case of similar languages and sparse training data. Hence, in this work we investigate multilingual NMT across Italian, Romanian, and English, and simulate low-resource conditions by limiting the amount of parallel data.

Our approach showed a BLEU increase in various language directions, in a low-resource setting. To compare a single language pair NMT models with a single multilingual NMT (M-NMT) model, we considered six translation directions (i.e English↔Italian, English↔Romanian, and Italian↔Romanian). For evaluating the zero-shot translation (i.e. a translation between language pair with no available parallel corpus), we removed the (Italian↔Romanian) language pairs. In the same way as the six-language-pairs, the performance of the four-language-pairs M-NMT model is comparable with the bilingual models for the language directions with parallel data.

We start in Section 2 with a brief description of NMT and state-of-the-art multilingual NMT approaches. In Section 3, we give a background on our M-NMT model. In Section 4, we present the experimental setting and the NMT model configurations. In Section 5, we show and discuss the results of the experiments. Finally, in Section 6 we present our conclusion and future works.

2 State of The Art

An NMT system consists of three different models called encoder, decoder and attention (Bahdanau et al., 2014). The encoder takes as an input a sequence of words $\mathbf{f} = f_1, \ldots, f_m$ in the form of vocabulary indexes, extract their embeddings and computes a contextual representation of the source words using an RNN implemented with an LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014):

$$\mathbf{h}_t = g(\mathbf{x}_t, \mathbf{h}_{t-1}) \quad t = 1, ..., m$$

where $\mathbf{x}_t$ is the embedding for the word at time step t and m is the length of the source sentence. The decoder receives as input the embedding of the target word at the previous decoding time step, and computes through a RNN a new representation of the current translation, given the representation in the previous step, and a relevant source context computed by the attention model. At each time step, the attention computes normalized weights for the source word positions according to the hidden state of the decoder, which are then used to compute the source context as a weighted sum of all the encoder hidden states. There are several strategies to implement a decoder but all of them end up computing the conditional probability of the next target word depending on the previously translated words and the source sentence:

$$p(e_i = k | e_{<i}, \mathbf{f})$$

The network is trained end-to-end to find the parameters $\hat{\mathbf{\Theta}}$ that maximizes the log-likelihood of the training set $\{(\mathbf{f}_s, \mathbf{e}_s) : s = 1, \ldots, S\}$:

$$\sum_{s=1}^{S} \log p(\mathbf{e}_s | \mathbf{f}_s; \Theta)$$

Based on the end-to-end training approach in NMT, M-NMT models translation across multiple languages with a single model. As such, a multilingual translation task can be categorized into many-to-one, one-to-many, or many-to-many directions, with increasing difficulty. By employing one of these scenarios, recent works in multilingual NMT have shown the possibility of translating language pairs never seen at training time, in addition to improving baseline bilingual NMT models (Ha et al., 2016) (Johnson et al., 2016).

The initial approaches to multilingual NMT required modifications on the standard encoder-decoder architecture (Zoph and Knight, 2016; Firat et al., 2016a; Firat et al., 2016b; Dong et al., 2015; Luong et al., 2015; Lee et al., 2016). Recently, state-of-the-art results are achieved by simply decorating the network inputs with special language tags, to direct the model to a preferred target

language at inference time. In this work, following (Johnson et al., 2016) we add a language token at the beginning of every source sentence. This token is unique for the target language and it is a way to impose the target language in which to translate (target-forcing).

3 M-NMT for Low-resource Languages

In this work, we show that it is possible to train a single NMT model for the translation task between multiple language pairs in a low-resource setting. In (Ha et al., 2016; Johnson et al., 2016) it has been shown that a multilingual system trained on a large amount of data improves over a baseline bilingual model, and it is also capable of performing zero-shot translation. In this work we focus on M-NMT in a resource-scarce (Koehn and Knowles, 2017) scenario and show how M-NMT is never worse than a bilingual system for each of the language directions used in the training phase. In fact, the multilinguality can be considered as a way to increase the available amount of data for language directions with small datasets. Moreover, only a single system is needed with respect to several bidirectional NMT systems, thus our setting also represents a way for saving training time and compresses the number of required parameters. The target language can be imposed on the network by using the previously described target forcing.

Furthermore, we use our multilingual model to perform zero-shot translation. We hope that by simply applying the target forcing in the zero-shot scenario, the system can generate sentences in the target language. An alternative zero-shot translation in a resource-scarce scenario can also be performed using a pivot language that is, using an intermediate language for translation. While this is a known technique in machine translation using two or more bilingual models, we expect to achieve a comparable pivoting results using a single multilingual model.

4 Experimental setting

Our NMT model uses embeddings with dimension 1024 and RNN layers based on GRUs of the same dimension. The optimization algorithm is Adagrad (Duchi et al., 2011) with an initial learning rate of 0.01 and mini-batches of size 100. Dropouts are used on every layer, with probability 0.2 on the embeddings and the hidden layers and 0.1 on the input and output layers. All experiments are done using the NMT toolkit Nematus[1] (Sennrich et al., 2017).

Pair	Train	Dev10	Test10	Test17
En-It	231619	1643	929	1147
En-Ro	220538	1678	929	1129
It-Ro	217551	1643	914	1127

Table 1: A total number of parallel sentences used for training and evaluation in a limited low-resource scenario.

For the training set, we used the dataset provided by the latest IWSLT2017[2] multilingual shared task for all possible language pair combinations between Italian, Romanian and English (Cettolo et al., 2012). At the preprocessing stage, we applied word segmentation by jointly learning the Byte-Pair Encoding (Sennrich et al., 2015), merging rules set to 39,500. There is a high overlap between the language pairs (i.e the English dataset paired with Romanian is highly similar to the English paired with Italian). Because of this overlapping, the actual unique sentences in the dataset are approximately the half of the total size. This consequently exacerbates the low-resource aspect in the multilingual models. The size of the vocabulary both in case of the bilingual and the multilingual models stays just under 40,000 sub-words. An evaluation script to determine the BLEU (Papineni et al., 2002) score is used to validate on the dev set and later to choose the best performing models.

