Language-aware Interlingua for Multilingual Neural Machine Translation

Changfeng Zhu, Heng Yu, Shanbo Cheng, Weihua Luo
Machine Intelligence Technology Lab, Alibaba Group
{changfeng.zcf,yuheng.yh,shanbo.csb,weihua.luowh}
@alibaba-inc.com

Abstract

Multilingual neural machine translation (NMT) has led to impressive accuracy improvements in low-resource scenarios by sharing common linguistic information across languages. However, the traditional multilingual model fails to capture the diversity and specificity of different languages, resulting in inferior performance compared with individual models that are sufficiently trained. In this paper, we incorporate a language-aware interlingua into the Encoder-Decoder architecture. The interlingual network enables the model to learn a language-independent representation from the semantic spaces of different languages, while still allowing for language-specific specialization of a particular language-pair. Experiments show that our proposed method achieves remarkable improvements over state-of-the-art multilingual NMT baselines and produces comparable performance with strong individual models.

1 Introduction

Neural Machine Translation (NMT) (Sutskever et al., 2014; Vaswani et al., 2017) has significantly improved the translation quality due to its end-to-end modeling and continuous representation. While conventional NMT performs single pair translation well, training a separate model for each language pair is resource consuming, considering there are thousands of languages in the world. Therefore multilingual NMT is introduced to handle multiple language pairs in one model, reducing the online serving and offline training cost. Furthermore, the multilingual NMT framework facilitates the cross-lingual knowledge transfer to improve translation performance on low resource language pairs (Wang et al., 2019).

Despite all the mentioned advantages, multilingual NMT remains a challenging task since the language diversity and model capacity limitations lead to inferior performance against individual models that are sufficiently trained. So recent efforts in multilingual NMT mainly focus on enlarging the model capacity, either by introducing multiple Encoders and Decoders to handle different languages (Firat et al., 2016; Zoph and Knight, 2016), or enhancing the attention mechanism with language-specific signals (Blackwood et al., 2018). On the other hand, there have been some efforts to model the specificity of different languages. Johnson et al. (2017) and Ha et al. (2016) tackle this by simply adding some pre-designed tokens at the beginning of the source/target sequence, but we argue that such signals are not strong enough to learn enough language-specific information to transform the continuous representation of each language into the shared semantic space based on our observations.

In this paper, we incorporate a language-aware Interlingua module into the Encoder-Decoder architecture. It explicitly models the shared semantic space for all languages and acts as a bridge between the Encoder and Decoder network. Specifically, we first introduce a language embedding to represent unique characteristics of each language and an interlingua embedding to capture the common semantics across languages. Then we use the two embeddings to augment the self-attention mechanism which transforms the Encoder representation into the shared semantic space. To minimize the information loss and keep the semantic consistency during transformation, we also introduce reconstruction loss and semantic consistency loss into the training objective. Besides, to further enhance the language-specific signal we incorporate language-aware positional embedding for both Encoder and Decoder, and take the language embedding as the initial state of the target side.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1650–1655
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

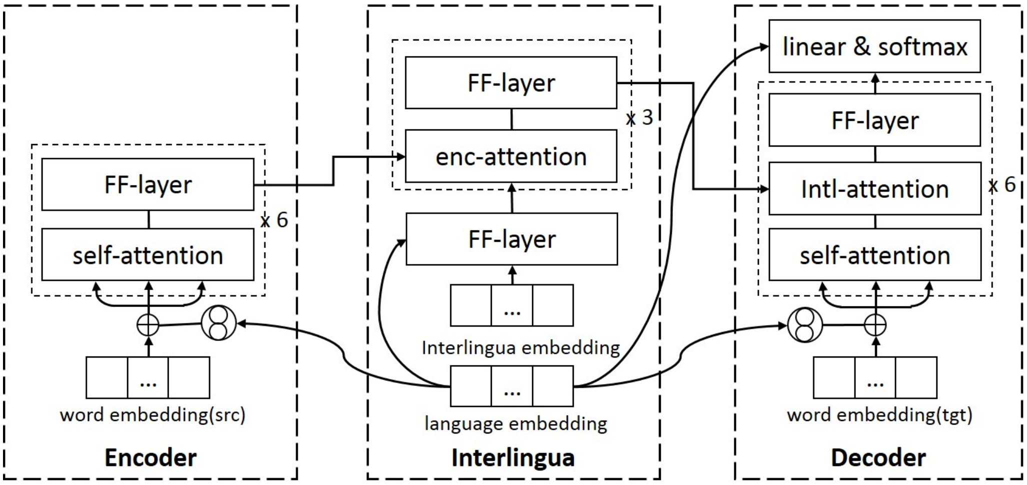

Figure 1: Our Encoder-Interlingua-Decoder architecture with a language-aware interlingua neural network.

