An Evaluation of Language-Agnostic Inner-Attention-Based Representations in Machine Translation

Alessandro Raganato,[*†] **Raúl Vázquez,**[*] **Mathias Creutz**[*] **and Jörg Tiedemann**[*]
[*]University of Helsinki, Department of Digital Humanities
[†]Basement AI
{name.surname}@helsinki.fi

Abstract

In this paper, we explore a multilingual translation model with a cross-lingually shared layer that can be used as fixed-size sentence representation in different downstream tasks. We systematically study the impact of the size of the shared layer and the effect of including additional languages in the model. In contrast to related previous work, we demonstrate that the performance in translation does correlate with trainable downstream tasks. In particular, we show that larger intermediate layers not only improve translation quality, especially for long sentences, but also push the accuracy of trainable classification tasks. On the other hand, shorter representations lead to increased compression that is beneficial in non-trainable similarity tasks. We hypothesize that the training procedure on the downstream task enables the model to identify the encoded information that is useful for the specific task whereas non-trainable benchmarks can be confused by other types of information also encoded in the representation of a sentence.

1 Introduction

Neural Machine Translation (NMT) has rapidly become the new Machine Translation (MT) paradigm, significantly improving over the traditional statistical machine translation procedure (Bojar et al., 2018). Recently, several models and variants have been proposed with increased research efforts towards multilingual machine translation (Firat et al., 2016; Lakew et al., 2018; Wang et al., 2018; Blackwood et al., 2018; Lu et al., 2018). The main motivation of multilingual models is the effect of transfer learning that enables machine translation systems to benefit from relationships between languages and training signals that come from different datasets (Ha et al., 2016; Johnson et al., 2017; Gu et al., 2018). Another aspect that draws interest in translation models is the effective computation of sentence representations using the translation task as an auxiliary semantic signal (Hill et al., 2016; McCann et al., 2017; Schwenk and Douze, 2017; Subramanian et al., 2018). An important feature that enables an immediate use of the MT-based representations in other downstream tasks is the creation of fixed-sized sentence embeddings (Cífka and Bojar, 2018).

However, the effects of the size of sentence embeddings and the relation between translation performance and meaning representation quality are not entirely clear. Recent studies based on NMT either focus entirely on the use of MT-based sentence embeddings in other tasks (Schwenk, 2018), on translation quality (Lu et al., 2018), on speed comparison (Britz et al., 2017), or only exploring a bilingual scenario (Cífka and Bojar, 2018).

In this paper, we are interested in exploring a cross-lingual intermediate shared layer (called *attention bridge*) in an attentive encoder-decoder MT model. This shared layer serves as a fixed-size sentence representation that can be straightforwardly applied to downstream tasks. We examine this model with a systematic evaluation on different sizes of the attention bridge and extensive experiments to study the abstractions it learns from multiple translation tasks. In contrast to previous work (Cífka and Bojar, 2018), we demonstrate that there is a correlation between translation performance and trainable downstream tasks when adjusting the size of the intermediate layer. The trend is different for non-trainable tasks that benefit from the increased compression that denser representations achieve, which typically hurts the translation performance because of the decreased capacity of the model. We also show that multilingual models improve trainable downstream tasks even further, demonstrating the additional abstraction that is pushed into the representations through additional translation tasks involved in training.

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 27–32
Florence, Italy, August 2, 2019.

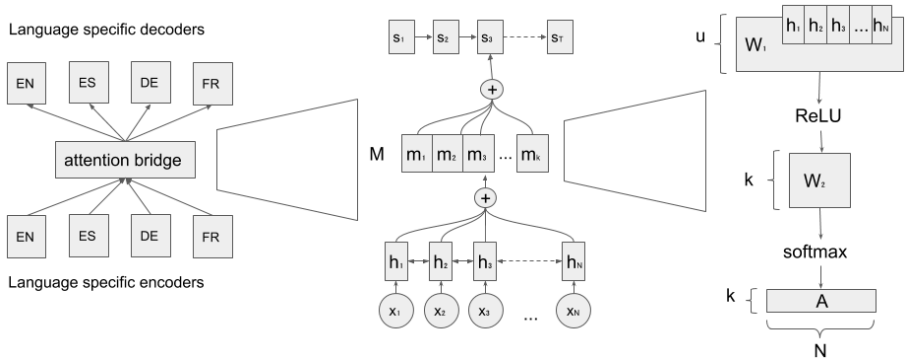

Figure 1: Architecture of our multilingual NMT system: (left) the attention bridge connects the language-specific encoders and decoders; (center) input $x_1 \dots x_n$ is translated into the decoder states $s_1 \dots s_t$ via the encoder states $H = h_1 \dots h_n$ and the attention bridge $m_1 \dots m_k$; (right) Computation of the hidden representation matrix A, needed to obtain the fixed-size attentive matrix $M = AH^T$.