We trained models for two different scenarios, the first is the multilingual scenario containing all the available language pairs, while the second scenario is the zero-shot using pivoting, which does not contain parallel sentences for the Romanian↔Italian language pairs. For development and evaluating the models, we used sets from the IWSLT 2010 (Paul et al., 2010) and IWSLT2017 evaluation campaign. The inference is performed using beam search of size 12.

5 Results

5.1 Bilingual Vs. Multilingual

In the first scenario, we compare the translation performance of independently trained bilingual

[1]Nematus- https://github.com/EdinburghNLP/nematus

[2]The International Workshop on Spoken Language Translation - http://workshop2017.iwslt.org/

models against the M-NMT model. In total there are six bilingual models, whereas the M-NMT is trained using the concatenation of all the six languages pair dataset, by just appending an artificial token on the source side. As shown in Table 2, the performance of our systems are evaluated on dev2010 and test2017.

Our preliminary experiments show that the M-NMT system favorably compares with the bilingual systems. Improvements are observed in several language directions, which are likely gained from the cross-lingual parameter transfer between the additional language pairs involved in the source and target side.

Direction	NMT	M-NMT
English→Italian	26.79	26.34
Italian→English	31.43	31.39
English→Romanian	21.55	**22.13**
Romanian→English	33.84	34.16
Italian→Romanian	15.60	15.92
Romanian→Italian	21.00	**21.60**

Table 2: Comparison between six bilingual models (NMT) against a single multilingual (M-NMT) model. A difference of ≥ 0.5 BLEU score is highlighted as bold.

Specifically, the M-NMT showed an improvement of $+0.58$ and $+0.60$ for En→Ro and It→Ro directions, while having only a small decrease in performance for the En→It and It→En directions (see Table 2).

Direction	NMT	M-NMT
English→Italian	27.44	**28.22**
Italian→English	29.9	**31.84**
English→Romanian	20.96	**21.56**
Romanian→English	25.44	**27.24**
Italian→Romanian	17.7	**18.95**
Romanian→Italian	19.99	**20.72**

Table 3: Comparison between six bilingual models (NMT) against a single multilingual (M-NMT) model on test2017.

For the evaluation using test2017, however, the M-NMT performed better in all directions than the NMT models (see Table 3). These results show that the M-NMT model performs either in a comparable way or outperforms the single language pair models in this resource-scarce scenario. Moreover, the simplicity of using a single model instead of six leaves a room for further improvements by incorporating more language pairs.

5.2 Pivoting using a Multilingual Model

The pivoting experiment is setup by dropping the Italian-Romanian language pairs from the six directions M-NMT model, which gives us a four directions multilingual model (we call it, PM-NMT), where all the configurations stays the same as in M-NMT. Our main aim is to analyze how a multilingual model can improve a zero-shot translation tasks using a pivoting mechanism, using English as a bridge language in the experiment. Moreover, the use of a multilingual model for pivoting is motivated by the results we acquired using the M-NMT.

Direction	P-NMT	PM-NMT	Δ BLEU
It→Ro	14.14	14.75	$+0.61$
Ro→It	20.16	19.72	-0.44

Table 4: Comparison of pivoting with two bilingual models (P-NMT) against pivoting one multilingual model (PM-NMT). Both approaches use English as the pivoting language. Italian-Romania data was excluded from the training data of the multi-lingual model.

The results in Table 4, show the potential, although partial, of using multilingual models with pivoting for unseen translation directions. The comparable results achieved in both directions speak to us in favor of training and deploying one M-NMT system instead of two distinct NMT.

Direction	P-NMT	PM-NMT	Δ BLEU
It→Ro	16.3	17.58	$+1.28$
Ro→It	18.69	18.66	-0.03

Table 5: Comparison of pivoting with two bilingual models (P-NMT) against pivoting one multilingual model (PM-NMT) using test2017 as the evaluation set.

From the evaluation results on *test2017*, we confirmed that M-NMT can achieve a comparable (Ro→It) or better (It→Ro) result over the two NMT systems used for pivoting. In future work, we will investigate if better performance in pivoting can be achieved by increasing the number of

languages covered by the M-NMT system (possibly related to the source and target languages), and/or by different choices of the bridging language.

6 Conclusions

In this paper, we used a multilingual NMT model in a low-resource language pairs scenario. We showed that a single multilingual system achieves comparable performances with the bilingual baselines while avoiding the need to train several single language pair models. Then, we showed how a multilingual model can be used for zero-shot translation by using a pivot language for achieving slightly lower results than a bilingual model trained on that language pair. As a future work we want to explore how the choice of different languages can enable a better parameter transfer in a single model, using more linguistic features of the surface word form, and how to achieve a direct zero-shot translation in a low-resource scenario without the pivoting mechanism.

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

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

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.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2016. The iwslt 2016 evaluation campaign. *Proc. of IWSLT, Seattle, pp. 14, WA, 2016.*

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*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL (1)*, pages 1723–1732.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multilingual neural machine translation. *arXiv preprint arXiv:1606.04164*.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 961–968. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

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

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

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.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the iwslt 2010 evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT) 2010*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. 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*.

Odlin. Terence. 1989. Language transfer-cross-linguistic influence in language learning. *Cambridge University Press. Cambridge Books Online.*, page 222, June.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *arXiv preprint arXiv:1601.00710*.

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