Tigrinya Dialect Identification

Anonymous authors
Paper under double-blind review

Abstract

Dialect Identification is an important topic of research in Natural Languge Processing (NLP) as it has broad implications in many real-world applications such as machine translation, speech recognition and document summarizing to name a few. In this work, we explore the problem of Tigrinya dialect identification using machine learning techniques. To that end, we have identified three Tigrinya dialects, namely: Z, L and D. Then we systematically collected data sets for each dialect. Finally, we perform experiments using classical machine learning and deep learning methods to quantify effectiveness of current approaches on the problem of Tigrinya dialect identification. We achieved a best score of 92.98% overall accuracy using character level CNNs.

1 Introduction

Tigrinya is the official regional language of Tigray, the northern part of Ethiopia. It is also among the official national languages of Eritrea Gebregzabiher (2010). In this work, we define the dialect identification problem as a multi-class classification task. Given a chunk of Tigrinya text T, we train models that determine which Tigrinya dialect T was written in. Our main contributions are: (i) unified categories of Tigrinya dialects and (ii) annotated dataset for the study of Tigrinya dialects.

2 Related Work

In ASFAW (2018) Rediat conducted a comparative study of automatic language identification for the same Ethio-Semitic languages - Amharic, Geez, Gurage and Tigrinya using character n-grams. The author compared Naïve Bayes with Cumulative Frequency Addition and concluded that the latter performed better. Similar to this, our work explores the use of character n-grams for dialect identification. In Feleke (2017) Tekabe examined the similarity and the mutual intelligibility between Amharic and Tigrinya using Levenshtein distance, intelligibility test and questionnaires. He found both Tigrinya varieties have almost equal phonetic and lexical distances from Amharic. In this work, we study the different dialects in Tigrinya. The other line of work that's directly relevant to ours is work done on dialect identification of other languages. In Mengistu & Alemayehu (2017) Abrham and Dagnachew studied four dialects of Amharic. They applied text independent back-propagation ANN, VQ (vector quantization), GMM and a composite of GMM and back-propagation ANN to categorize Amharic speakers into four dialects. They used MFCC (Mel Frequency Cepstral Coefficients) as feature extractor and the blend of GMM and back-propagation ANN that scored 95.7% accuracy. In Negesse (2015) Feda employed Levenshtein Algorithm to compute lexical distance and agglomerative clustering method to categorize Afaan Oromo linguistic varieties into six dialect clusters. Tigrinya Dialects - Previous works on Tigrinya dialects have categorized the task mainly using geography of speakers and characteristics of the dialects in Tigrinya. In [2], Niguss listed three (L, D and Z) Tigrinya dialects whereas only two dialects (Tigray and Eritrea) are mentioned as common dialects in Feleke (2017). In Mengesha (2009), Tsehaye, who compared Wajerat Tigrinya with mainstream Tigrinya, states that Tigrinya has multiple varieties that broadly differ from one another, not only from one zone to the other but within one zone as well.

3 Dataset

There are no available datasets to study Tigrinya dialects. In this section we describe how we generated a dataset for the purpose of identifying Tigrinya dialects from text. Our approach is as follows: (i) We took sample sources, e.g., books, from each dialect when available then (ii) For a sample text from each source, we translated them to the other dialects with natives in the target dialect and (iii) Finally, we performed quality checks.

Concretely our implementaiton is as follows: For the Z variety, we found a book (Kilte Zantatat) by Gidey Gebrekidan (Gebrekidan, 2021) which is written in the Z variety. The book was subdivided into 100 parts and the pages were reshuffled. Out of the reshuffled text, we randomly took 10 of them and distributed them among selected participants that volunteered to translate them into L and D varieties. Out of the ten distributed text files, seven of them were successfully translated into L and D varieties. For the L variety, we were able to acquire book chapters from two books written in that variety, (Fato) by Tesfay Gebregizeabher (Gebre-Egziabher, 2017) and ErquiEnderta by Meles Solomon (Solomon, 2020), both of which were translated to Z and D varieties. For the D variant, we could not find a book. Instead, we collected data form two Facebook users, Akeza Awalom Awealom and Guraya Asadi Raya Raya, that consistently write in that variety and the collected posts were translated to Z and L varieties in a similar fashion.

This way we gathered a total of 2,960 source sentences in each dialect that were then translated to the other two dialects (8,880). The source sentence breakdwon per source is: L (1041) D (764) and and Z (1155).

4 Experiments and Results

We cast the problem of identifying dialects as a multi-class classification task. As baselines, we used three classical text classification approaches namely, Naïve Bayes, Linear SVM and ANN. We used character n-grams as features. In addition to the experiments conducted on the classical text classifiers, we also conducted similar experiments with deep-learning approaches, namely: CNN, BiLSTM and combination of the two models (CNN-BiLSTM) using character sequences. We used NLTK[1] along with Scikit-Learn[2] libraries for the classical machine learning algorithms. To train the deep learning models, we used Keras[3] with Tensorflow API as a backend. Table 1 shows the result of our experiments. Using ten-fold cross-validation, we showed that the CNN based model achieved 92.98% overall accuracy.

Table 1: Weighted average Precision, Recall and F1-measure as well as overall Accuracy of dialect identification for Tigrinya

	Precision	Recall	F1-measure	Accuracy
NB	0.85	0.83	0.82	0.83
SVM	0.87	0.87	0.87	0.87
ANN	0.89	0.89	0.89	0.89
BiLSTM	0.89	0.90	0.89	0.90
CNN-BiLSTM	0.92	0.92	0.92	0.92
CNN	0.93	0.93	0.93	0.93

In this work, we have seen dialect identification for Tigrinya. This has practical uses for real-world applications. We also believe it's an important research to document and preserve a language. Future work should look into expanding the dataset. We hope this will also contribute to the standardization efforts of the language that should include from all dialects.

[1] https://www.nltk.org/
[2] https://scikit-learn.org/stable/
[3] https://keras.io/

References

REDIAT BEKELE ASFAW. A comparative study of automatic language identification on ethio-semitic languages. 2018.

Akeza Awealom. Facebook. URL https://www.facebook.com/akeza.awealom.

Tekabe Legesse Feleke. The similarity and mutual intelligibility between amharic and tigrigna varieties. In Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), pp. 47–54, 2017.

Tesfay Gebre-Egziabher. Fato (First Edition). Planography, Mekelle, Tigray, 2017.

Teklay Gebregzabiher. Part of speech tagger for tigrigna language. Master's Thesis, Addis Ababa University, Addis Ababa, unpublished, 2010.

Gidey Gebrekidan. Kilte Zantatat. snippet taken from draft, Mekelle, Tigray, 2021.

Tsehaye Kiros Mengesha. A comparison of wajerat tigrigna vs. standard tigrigna, 2009.

Abrham Debasu Mengistu and Dagnachew Melesew Alemayehu. Speech processing for text independent amharic language dialect recognition. Indonesian Journal of Electrical Engineering and Computer Science, 5(1):115–122, 2017.

Feda Negesse. Classification of oromo dialects: A computational approach. International Journal of Computational Linguistics (IJCL), 6(1):1–10, 2015.

Guraya Asadi Raya. Facebook. URL https://www.facebook.com/guraya.asadiraya.

Meles Solomon. Erqi Enderta. snippet taken from draft, Mekelle, Tigray, 2020.