We conduct experiments on both standard WMT data sets and large scale in-house data sets. And our proposed model achieves remarkable improvements over state-of-the-art multilingual NMT baselines and produces comparable performance with sufficiently trained individual models.

2 Model Architecture

As shown in Figure 1, we propose a universal Encoder-Interlingua-Decoder architecture for multilingual NMT. The Encoder and Decoder are identical to the generic self-attention TRANSFORMER (Vaswani et al., 2017), except some modifications in the positional embedding. The Interlingua is shared across languages, but with language-specific embedding as input, so we call it language-aware Interlingua. The Interlingua module is composed of a stack of N identical layers. Each layer has a multi-head attention sub-layer and a feed-forward sub-layer.

2.1 Interlingua

The Interlingua module uses multi-head attention mechanism, mapping the Encoder output H_{enc} of different languages to a language-independent representation I.

$$I = \mathbf{FFN}(\mathbf{ATT}(Q, K, V)) \quad (1)$$

$$Q = \mathbf{FFN}(L_{emb}, I_{emb}) \in \mathbb{R}^{d\times r} \quad (2)$$

$$K, V = H_{enc} \in \mathbb{R}^{d\times n} \quad (3)$$

The H_{enc} denotes the hidden states out of the Encoder, while the d is the hidden size, and the n denotes the length of the source sentence. $\mathbf{ATT}(.)$ is the multi-head attention mechanism (Vaswani et al., 2017). The (K, V) here are computed from the hidden states of the Encoder output H_{enc}. The Q is composed of two parts in simple linear combination. One part is from the language-specific part L_{emb}, and the other part is a shared matrix I_{emb}, which we called interlingua embedding. Note that, the interlingua embedding I_{emb} has a fixed size of $[d \times r]$. the i-th column of I_{emb} represents a initial semantic subspace that guides what semantic information of the H_{enc} should be attended to at the corresponding position i of the Interlingua output. The r means every Encoder H_{enc} will be mapped into a fixed size representation of r hidden states, and it is set to 10 during all of our experiments, similar to the work of (Vázquez et al., 2018). By incorporating a shared interlingua embedding, we expect that it can exploit the semantics of various subspaces from encoded representation, and the same semantic components of different sentences from both same and different languages should be mapped into the same position $i \in [1, r]$. Language embedding L_{emb} is used as an indicator for the Interlingua that which language it is attending to, as different languages have their own characteristics. So we call the module language-aware Interlingua. $\mathbf{FFN}(.)$ is a simple position-wise feed-forward network. By introducing Interlingua module into the Encoder-Decoder structure, we explicitly model the intermediate semantic. In this framework, the language-sensitive Enc is to model the characteristics of each language, and the language-independent Interlingua to enhance cross-language knowledge transfer.

2.2 Language Embedding as Initial State

The universal Encoder-Decoder model (Johnson et al., 2017) use a special token (e.g. <2en>) at the beginning of the source sentence, which gives a signal to the Decoder to translate sentences into the right target language. But it is a weak signal as the language information must go through $N = 6$ Encoder self-attention, and then $N = 6$ Encoder-Decoder attention before the Decoder attends to it. Inspired by Wang et al. (2018), we build a language embedding explicitly, and directly use it as the initial state of the Decoder.

2.3 Language-aware Positional Embedding

Considering the structural differences between languages, each language should have a specific positional embedding. Wang et al. (2018) use trigonometric functions with different orders or offsets in the Decoder for different language. Inspired by this, we provide language-aware positional embedding for both Encoder and Decoder by giving language-specific offsets to the original $sine(x)$, $cosine(x)$ functions in TRANSFORMER. The offset is calculated from $W_L L_{emb}$, where W_L is a weight matrix and L_{emb} is the language embedding.

2.4 Training Objective

We introduce three types of training objectives in our model, similar to (Escolano et al., 2019).

(i) Translation objective: Generally, a bilingual NMT model adopts the cross-entropy loss as the training objective, which we denote as $\mathcal{L}_{s2t}$, meanwhile, we incorporate another loss $\mathcal{L}_{t2s}$ for translation from the target to the source.

