What do Multilingual Neural Machine Translation Models learn about Typology?

Ryokan Ri and Yoshimasa Tsuruoka
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{li0123,tsuruoka}@logos.t.u-tokyo.ac.jp

1 Introduction

Unlike traditional statistical machine translation, neural machine translation (NMT) has enabled translation between multiple languages using a single model (Ha et al., 2016; Johnson et al., 2017). It enjoys easier maintainability and production deployment, without changing the model architecture and hurting performance so much.

However, its simplicity raises a natural question: how the multi-lingual model handle the multilingualism? Recent work has shown that the representations learned in neural machine translation models encode a great deal of linguistic information, such as morphology (Belinkov et al., 2017a; Bisazza and Tump, 2018), syntax (Shi et al., 2016), and semantics (Belinkov et al., 2017b; Poliak et al., 2018). However, most analyses are for models which translate in one direction, and only little is known about the linguistic competence of multilingual NMT models.

The goal of this work is to understand the following question: how much do multilingual NMT models capture the universality and variation of languages? Specifically, we will answer the following questions in this paper:

- How much typological information does each module in a model contain?
- How do the architectural choice of a model, specifically the subword or character model, affect the ability to capture linguistic typology?

The experimental design in this work is similar to Malaviya et al. (2017), where they trained a many-to-one multilingual NMT system to predict the typological features of the languages. However, we differ in that, while they focus on learning representation that can be used to predict missing features in typological databases, we aim to analyze the linguistic property of multilingual NMT models and conducted more fine-grained analyses.

2 Methodology

2.1 NMT Training

For a fair comparison among languages, we need sentences aligned in multiple languages. In this experiment, we use the Bible Corpus (Christodouloupoulos and Steedman, 2015), which contains translations of the Bible in 100 languages. We extracted verses aligned among 58 languages, and then split them into train/dev/test sets, which resulted in 23,555/455/455 verses respectively. The test data is used afterward for the following typological prediction task.

The NMT model is the attentional encoder-decoder model similar to Luong et al. (2015). The model has two stacked LSTM layers for the encoder and decoder, and the sizes of embeddings and hidden states are set to 500. The model is trained in the many-to-one scheme, *i.e.*, translating from multiple languages into one single target language. Following Johnson et al. (2017), we do not explicitly specify the source language which the model is translating.

Sentences are segmented by sentencepiece (Kudo, 2018), a language-agnostic tokenizer. For the source languages (57 languages in total), we created a shared vocabulary with the size of 32,000. For the target language (English), the vocabulary size is set to 8,000. We also experimented with character tokenization.

2.2 Probing Task

We investigate the extent to which the NMT model captures the typology of the source languages. We use the URIEL Typological Database

(Littell et al., 2017), which compiles typological features of languages extracted from multiple linguistics sources. We only use the syntactical features, which amount to 103 features, from the database (*e.g.*, S_SUBJECT_BEFORE_VERB, S_PLURAL_PREFIX), as we focus on features that would be directly learned in translation.

Our approach utilizes a probing task (Adi et al., 2017; Conneau et al., 2018). We train a logistic regression classifier with the sentence representations extracted from the trained model to predict the typological features of the source language. We performed 10-fold cross-validation, with no overlapping of languages in each train/test set. As the data in the test set is of the languages unknown to the classifier, the accuracy indicates how much the extracted representation generalizes about the typological features across languages.

3 Results and Discussion

This section presents the results for the two questions we asked: How much typological information does each module in a model contain?; How do the different architectures of a model, specifically subwords or characters model, affect the ability to capture linguistic typology? Table 1 summarizes the result with the majority baseline.

	Encoder	*Decoder*	*Attention*
majority	80.90%		
subword	84.90%	80.10%	81.30%
character	87.00%	80.10%	84.90%

Table 1: The accuracy of typology prediction using features extracted from different layers of the model. The values are averaged across all the predicted typological features and languages. *Encoder* and *Decoder* represents the output from the top layer of the encoder and decoder respectively. The lower layers gave lower accuracy in most cases. *Attention* is the representation after computing attention, before the output projection layer.

Effect of module

The representations from the encoder predict the typological feature of the source language significantly better than the majority baseline, whereas the decoder sees almost no improvements from the baseline. This indicates the encoder is aware of what language it encodes, whereas the decoder is ignorant of the source and focuses on generating the target language.

However, although the decoder is unaware of the source language, the representation from the attention, again, contains the typological information on the source language. This indicates the limitation of the current shared-attention architecture. Ideally, to achieve the most efficient parameter sharing in multilingual translation systems, target sequence generation should be ignorant of the source language properties, as the sentences with the same meaning are eventually mapped to the same target sequence regardless of the source language. In other words, the decoder has to generate a word sequence based on *interlingua*, *i.e.*, shared meaning representation across all languages (Richens, 1958; Schwenk and Douze, 2017; Johnson et al., 2017). This result confirms that attention is one of the obstacles to language-agnostic generation of the decoder, and is in line with recent efforts to improve multilingual NMT by seeking *neural interlingua* (Lu et al., 2018; Cífka and Bojar, 2018).

Subword vs. characters

The character model is more predictive of typological properties than the subword model in every layer. This can be attributed to the ability of the character model to capture morphology (Qian et al., 2016; Belinkov et al., 2017a), which is verified by the top 5 typological features that see improvement from the subword model to the character model. Three of them are features concerning part-of-speech (ADJECTIVE and NOUN), and one is about dependency marking:

- S_ADJECTIVE_AFTER_NOUN
- S_ADJECTIVE_BEFORE_NOUN
- S_INDEFINITE_WORD
- S_ADJECTIVE_WITHOUT_NOUN
- S_TEND_DEPMARK.

4 Conclusion

We will continue to experiment with another dataset, larger models, and different target languages to verify the observation in this paper and conduct further analyses. We also intend to use other probing tasks, such as universal part-of-speech tagging and natural language inference, to investigate the generalization ability of multilingual NMT models across languages.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Predicition Tasks. In *Proceedings of the International Conference on Learning Representations*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks . In *Proceedings of the International Joint Conference on Natural Language Processing*.

Arianna Bisazza and Clara Tump. 2018. The Lazy Encoder: A Fine-Grained Analysis of the Role of Morphology in Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Ondřej Cífka and Ondřej Bojar. 2018. Are BLEU and Meaning Representation in Opposition? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. *arXiv.org*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viegas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates . In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Patrick Littell, David Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

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 Conference on Machine Translation*.

Minh-Yhang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning Language Representations for Typology Prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018. On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Investigating Language Universal and Specific Properties in Word Embeddings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

R H Richens. 1958. Interlingual Machine Translation. *The Computer Journal*, 1(3):144–147.

Holger Schwenk and Matthijs Douze. 2017. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. In *Proceedings of the Workshop on Representation Learning for NLP*.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does String-Based Neural MT Learn Source Syntax? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.