Leveraging related languages for the translation of a low-resourced language pair: Manipuri to English

Rudali HUIDROM and Yves LEPAGE
Graduate School of Information, Production and Systems / EBMT/NLP Laboratory
Waseda University, Japan
rudali.huidrom@ruri.waseda.jp, yves.lepage@waseda.jp

Abstract

Low resourced languages are a challenge for neural machine translation, because neural methods require large amounts of data. We show that leveraging translation knowledge from languages related to the source language can improve translation accuracy for a low-resourced language, Manipuri.

Topics: information and communication model, language and media information

Introduction and background

We tackle the problem of machine translation from a low-resourced language called Manipuri (or Meiteilon) into English. Manipuri is an Indian language. It is a highly agglutinative language from the Sino-Tibetan language family. It uses the same writing system as Assamese and Bengali. Hindi is yet another Indian language. Although these Indian languages belong to different language families, they exert some lexical influence one onto another.

To improve the quality of translation from Manipuri into English, we propose to leverage translation knowledge from these three other Indian languages: Assamese, Bengali and Hindi.

Goals, proposal, method

We build different machine translation systems by adding knowledge from the three languages in combination with Manipuri and learn translation models into English. We measure by how much we can increase translation accuracy, in comparison with a baseline system where only Manipuri is used.

Manipuri	Assamese	Bengali	Hindi	BLEU	Confidence Interval
■				4.3	± 0.6
■	■			6.7	± 0.8
■		■		5.9	± 0.7
■			■	6.6	± 0.7
■	■	■		10.5	± 1.1
■	■		■	10.1	± 1.1
■		■	■	9.4	± 0.9
■	■	■	■	12.7	± 1.3

Table 1: Translation accuracies as measured with BLEU with 95 % confidence.

Experiments, results, analysis of the results

We use the OpenNMT-py toolkit (Klein et al., 2017) to implement our model: a two layered RNN model with bidirectional RNN as the encoder and a simple RNN as the decoder. We use the default values of the OpenNMT-py toolkit for the parameters.

We use data from the pmindia dataset (Haddow et al., 2020): 5,000 sentence pairs for training data and 1,000 sentence pairs each for test and validation in all the combinations. Since the amount of data for train, test, validation is still very small, we use a drop-out rate of 0.3.

The results in Table 1 show that the combinations perform better than the baseline (first line, Manipuri only). As we increase the number of similar languages on the source side, the accuracy of translation increases. The scores can be multiplied by more than 2, with statistical significance, when adding two languages; and almost by 3 when adding three languages (last line, all languages).

Conclusion

We showed how to improve the translation accuracy of a less-resourced language pair: Manipuri--English. For that, we exploited languages with very small resources which are related to Manipuri. As the resources we used are still small, in the future, we intend to use back-translation (Sennrich et al., 2016) to enlarge these resources and to study the application of unsupervised approach using adversarial training to learn a mapping from source to target without parallel data or anchor points (Lample et al., 2018).

References

Barry, H., & Faheem, K. (2020). A collection of parallel corpora of language of India. ArXiv, abs/2001.09907.

Guillaume, K., Yoon, K., Yuntian, D., Jean, S., & Alexander, R. (2017). Open-source toolkit for neural machine translation. *In Proceedings of ACL 2017* , (pp. 67-72). Vancouver.

Guillaume, L., Alexis, C., Marc'Aurelio, R., Ludovic, D., & Jégou, H. (2018). Word translation without parallel data. *In Conference Track Proceedings of ICLR 2018.* Vancouver.

Rico, S., Haddow, b., & Birch, A. (2016). Improving neural machine translation models with monolingual data. *In Prodeedings of ACL 2016.* Berlin.