Finding the Most Helpful Language to Adapt From for Endangered Languages

Chian-Yu Chen, Jean Lee, Zirui Li, Yu-Hsiang Lin, Yuyan Zhang, Graham Neubig
Language Technologies Institute
Carnegie Mellon University
{chianyuc, jeanl1, ziruil, yuhsianl, yuyanz1}@andrew.cmu.edu
gneubig@cs.cmu.edu

Abstract

Abstract.

1 Introduction

The common challenge of applying natural language processing (NLP) techniques to documenting the endangered languages is lack of language data. Moreover, among the limited data, there is often only a small portion of it that is annotated. Because the latest NLP technologies such as machine translation or speech recognition usually depends on a large quantity of annotated data, their performance is poor when directly applied to the endangered languages.

It has been shown that by using multi-lingual learning one can leverage one or more similar high-resource languages to improve the performance on the low-resource languages in several NLP tasks. One example is that by combining the training data of one or more high-resource languages with that of the target low-resource language to form a larger training dataset, one can obtain higher BLEU score in machine translation tasks (Neubig and Hu, 2018). It is therefore compelling to conduct a thorough investigation on the effective way of performing language adaptation in several common NLP tasks.

2 Finding the Most Helpful Language for Adaptation

The general question is: given a NLP task, a target low-resource language and its dataset, and some high-resource languages and their datasets, how can one find out which auxiliary high-resource language is the most helpful to adapt from, without exhaustively performing the task on all possible choices? To answer this question, we look at a few attributes that may be representative for the language and/or the particular dataset, and try to find the correlation of them to the quality of adaptation. The NLP tasks we consider are machine translation, entity linking, and [SOME TASK].

3 Experiments

Experiments.

4 Related Works

Related Works.

5 Conclusion

Conclusion.

References

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.