2 Architecture

Our architecture follows the standard setup of an encoder-decoder model in machine translation with a traditional attention mechanism (Luong et al., 2015). However, we augment the network with language specific encoders and decoders to enable multilingual training as in Lu et al. (2018), plus we introduce an inner-attention layer (Liu et al., 2016; Lin et al., 2017) that summarizes the encoder information in a fixed-size vector representation that can easily be shared among different translation tasks with the language-specific encoders and decoders connecting to it. The overall architecture is illustrated in Figure 1 (see also Vázquez et al., 2019). Due to the attentive connection between encoders and decoders we call this layer *attention bridge*, and its architecture is an adaptation from the model proposed by Cífka and Bojar (2018). Finally, each decoder follows a common attention mechanism in NMT, with the only exception that the context vector is computed on the *attention bridge*, and the initialization is performed by a mean pooling over it. Hence, the decoder receives the information only through the shared attention bridge.

The fixed-sized representation coming out of the shared layer can immediately be applied to downstream tasks.[1] However, selecting a reasonable size of the attention bridge in terms of *attention heads* (m_i in Figure 1) is crucial for the performance both in a bilingual and multilingual scenario as we will see in the experiments below.

3 Experimental setup

All models are implemented using the OpenNMT framework (Klein et al., 2017) trained using the same set of hyper-parameters.[2] We use embedding layers of 512 dimensions, two stacked bidirectional LSTM layers with 512 hidden units (256 per direction) and an attentive decoder composed of two unidirectional LSTM layers with 512 units. Regarding the attention bridge, we experimented with four different configurations: 1, 10, 25 and 50 *attention heads* with 1024 hidden units each. For multilingual models, we used a language-rotating scheduler, in which each mini-batch contains sentences from a different language pair, cycling through all the language pairs uniformly. We selected the best model according to the BLEU score on the validation set. We train all the models using the Europarl Corpus v7 (Koehn, 2005), focusing on 4 languages: English (EN), French (FR), German (DE) and Spanish (ES). First we train bilingual models for EN→DE; then we train multilingual models {DE,ES,FR}↔EN; lastly we train a final Many-to-Many model using the biggest size, i.e., 50 *attention heads*, involving all translation directions between the three languages, i.e., we also include DE–ES, DE–FR and ES–FR.

To evaluate the sentence representations we utilize the SentEval toolkit (Conneau and Kiela, 2018) that combines various established downstream tasks for testing representations of English

[1] As in Lu et al. (2018), we note that the attention bridge is independent of the underlying encoder and decoder. While we use LSTM, it could be easily replaced with a transformer type network (Vaswani et al., 2017) or with a CNN (Gehring et al., 2017).

[2] Our fork implementation is available at https://github.com/Helsinki-NLP/OpenNMT-py/tree/att-brg.

		SNLI	SICK-E	AVG
en→de	k=1	63.86	77.09	71.46
en→de	k=10	**65.30**	78.77	72.02
en→de	k=25	65.13	79.34	**72.68**
en→de	k=50	**65.30**	**79.36**	72.60
Multilingual	k=1	65.56	77.96	72.67
Multilingual	k=10	67.01	79.48	72.89
Multilingual	k=25	66.94	79.85	**73.67**
Multilingual	k=50	**67.38**	**80.54**	73.39
Many-to-Many	k=50	**67.73**	**81.12**	**74.33**
Most frequent baseline†		34.30	56.70	48.19
GloVe-BOW†		66.00	78.20	75.81
Cífka and Bojar (2018) en→cs†		69.30	80.80	73.40

Table 1: Accuracy of different models on two SentEval tasks as well as the overall average accuracy on all of them. The general trend is that a higher number of attention heads and multilingual models are beneficial. Results with † taken from Cífka and Bojar (2018).