(ii) Reconstruction objective: The Interlingua transforms the Encoder output into an intermediate representation I. During translation, the Decoder only uses the I instead of any Encoder information. Inspired by Lample et al. (2017), Tu et al. (2017) and Lample et al. (2018), we incorporate an reconstruction loss for the purpose of minimizing information loss. We denote the $X' = \textbf{Decoder}(\textbf{Interlingua}(\textbf{Encoder}(X)))$ as the reconstruction of X. So we employ cross-entropy between X' and X as our reconstruction loss, and denote $\mathcal{L}_{s2s}$ for the source, $\mathcal{L}_{t2t}$ for the target.

(iii) Semantic consistency objective: Obviously, sentences from different languages with the same semantics should have the same intermediate representation. So we leverage a simple but effective method, $cosine$ similarity to measure the consistency. Similar objectives were incorporated in zero-shot translation (Al-Shedivat and Parikh, 2019; Arivazhagan et al., 2019)

$$sim(I^s, I^t) = \frac{1}{r}\sum_{i=1}^{r} \frac{I_i^s \cdot I_i^t}{\|I_i^s\| \|I_i^t\|} \tag{4}$$

Where, I^s and I^t denote the Interlingua representation of the source and target sides respectively. I_i is the i-th column of matrix I. $\mathcal{L}_{dist} = 1 - sim(I^s, I^t)$ is used as distance loss in our training objective.

Finally, the objective function of our learning algorithm is thus:

$$\mathcal{L} = \mathcal{L}_{s2t} + \mathcal{L}_{t2s} + \mathcal{L}_{s2s} + \mathcal{L}_{t2t} + \mathcal{L}_{dist} \tag{5}$$

3 Experiments

3.1 Experimental Settings

We conduct our experiments on both WMT data and in-house data. For WMT data, we use the WMT13 English-French (En-Fr) and English-Spanish (En-Es) data. The En-Fr and En-Es data consist of 18M and 15M sentence pairs respectively. We use newstest2012 and newstest2013 as our validation set and test set. Our in-house data contains about 130M parallel sentences for each language pair in En-Fr, En-Es, En-Pt (Portuguese), and 80M for En-Tr (Turkish). During all our experiments, we follow the settings of TRANSFORMER-base (Vaswani et al., 2017) with hidden/embedding size 512, 6 hidden layers and 8 attention heads. We set 3 layers for Interlingua, and $r = 10$ similar to the work of (Vázquez et al., 2018). We apply sub-word NMT (Sennrich et al., 2015), where a joint BPE model is trained for all languages with 50,000 operations. We used a joint vocabulary of 50,000 sub-words for all language pairs.

3.2 Experimental Results

3.2.1 Multilingual NMT vs Bilingual NMT

We take the UNIV model introduced by Johnson et al. (2017) as our multilingual NMT baseline, and individual models trained for each language pair as our bilingual NMT baseline.

The experimental results on WMT data are shown in Table 1. Compared with the UNIV

	one-to-many			many-to-one			zero-shot		
	En-Fr	En-Es	AVG	Fr-En	Es-En	AVG	Fr-Es	Es-Fr	AVG
INDIV/Pivot	**35.09**	**34.54**	**34.82**	32.91	33.48	33.20	**30.36**	**31.64**	**31.00**
UNIV	33.72	32.78	33.25	32.11	32.38	32.25	15.20	16.18	15.69
INTL	34.15	33.67	33.91	33.68	33.97	33.83	22.48	23.92	23.20
INTL+REC	34.97	34.28	34.63	**33.72**	**34.10**	**33.91**	23.69	25.16	24.43
INTL+SIM	34.09	33.56	33.83	33.54	33.95	33.75	25.93	26.81	26.37
INTL+REC+SIM	34.83	34.15	34.49	33.63	34.06	33.85	26.87	27.24	27.01

Table 1: BLEU scores on newstest2013. **INDIV** denotes direct model. **Pivot** is bridge translation system; **UNIV** denotes the universal framework introduced by Google (Johnson et al., 2017), but with a 9-layer Encoder. **INTL** refers to Interlingua model with only translation objective, and **REC**, **SIM** represent the reconstruction objective and the semantic consistency objective respectively.

