Point of pivot: cross-lingual embeddings calibration for Southern Nguni and Niger-Congo low-resourced languages.

Thapelo Sindane[1], Michelle Terblanche[1] , Vukosi Marivate [1 2]
[1]University of Pretoria
[2]Lelapa AI

Abstract

Analytics for transfer learning on low resourced languages are still grounded by the English-X outset even though recent research shows that English is not always the best source language. Causality can be traced back to accessibility, availability, trend, and most importantly expertise of Natural Language Processing resources. However, with some constraints such as the availability of digital data being gradually loosened in research, the need to explore other pivot alternatives becomes unequivocal. However, the point of pivot language has become a critical concern even though linguistic insights on language intelligibility is sometimes available. In this paper, we create all possible pairs of cross-lingual embeddings for Southern Nguni and Niger-Congo languages of South Africa and investigate analyses of all combinations on word similarity and downstream tasks relative to transfer learning. Our preliminary intrinsic evaluations indicate a conflicting outcome that mutually intelligible languages do not always generate supreme entangled representations. That is, languages belonging to the same language family generate sub-standard cross-lingual representations. Intrinsic evaluation will consider available annotated downstream tasks such as Named Entity Recognition, Part of Speech Taggings, Machine Translation with the intentions of establishing point of pivot insights for South African Languages

Keywords— Monolingual Embeddings, Cross-lingual Embeddings, Transfer learning, Intrinsic and Extrinsic evaluations, Cosine similarity, Named Entity Recognition and Part of Speech tagging

1 Introduction

The Constitution of South Africa recognises 11 official languages. Figure 1.1 shows the distribution of the different languages according to first (L1) and second (L2) language speakers . As can be seen, a big portion of the African languages are spoken by fewer people hence deeming them as low-resourced with low discoverability [1]. Recent research into zero-shot and few-shot learning has focused on utilising the knowledge from a high resource language such as English to apply similar word representations to a lower resource language being typologically similar to the high resource language. This avenue paves the way to introduce and improve NLP applications in communities that were previously lacking exposure to and the benefits of such technologies.

[6] first introduced the concept of representing words as vectors which evolved into using the approach the develop representations for sentences and even documents. In layman's terms, words/sentences of similar meaning should have similar vector representations. Similarity is measured using cosine similarity1: the inner product of the vectors. When visualising these vectors using a dimensionality reduction technique such as t-distributed Stochastic Neighbor Embedding (t-SNE) will show clusters for similar words or group of words.

The aim of this work is to highlight the South African languages that transfer well and that can be more easily exploited to develop mutual resources for enhancing NLP research using the measure of similarity. Furthermore, the intrinsic evaluation will allow us to understand where languages remain under-represented and future work should be aimed.

2 Related Works

Generating entangled representations using cross-lingual mathematical models have shown practical immanence, especially when supplemented by accompanying resources such as bilingual lexicons [7, 4, 1]. However, the choice of pivot language from which to base the projections from together with language closeness evaluations and determination remains unattempted. Recent works rely on linguistic insights for determining language intelligibility

[1]https://www.masakhane.io/

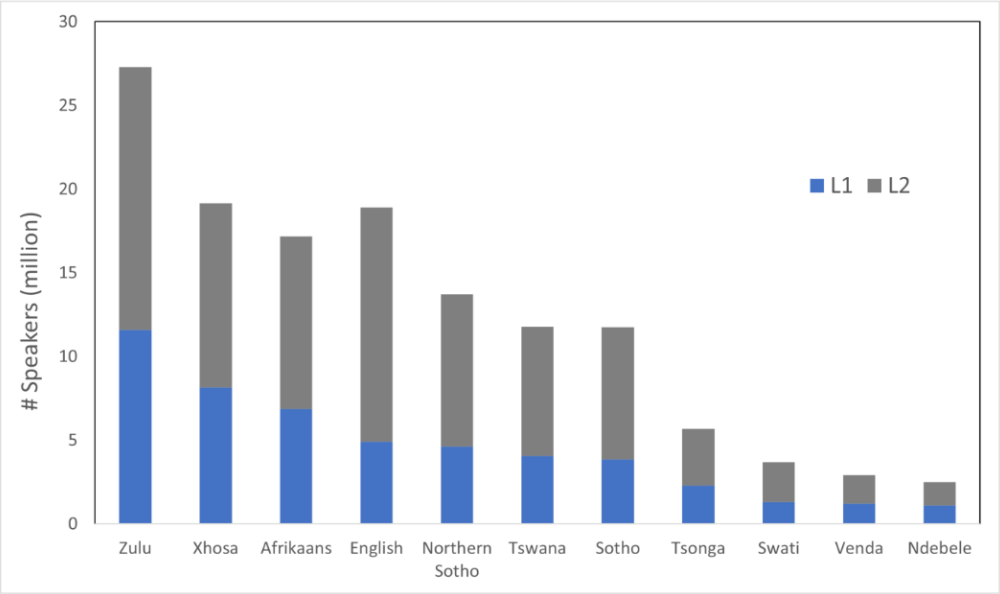


Figure 1: Caption

for which this knowledge is used to select a pivot language. Regardless, practical evidence of mutually intelligible languages being the best choice for point of pivot remains sparse. In this work, we explore a brute-force approach to investigate language combinations with better projection capabilities and later transfer performance.

