Instructor: Rajesh Kumar Mundotiya

Aligning Word Vectors on Low-Resource Languages with Wiktionary

Group F

Vishesh Thakur

Shaleen Malik

Vijay Rakshit

Submitted: November 23, 2023

Contents

1 Introduction

This review report references the paper Aligning Word Vectors on Low-Resource Languages with Wiktionary by Mike Izbicki. Most alignment and evaluation tasks to date use the MUSE lexicon collection which provides bilingual lexicons between 45 languages and English. This lexicon is generated using a machine translation system and therefore has several errors like repetitive words, excessive proper nouns, adjectives etc. So, this paper delves into an innovative methodology that uses the high quality user-annotated Wiktionary corpus to create BLI datasets in 298 languages, 253 of which have not been made earlier. Through this paper, the datasets are also benchmarked against the MUSE corpus mentioned above to test their quality and competitive standing. The paper also focuses on the results achieved by training BLI models on these datasets and comparing the accuracy achieved.

2 Problem Statement

This paper introduces a novel bilingual lexicon collection derived from Wiktionary, a collaboratively edited resource encompassing over 7 million words across 8166 languages and involving contributions from 3.9 million users. The key contributions of this work are twofold:

1. Creation of High-Quality Bilingual Lexicons:

    - The authors leverage Wiktionary to construct bilingual lexicons tailored for training and evaluating Bilingual Lexicon Induction (BLI) models for 298 languages into English.
    - A noteworthy aspect is the inclusion of languages that are highly low-resource and, in some cases, extinct. The dataset not only introduces the first-ever BLI

datasets for 253 of these languages but also enhances the quality of existing datasets for the remaining 45 languages.

2. Training the Largest Collection of BLI Models:

   - The paper surpasses existing efforts by training the most extensive collection of BLI models to date.
   - Building upon the work of Grave et al. (2018), who provided pretrained word vectors in 157 languages, the authors extend this by training BLI models between each of these languages and English.
   - Notably, 112 of these languages had not been studied in the context of BLI before due to the absence of training/evaluation data. The paper identifies 15 of these previously unstudied languages as exhibiting particularly commendable performance, contributing valuable insights into the effectiveness of BLI models across diverse linguistic contexts.

3 Earlier Works

Previous endeavors, were predominantly reliant on the MUSE lexicon collection. This collection provides bilingual lexicons between 45 languages and English. This lexicon is generated from a machine translation system, and so suffers from a number of problems. First, many of the mappings in the lexicon do not contain real words in either the source or target language. Second, the distribution of words is inconsistent between languages, with many languages containing only proper nouns in their training and test sets erroneous mappings, distributional inconsistencies, and the glaring limitation of coverage for low-resource languages. Recognizing the pressing need for datasets that facilitate meaningful cross-lingual performance evaluations in the BLI domain, this report takes a bold departure from conventional sources and explores the untapped potential of Wiktionary, a

collaborative and expansive linguistic repository.

4 Novelty

The Existing dataset consisted 18 Indian languages out of the total of 22 scheduled Indian languages. As a novelty, we have included Bodo as a new language, which is extremely low resource. We have prepared a small dataset on Bodo, and evaluated it using the scored mentioned in the paper.

The Alignment MSE that we obtained 3.037.

We have used a new UMAP algorithm to learn a transformation from the embeddings of Bodo to English.

We used Procrustes and UMAP for alignment part, and our model seems to be working almost equally well as compared to Procrustes.

We also tried to experiment with the behaviour that what is the comparison of MSE loss when we gradually increase the size of the dictionary. We experimented with dictionary sizes of 20, 40, 50, 60, 70 and 80.

The results of all these experiments have been added separately.

5 Experimental Architecture

The experimental architecture unfolds with the systematic extraction of multifaceted data from Wiktionary, capturing essential elements such as language, part of speech, and nuanced English definitions. A meticulous division of the dataset into training and test sets ensues, ensuring a delicate balance of uniformity and fairness across languages.

- Impact of size of BLI training dataset on model performance
- Comparing quality of MUSE and Wiktionary lexicons

- Training BLI models on 112 new, previously unstudied languages

For each of these, the VecMap model has been trained using the common crawl vectors in the language that are then aligned to English language vectors trained on the common crawl data. The iterative normalization processing procedure used to transform the source and target language vectors before learning.

For the last experiment, in addition to the VecMap model, the Procrustes and Bootstrap Procrustes models are also used. These models are trained in all the languages where the parallel MUSE lexicons are also available. In addition to this, the results are also calculated for a smaller test set to ensure that the alignment of lower-resource languages is also being judged properly.

The results for all these cases are detailed out in the paper and some conclusive inferences can definitely be drawn.

6 Experimental Settings

The experiments unravel using the common crawl vectors provided by Grave et al. (2018), meticulously aligned to English-language vectors trained on the common crawl dataset by Mikolov et al. (2018). The iterative normalization preprocessing procedure takes center stage, systematically transforming both source and target language vectors. This has been done for all the 298 languages wherever possible. The previously trained high-quality embeddings are used along with fastText models available.

7 Results and Analysis

- Training Dictionary Size

  The experimental findings elucidate that BLI accuracy experiences rapid improvement until the number of training samples reaches a critical threshold of 5k. Beyond

this point, the incremental gains taper off, marking a pragmatic balance between computational efficiency and statistical robustness till 20K. Pragmatically, truncating the training set size to 20k is deemed computationally efficient because beyond that point the accuracy hardly shows any increase. Thus, for all the following experiments, only 5K points are used to make the experiments on such a large number of languages efficient and computationally feasible.

- MUSE Corpus vs Wiktionary Corpus

  Intriguingly, the MUSE training set outperforms the Wiktionary training set for 22 out of 45 languages. This unexpected outcome suggests that, despite the high-quality nature of Wiktionary, additional data from diverse sources could further enhance vector space alignment. The nuanced interplay between the size of training sets and the richness of lexical information emerges as a key focal point, underscoring the dynamic nature of linguistic data. It is hypothesized by the author that this might be happening because of a combination of the below two factors:

  - large size of the MUSE corpus in comparison to the Wiktionary corpus as the effect is mostly observed in high resource languages.
  - bias of the Wiktionary corpus towards the unconjugated or the dictionary form of words, which is not a problem with the MUSE dataset. This also contributes to the larger size of MUSE corpus.

  Therefore, we infer that even though accuracy does start to stagnate after the 5K or the 20K mark as seen in the first experiment, but still, for high-resource languages, it does have a significant effect.

- The Grave et al. (2018) Languages

  The Wiktionary corpus emerges as the pioneering publicly available dataset for training and testing alignment models in 112 languages not covered by the MUSE corpus.

Fifteen languages exhibit competitive performance, showcasing the adaptability and robustness of the proposed methodology. These languages, including Esperanto, Galician, and Armenian, present a significant breakthrough, indicating their suitability for downstream cross-lingual tasks. The observed disparities in performance between high-resource and low-resource languages are ascribed to the quality of word vectors trained on smaller datasets, thereby shedding light on the intricate balance between data quality and linguistic complexity.

We can also observe from the results displayed that the VecMap model mostly gives a better accuracy on the high-resource languages while the Procrustes model generally outperforms other models on the lower-resource language sets.