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

TODO: Worth citing? (Kocmi and Bojar, 2018)

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 questions we try to answer are:

1. 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?

2. Does there exist language or dataset features that are common strong indicators across multiple tasks? Or the strong features are highly task-dependent?

3. How does the performance of the method scales as the size of the dataset decreases? Could it be applied to resource-constrained endangered languages?

To answer the first question, we look at a few features that may be representative for the language and/or the particular dataset, and try to find the correlation between them and the quality of adaptation. More precisely, the features we consider include:

1. Dataset size
   For each corpus in different languages, we define the dataset size for that language the number of total word tokens in that corpus.

2. Type-token ratio (TTR) of the dataset
   The type-token ratio of a dataset is defined as the ratio of the types (the amount of unique words) to the number of total word tokens (Richards, 1987). It is a measure for lexical diversity, as lower TTR represents lower lexical variation and higher TTR represents the opposite.

3. Word-level/character-level overlap ratio between the target language dataset and auxiliary language dataset
   For the word-level overlap ratio of the datasets, we count the number of the tokens that appeared to be the same in both target and auxiliary language datasets. As for the character-level overlap ratio of the datasets, it is defined to be the overlapping of the sequence of characters in both target and auxiliary language datasets. For different lan-

guages, these sequences of characters are modelled separately.

4. URIEL distance between task and auxiliary language (dataset independent)
   The URIEL typological database (Littell et al., 2017) provides varies kinds of information and features for different languages. Features such as geographical distances between languages, and also vectors such as phylogeny vectors and typological vectors are all provided. We queried multiple kinds of features for each languages from the URIEL knowledge database and took them into consideration. These feature are dataset independent, so it is relatively easier to extract even for low-resource languages.

5. Earth-mover distance

We formulate our problem as: given the dataset of the low-resource task language, and a set of datasets of the high-resource auxiliary languages, predict which auxiliary language would help improving the performance most. There are at least three possible paradigms to address this prediction problem:

1. Regression: directly predict the task metric score.

2. Ranking: predict the order of the auxiliary language according to how much they improve the performance.

3. Binary classification: only predict which one language will be the most helpful.

To answer the second question, we consider the following common NLP tasks: machine translation, entity linking, and [SOME TASK]. In choosing the models, we prefer the ones that are easy to interpret, so that it is easier to tell the relative importance among the features. In this work, we consider decision trees as our models.

To answer the third question, we start with languages that include both low-resource and high-resource languages, and decrease the data size to observe the effect. We use TED dataset that include 54 languages.

3 Experiments

Experiments.

4 Related Works

Related Works.

5 Conclusion

Conclusion.

References

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proc. WMT*.

Patrick Littell, David R 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 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.

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.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.