3 Methodology

Bilingual-lexicons aided cross-lingual representation learning have shown better results compared to unsupervised learning counterparts. For this reason, all our cross-lingual representation learning focuses on the use of bilingual lexicons for training our three projection techniques: Muse, VecMap, and Canonical correlation analysis. In this section, we discuss the datasets collected (bilingual lexicons and monolingual data), projections models, and evaluation strategies.

3.1 Data

This section discusses various types of datasets collected in this study.

3.1.1 Bilingual lexicons

Our bilingual lexicons are collected from the government public school repositories [2], cput [3], etc. Figure 2, shows the distributions of all lexicons collected for all combinations of 11 languages in South Africa.

3.1.2 Monolingual data

Our monolingual data used to train our FastText monolingual embeddings are sourced from multiple repositories including Flores, WMT, MC4, NCHLT, and African Dataset.

3.1.3 Cross-lingual models

We investigate three projection strategies for creating cross-lingual embeddings, namely, Muse [5], VecMap [2], and Canonical correlation analysis [3].

3.2 Evaluations

Our evaluation strategy is two-tiered. First, we intrinsically evaluate our projected representations using cosine similarities. That is, we expect sampled translation pairs to have a high cosine similarity score. Likewise, non-translation-pairs are expected to have a low similarity score. This is done to evaluate if the projection model was able to project similar embedding of the same space and dissimilar embeddings to dislocated spaces. Our last tier