	one-to-many					many-to-one				
	En-Fr	En-Es	En-Pt	En-Tr	AVG	Fr-En	Es-En	Pt-En	Tr-En	AVG
INDIV	**53.96**	34.53	52.97	**40.14**	45.40	59.01	36.92	53.87	38.63	47.11
UNIV	53.12	34.03	52.98	39.43	44.89	59.25	37.36	54.62	38.32	47.39
Ours	53.91	**34.71**	**53.95**	40.13	**45.68**	**60.15**	**38.27**	**55.57**	**38.77**	**48.19**

Table 2: BLEU scores on the 470M in-house data of four language pairs. **Ours** denotes Interlingua model with all training objectives

model (Johnson et al., 2017), our model get statistically significant improvements in both many-to-one and one-to-many translation directions on WMT data. Note that we set the Encoder of the UNIV model to 9 layers, which makes it comparable to this work in the term of model size. Compared with the individual models, our model is slightly better for Fr/Es-En in many-to-one scenario. In the one-to-many scenario, the individual models get the best BLEU score, while our model outperforms the universal model in all language pairs. Similarly, the experimental results on in-house large-scale data are shown in Table 2. In one-to-many settings, our model acquires comparable BLEU scores with the bilingual NMT baselines (Individual model), and around 1 BLEU point improvement in En-Pt translation. Our model gets the best BLEU score in many-to-one directions for all language pairs. Besides, the proposed model significantly exceeds the multilingual baseline (Universal model) in all directions. The results show that multilingual NMT models perform better in big data scenarios. This might the reason that intermediate representation can be trained more fully and stronger in a large-scale setting.

3.2.2 Zero-shot Translation

To examine whether our language-aware Interlingua can help cross-lingual knowledge transfer, we perform zero-shot translation on WMT data. The Fr-Es and Es-Fr translation directions are the zero-shot translations. As shown in Table 1, our method yields more than 10 BLEU points improvement compared with the universal Encoder-Decoder approach and significantly shortens the gap with sufficiently trained individual models.

3.2.3 Ablation study on training objectives

We further verify the impact of different training objectives in Table 1. Compared with the INTL baseline, the REC training objective can further improve the translation quality of both supervised and zero-shot language pairs. However, the SIM objective contributes to zero-shot translation quality significantly, with a slight decrease in supervised language pairs. The integration of both REC and SIM in INTL ultimately achieves balance increments between supervised and zero-shot language pairs. This suggests that constraints on Interlingua can lead to better intermediate semantic representations and translation quality.

4 Related Work

Multilingual NMT is first proposed by Dong et al. (2015) in a one-to-many scenario and generalized by Firat et al. (2016) to many-to-many scenario. Multilingual NMT suffered from the language diversity and model capacity problem. So one direction is to enlarge the model capacity, such as introducing multiple Encoders and Decoders to handle different languages (Luong et al., 2015; Dong et al., 2015; Firat et al., 2016; Zoph and Knight, 2016), or enhancing the attention mechanism with language-specific signals (Blackwood et al., 2018). The other direction is aimed at a unified framework to handle all language pairs (Ha et al., 2016; Johnson et al., 2017). They try to handle diversity by enhancing language-specific signals, by adding designed language tokens (Ha et al., 2016) or language-dependent positional encoding (Wang et al., 2018). Our work follows the second line by explicitly building a language-aware Interlingua network which provides a much stronger language signal than the previous works.

In regards to generating language-independent representation, Lu et al. (2018) and Vázquez et al. (2018) both attempted to build a similar language-independent representation. However, their work is all based on multiple language-dependent LSTM Encoder-Decoders, which significantly increase the model complexity. And they don't have the specially designed training objective to minimize the information loss and keep the semantic consistency. Whereas our work is more simple and effective in these regards and testified on a much stronger TRANSFORMER based system.

5 Conclusion

We have introduced a language-aware Interlingua module to tackle the language diversity problem for multilingual NMT. Experiments show that our method achieves remarkable improvements over state-of-the-art multilingual NMT baselines and produces comparable performance with strong individual models.

References

Maruan Al-Shedivat and Ankur P Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. *arXiv preprint arXiv:1904.02338*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. *arXiv preprint arXiv:1806.03280*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732.

Carlos Escolano, Marta R Costa-jussà, and José AR Fonollosa. 2019. From bilingual to multilingual neural machine translation by incremental training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 236–242.

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

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

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 for Computational Linguistics*, 5:339–351.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. *arXiv preprint arXiv:1804.08198*.

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*.

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

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.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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. 2018. Multilingual nmt with a language-independent attention bridge. *arXiv preprint arXiv:1811.00498*.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. *arXiv preprint arXiv:1902.03499*.

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.

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