TIGRINYA DIALECT IDENTIFICATION

Asfaw Gedamu Haileslasie
Lesan AI
asfawg@gmail.com

Asmelash Teka Hadgu
Lesan AI
asme@lesan.ai

Solomon Teferra Abate
Addis Ababa University
solomon.teferra@aau.edu.et

ABSTRACT

Dialect Identification is an important topic of research in Natural Language Processing (NLP) as it has broad implications in many real-world applications such as machine translation, speech recognition and chatbots to name a few. In this work, we investigate 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 datasets for each dialect. Finally, we perform experiments using classical machine learning and deep learning methods to quantify effectiveness of current methods on the problem of Tigrinya dialect identification. The highest overall accuracy of 92.98% was achieved using character-level Convolutional Neural Networks (CNNs).

1 INTRODUCTION

Tigrinya is the official regional language of Tigray, in 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, we train models that determine which Tigrinya dialect the text is 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

Tigrinya Dialects: Previous works have categorized Tigrinya dialects, mainly using geography of speakers and characteristics of the different dialects. Weldezgu Mehari (2021) listed three Tigrinya dialects namely northern, central and southern dialects whereas only two dialects (Tigray and Eritrea) are mentioned as common dialects in Feleke (2017). Mengesha (2009) compared Wajerat Tigrinya with mainstream Tigrinya, and stated 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. We have synthesized three dialects: Z, L, D based on these previous works that had concrete data (examples and dialect characteristics) to support each dialect.
Dialect Identification: Mengistu & Alemayehu (2017) 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. Negesse (2015) employed Levenshtein Algorithm to compute lexical distance and agglomerative clustering method to categorize Afaan Oromo linguistic varieties into six dialect clusters.
Language Identification: ASFAW (2018) conducted a comparative study of automatic language identification on Ethio-Semitic languages – Amharic, Ge'ez, Gurage and Tigrinya using character n-grams. The author compared Naïve Bayes with Cumulative Frequency Addition and concluded that the latter performed better. Feleke (2017) 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.

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 native speakers in the target dialect and (iii) Finally, we performed quality checks.

Concretely our implementation is as follows: For the Z variety, we used snippets from the book Kilte Zantatat (Gebrekidan, 2021). For the L variety, we used book chapters from Fato (Gebre-Egziabher, 2017) and Erqi Enderta (Solomon, 2020). These were then 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, 2021) and Guraya Asadi Raya (Raya, 2021), that consistently write in that variety. The collected posts were translated to Z and L varieties in a similar fashion. Our approach makes it possible to study the dialect identification on a corpus with similar content across dialects, avoiding effects such as author or category of text.

This way we gathered a total of 2,964 source sentences from all three dialects. The source sentence breakdown per dialect is: Z (n=1125), L (n=1075) and D (n=764). These were then translated to the other two dialects resulting in a total of 8,892 sentences. 14 speakers in total that are native in the target dialect participated in the translation task. We make the dataset publicly available[1] for other researchers to build up on.

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 running from one to five(n=1- 5) 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[2] along with Scikit-Learn[3] libraries for the classical machine learning algorithms. To train the deep learning models, we used Keras[4] 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
BiLSTM	0.88	0.88	0.88	0.88
ANN	0.89	0.89	0.89	0.89
CNN-BiLSTM	0.91	0.91	0.91	0.91
CNN	**0.92**	**0.92**	**0.92**	**0.92**

In this work, we have demonstrated the application of machine learning techniques for dialect identification in Tigrinya text. Our findings contribute to the ongoing efforts to standardize the language by incorporating all dialects. Furthermore, we believe that this research plays a crucial role in documenting and preserving the language. In future work, we plan to expand the dataset and explore additional modalities, such as audio. We will also study the effect of dialect identification on a real-world machine translation application for Tigrinya to and from Amharic and English (Hadgu et al., 2021).

[1] https://zenodo.org/record/7832329#.ZDtFk9JBx3Q

[2] https://www.nltk.org/

[3] https://scikit-learn.org/stable/

[4] https://keras.io/

REFERENCES

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

Akeza Awealom. Akeza awealom [facebook page], 2021. URL https://www.facebook.com/akeza.awealom. Accessed: 2021-03-06.