		SICK-R	STSB	AVG
en→de	k=1	0.74 / 0.67	**0.69 / 0.69**	**0.57**
en→de	k=10	0.76 / 0.71	**0.69 / 0.69**	0.52
en→de	k=25	**0.78 / 0.73**	0.67 / 0.66	0.49
en→de	k=50	0.78 / 0.72	0.65 / 0.64	0.46
Multilingual	k=1	0.76 / 0.71	0.69 / 0.68	**0.50**
Multilingual	k=10	0.78 / 0.74	**0.69 / 0.69**	0.48
Multilingual	k=25	0.78 / 0.74	0.68 / 0.67	0.43
Multilingual	k=50	**0.79 / 0.74**	0.66 / 0.64	0.40
Many-to-Many	k=50	**0.79 / 0.74**	**0.69 / 0.68**	0.40
InferSent†		0.88 / 0.83	0.76 / 0.75	0.66
GloVe-BOW†		0.80 / 0.72	0.64 / 0.62	0.53
Cífka and Bojar (2018) en→cs†		0.81 / 0.76	0.73 / 0.73	0.45

Table 2: Results from supervised similarity tasks (SICK-R and STSB), measured using Pearson's (r) and Spearman's (ρ) correlation coefficients (r/ρ). The average across unsupervised similarity tasks on Pearson's measures are displayed in the right-most column. Results with † taken from Cífka and Bojar (2018).

sentences.[3] In order to obtain a sentence vector out of multiple attention heads we apply mean pooling over the *attention bridge*.

We are also interested in the translation quality to verify the appropriateness of our models with respect to the main objective they are trained for. For this, we adopt the in-domain development and evaluation dataset from the ACL-WMT07 shared task. Sentences are encoded using Byte-Pair Encoding (Sennrich et al., 2016), with 32,000 merge operations for each language.

4 SentEval: Classification tasks

Table 1 shows the performance of our models on two popular tasks (SNLI and SICK-E) as in Cífka and Bojar (2018) as well as the average of all 10 SentEval downstream tasks. The experiments reveal two important findings:

(1) In contrast with the results from Cífka and Bojar (2018), our scores demonstrate that an increasing number of attention heads is beneficial for classification-based downstream tasks. All models perform best with more than one attention head and the general trend is that the accuracies improve with larger representations. The previous claim was that there is the opposite effect and lower numbers of attention heads lead to higher performances in downstream tasks, but we do not see that effect in our setup, at least not in the classification tasks.

(2) The second outcome is the positive effect of multilingual training. We can see that multilingual training objectives are generally helpful for the trainable downstream tasks.

Particularly interesting is the fact that the Many-to-Many model performs best on average even though it does not add any further training examples for English (compared to the other multilingual models), which is the target language of the downstream tasks. This suggests that the model is able to improve generalizations even from other language pairs (DE–ES, FR–ES, FR–DE) that are not directly involved in training the representations of English sentences.

Comparing against benchmarks, our results are in line with competitive baselines (Arora et al., 2017). While our aim is not to beat the state of the art trained on different data, but rather to understand the impact of various sizes of attention heads in a bi- and multilingual scenario, we argue that a larger attention bridge and multilinguality constitute a preferable starting point to learn more meaningful sentence representations.

5 SentEval: Similarity tasks

Table 2 summarizes the results using Pearson's and Spearman's coefficient on the two SentEval supervised textual similarity tasks, SICK-R and STSB, and the average Pearson's measure on the remaining unsupervised similarity tasks.

Two different trends become visible: i) On the unsupervised textual similarity tasks, having fewer attention heads is beneficial. Contrary to the results in the classification tasks, the best overall

[3] Due to the large number of SentEval tasks, we report results on natural language inference (SNLI, SICK-E/SICK-R) and the average of all tasks.

		k=1	k=10	k=25	k=50	M-to-M	att.
en	de	14.66	19.87	20.61	20.83	20.47	22.72
	es	21.82	27.55	28.41	28.13	27.6	30.28
	fr	17.8	23.35	24.36	23.79	24.15	25.88
de	en	16.97	21.39	23.42	24	24.4	24.28
es		18.38	25.39	27.01	27.12	26.98	28.16
fr		17.52	21.93	24.4	23.9	24.47	25.39

Table 3: BLEU scores for multilingual models. Baseline system in the right-most column.