[2]https://github.com/dsfsi/gov-za-multilingual
[3]https://mlg.cput.ac.za/

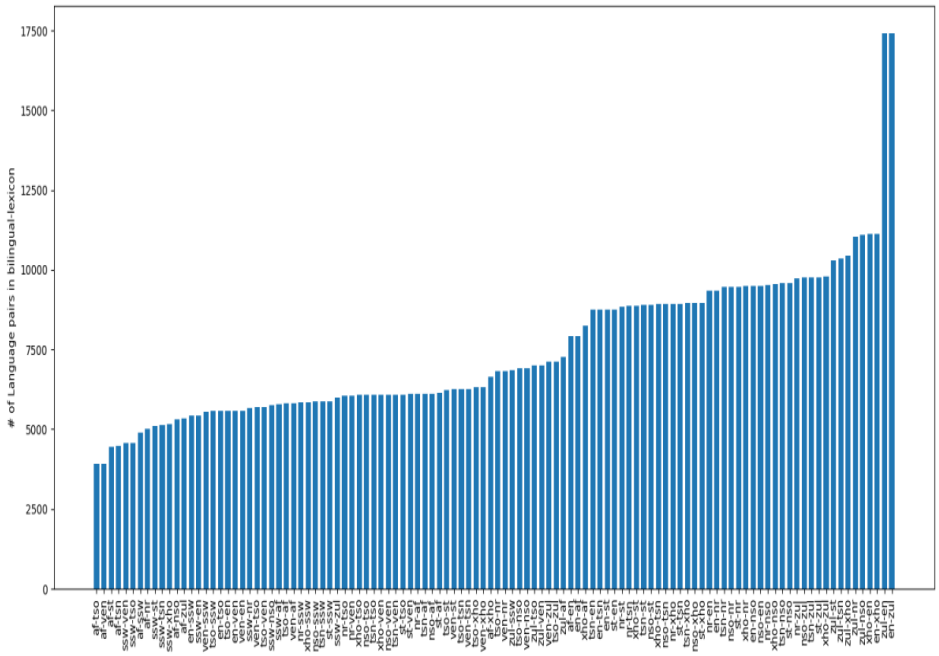


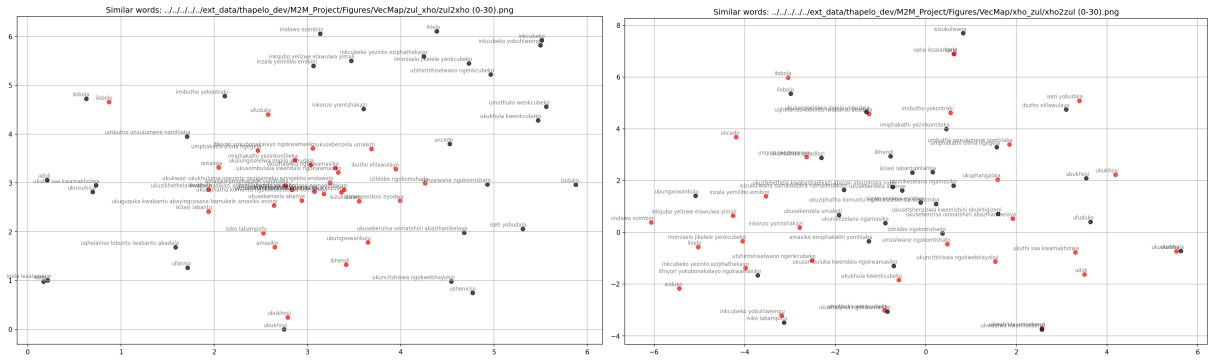
Figure 2: Caption

looks at Extrinsic evaluation, where cross-lingual embeddings are evaluated on downstream tasks: NER, POS, and Machine Translation.

4 Results

Figure 2 shows the distribution of our massively multilingual lexicons set collected for this study. The multilingual lexicons cover all 11 official South African languages and all possible combinations.

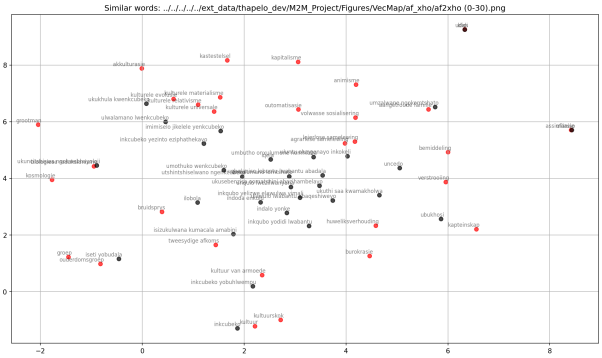
Figure 4, and 4 present a sample of the 110 plot for each of the possible combinations creations in this study. Preliminary results show that, the point of pivot is important in creating shared representations and IsiXhosa to IsiZulu generates better representations compared to IsiZulu to IsiXhosa. Additionally, our results indicated that language intelligibility is not always a good indicator of projection robustness as some mutually intelligible embedding show sub-optimal performance. However, we still aim to explore this as it may be a result of model incapacity. That is, these results are based on one of the models (VecMap).



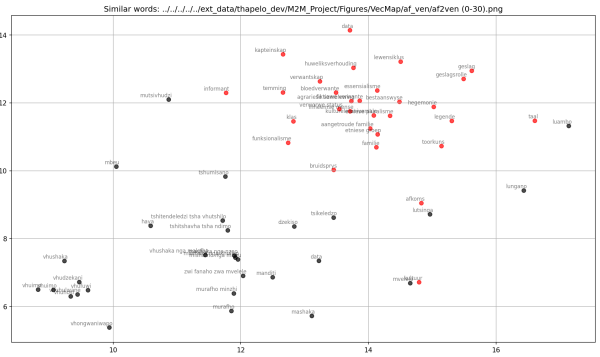
(a) Zulu to Xhosa

(b) Xhosa to Zulu

Figure 3: IsiXhosa and IsiZulu as pivot languages.



(a) Afrikaans and IsiXhosa projections



(b) Afrikaans and Tshivenda projections

Figure 4: Afrikaans as pivot languge

5 Conclusion

We created 110 combinations of cross-lingual embeddings from 11 South African languages and evaluated them with cosine similarity scores. Our preliminary results indicate that language intelligibility is not always a common indicator of increased projection performance. This is indicated by languages belonging to the same language family having dislocated word representations in the cross-lingual space. We intend to confirm if this is consistent with other projections techniques or it is primarily the limitation of the initial projection model.

References

[1] Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C. and Smith, N. A. [2016], 'Massively multilingual word embeddings', *arXiv preprint arXiv:1602.01925* .

[2] Artetxe, M., Labaka, G. and Agirre, E. [2018], A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 789–798.

[3] Faruqui, M. and Dyer, C. [2014], Improving vector space word representations using multilingual correlation, *in* 'Proceedings of EACL'.

[4] Glavas, G., Litschko, R., Ruder, S. and Vulic, I. [2019], 'How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions', *arXiv preprint arXiv:1902.00508* .

[5] Lample, G., Conneau, A., Denoyer, L. and Ranzato, M. [2017], 'Unsupervised machine translation using monolingual corpora only', *arXiv preprint arXiv:1711.00043* .

[6] Mikolov, T., Chen, K., Corrado, G. and Dean, J. [2013], 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

[7] Mikolov, T., Le, Q. V. and Sutskever, I. [2013], 'Exploiting similarities among languages for machine translation', *arXiv preprint arXiv:1309.4168* .


```
!!! File not found or not readable: output.tex !!!
(errors:1)
```