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.

Asmelash Teka Hadgu, Abel Aregawi, and Adam Beaudoin. Lesan–machine translation for low resource languages. In *NeurIPS 2021 Competitions and Demonstrations Track*, pp. 297–301. PMLR, 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. Guraya asadi raya [facebook page], 2021. URL https://www.facebook.com/guraya.asadiraya. Accessed: 2021-03-02.

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

Niguss Weldezgu Mehari. A grammar of rayya tigrinya, 2021. URL http://etd.aau.edu.et/bitstream/handle/123456789/24899/Niguss%20W.%20PhD%20Dissertation%20Final%20Printed%20%40%2027.pdf?sequence=1&isAllowed=y.

A APPENDIX

Source	Z	L	D
Z	ዝደልዮ ንባባት ምቕራብ እውን ኣቓሪፃ ኣላ።	ልደልዮ ንባባት ምቕራብ ለ ኣቓሪፃ እኒሀ።	ድደልዮ ንባባት ምምፃእ ለ ሓድጋቶ እንሃ።
	ኣይመስለንን፤ እንተወሓደ ንዓርሰይ ድሕሪ ሓሙሽተ ዓመት ኣበይ ክረኽቦ ከም ዝደሊ ሕዚ ክረኣየኒ ክኢሉ ኣሎ።	ኣይመስለንይ፤ ተወሓደ ልባዕለይ ድሕሪ ሓሙሽተ ዓመት ኣበይ ከግንዮ ሃም ልደሊ ኸዚ ክረኣየኒ ኽኢሉ።	የመስለንይ፤ ተንኣሰ ድባዕለይ ካሓሙሽተ ዓመት ድደሓር ኣበይ ካግንዮ ሃም ድደሊ ሐዚ ክርእየኒ ጀምሩ እንሆ።
	ንምንታይ ከም ዝኾነ ኣይፈልጥን።	ልምንታይ ሃም ልኾነ ኣይፈልጥይ።	ድምንታይ ሃም ድኾነ ኣይፈልጥይ።
L	ብዕርቂ እንደርታ ዝፍትሑ ቀንዲ መበገሲ ጕንፅታት÷	ብዕርቂ እንደርታ ልፍትሑ ቀንዲ መልዐሊ ጕንፅታት÷	ብዕርቂ እንደርታ ድፍትሑ ቀንዲ መልዐሊ ጕንፅታት÷
	ናብ ታክሲ ውሽጢ ምስ ኣተወ ብሞስኮት ክሳዱ ውጥጥ ኢሉ “ራዛ?	ላብ ታክሲ ውሽጢ ብልኣተወ ብሞሽኮት ኽሳዱ ውጥጥ ኢሉ “ራዛ?	ዳብ ታክሲ ውሽጢ ብድጉበእ ብሞሽኾት ኽሳዱ ውጥጥ ውሉ “ራዛ?
	ንሳቶም እውን ከም'ዚ ዝስዕብ ተገሊፆም ኣለው።	እሳቶም ለ ሃም'ዚ ልስዕብ ተገሊፆም እኒሀዉ።	እሳቶም ለ ሃ''ዚ ድስዕብ ትገልፆም እንሀዩ።
D	ፍረ ለካቲት ማለት ከምዚ ዝበለ ተግባር እዩ።	ፍረ ለካቲት ማለት ሃምዚ ልበለ ተግባር እዩ።	ፍረ ለካቲት ማለት ሃምዚ ዳበለ ተግባር ዩኡ።
	ምስ በላዕኻ ተመራረቅካ ናብ ዳስ ትምለስ።	ብል በላዕኻ ተመራረቕኻ ላብዳስ ትምለስ።	ብድ በላዕኻ ተምራረቕዋ ዳዳስ ትምለስ።
	ንኩሉ ግን ዕድመን ጥዕናን ይሃበና።	ልኹሉ ምሉይ ዕድመይ ጥዕናይ ይሃበና።	ድኹሉይ ምሉይ ዕድመይ ጥዕኛይ ይሃበና።

Figure 1: A random sample of text examples for different Tigrinya dialects.