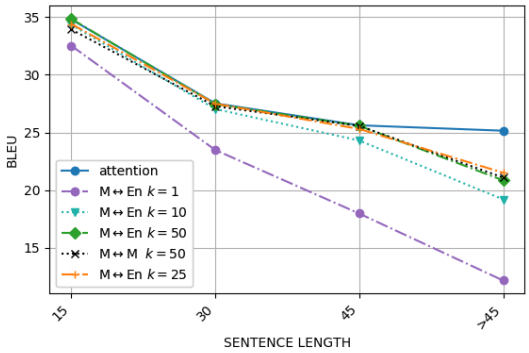

Figure 2: The BLEU scores obtained by the multilingual models and baseline system with respect to different sentence length.

model is provided by a bilingual setting with only one attention head. This is in line with the findings of Cífka and Bojar (2018) and could also be expected as the model is more strongly pushed into a dense semantic abstraction that is beneficial for measuring similarities without further training. More surprising is the negative effect of the multilingual models. We believe that the multilingual information encoded jointly in the attention bridge hampers the results for the monolingual semantic similarity measured with the cosine distance, while it becomes easier in a bilingual scenario where the vector encodes only one source language, English in this case.

ii) On the supervised textual similarity tasks, we find a similar trend as in the previous section for SICK: both a higher number of attention heads and multilinguality contribute to better scores, while for STSB, we notice a different pattern.

This general discrepancy between results in supervised and unsupervised tasks is not new in the literature (Hill et al., 2016). We hypothesize that the training procedure is able to pick up the information needed for the task, while in the unsupervised case a more dense representation is essential.

6 Translation quality

Finally, we also look at the translation performance of the multilingual models we have introduced above compared with a baseline, an standard encoder-decoder model with attention (Luong et al., 2015). In this section, we verify that the attention bridge model is stable and successfully learns to translate in the multilingual case.

Table 3 shows the comparison between the multilingual models. In general, we observe the same trend as in the bilingual evaluation concerning the size of the attention bridge. Namely, more attention heads lead to a higher BLEU score. The model with 50 heads achieves the best results among our models. It obtains scores that range in the same ballpark as the baseline, only in a few cases there is a degradation of few BLEU points. Notably, we do not see any increase in translation quality from the {DE,ES,FR}↔EN model to the Many-to-Many model; the BLEU scores are statistically equivalent for all six translation directions.

One of the main motivations for having more attention heads lies in the better support of longer sentences. To study the effect, we group sentences of similar length and compute the BLEU score for each group. As we can see from Figure 2 a larger number of attention heads has, indeed, a positive impact when translating longer sentences. Interestingly enough, on sentences with up to 45 words, there is no real gap between the results of the baseline model and our bridge models with a high number of attention heads. It looks like the performance drop of the attention bridge models is entirely due to sentences longer than 45 words.

We hypothesize that this might be due to the increasing syntactic divergences between the languages that have to be encoded. The shared self-attention layer needs to learn to focus on different parts of a sentence depending on the language it reads and, with increasing lengths of a sentence, this ability becomes harder and more difficult to pick up from the data alone.

7 Conclusion

We have shown that fixed-size sentence representations can effectively be learned with multilingual machine translation using a inner-attention layer and scheduled training with multiple translation tasks. The performance of the model heavily depends on the size of the intermediate representation layer and we show that a higher number of attention heads leads to improved translation and stronger representations in supervised

downstream tasks (contradicting earlier findings) and multilinguality also helps in the same downstream tasks. Our analysis reveals that the attention bridge model mainly suffers on long sentences. The next steps will include a deeper linguistic analysis of the translation model and the extension to multilingual models with more languages with greater linguistic diversity.

Acknowledgments

This work is part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113).

The authors gratefully acknowledge the support of the Academy of Finland through project 314062 from the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence and project 270354/273457. Finally, We would also like to acknowledge NVIDIA and their GPU grant.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *ICLR*.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122. Association for Computational Linguistics.

Ondej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurlie Nvol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors. 2018. *Proceedings of the Third Conference on Machine Translation*. Association for Computational Linguistics, Belgium, Brussels.

Denny Britz, Melody Guan, and Minh-Thang Luong. 2017. Efficient attention using a fixed-size memory representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 392–400.

Ondřej Cífka and Ondřej Bojar. 2018. Are bleu and meaning representation in opposition? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1362–1371. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proc. of ICML*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 344–354.

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

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics*, 5(1):339–351.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652. Association for Computational Linguistics.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *ICLR*.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv preprint arXiv:1605.09090*.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 228–234.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *ACL workshop on Representation Learning for NLP*.

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 (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *ICLR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. Multilingual NMT with a language-independent attention bridge. In *Proceedings of The Fourth Workshop on Representation Learning for NLP (RepL4NLP)*. Association for Computational Linguistics.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960. Association for Computational Linguistics.