MENYO-20k: A Multi-domain English-Yorùbá Corpus for Machine Translation

David I. Adelani[12], **Jesujoba O. Alabi**[3], **Damilola Adebonojo**[4], **Adesina Ayeni**[5], **Mofe Adeyemi**[4], **Ayodele Awokoya**[4]

[1]Spoken Language System Group (LSV)
[2]Saarland Informatics Campus, Saarland University, Germany.
[3]Max Planck Institute for Informatics, Germany.
[4]Masakhane NLP
[5]Yobamoodua Cultural Heritage (YMCH).

1 Introduction

Machine translation is a popular Natural Language Processing (NLP) task which involves the automatic translation of sentences from a source language to a target language. In practice, training machine translation models involves collecting huge parallel sentences like a hundred thousand sentences to achieve a moderate translation performance. There are several parallel sentences online for high-resource languages like some European languages. However, for low-resource languages like Yorùbá, there are a few parallel sentences online. The available corpora for training low-resource machine translation systems are the Bible – the Bible is the most available resource for low-resourced languages (Resnik et al., 1999) and JW300 (Agić and Vulić, 2019), but they are both in the religious domain. They do not generalize very well to the domains of interest to most users like news domain and daily conversations (∀ et al., 2020). In this paper, we address this problem by creating a multi-domain parallel dataset to properly evaluate the generalization of machine translation models trained on JW300 and the bible on new domains while exploring some transfer learning approaches that can make use of few thousand sentences for domain adaptation. The dataset we are creating comprise of texts obtained from news articles, ted talks, movie transcripts, radio transcripts, science and technology texts, and other short articles curated from the web and professional translators.

2 The Yorùbá Language

The Yorùbá language is the third most spoken language in Africa, and is native to the south-western Nigeria and the Republic of Benin. It is one of the national languages in Nigeria, Benin and Togo, and it is also spoken in other countries like Ghana, Côte d'Ivoire, Sierra Leone, Cuba, Brazil and by a significant Yorùbá diaspora population in the US and United Kingdom mostly from the Nigerian ancestry. The language belongs to the Niger-Congo family, and is spoken by over 40 million native speakers (Eberhard et al., 2019).

Yorùbá has several dialects but the written language has been standardized by the 1974 Joint Consultative Committee on Education (Asahiah et al., 2017), it has 25 letters without the Latin characters (c, q, v, x and z) and with additional characters (ẹ, gb, ṣ , ọ). There are 18 consonants (b, d, f, g, gb, j[dz], k, l, m, n, p[kp], r, s, ṣ, t, w y[j]), 7 oral vowels (a, e, ẹ, i, o, ọ, u), five nasal vowels, (an, ẹn, in, ọn, un) and syllabic nasals (m̀, ḿ, ǹ, ń). Yorùbá is a tonal language with three tones: low, middle and high. These tones are represented by the grave ("\"), optional macron ("−") and acute ("/") accents respectively. These tones are applied on vowels and syllabic nasals, but the mid tone is usually ignored in writings. The tones are represented in written texts along with a modified Latin alphabet. A few alphabets have underdots (i.e. "ẹ", "ọ", and "ṣ"), we refer to the tonal marks and underdots as diacritics. It is important to note that tone information is needed for correct pronunciation and to have the meaning of a word (Asahiah et al., 2017; Adegbola and Odilinye, 2012).

As noted in (Asahiah, 2014), most of the Yorùbá texts found in websites or public domain repositories either use the correct Yorùbá orthography or replace diacriticized characters with un-diacriticized ones. Often time, articles written online including news articles[1] like BBC and VON ignore diacritics. Ignoring

[1] `https://www.von.gov.ng/yoruba/`, and `https://www.bbc.com/yoruba`

diacritics makes it difficult to identify or pronounce words except they are in a context. For example, *owó* (money), *ọ*wọ̀ (broom), *ò*wò (business), *ọ̀*wọ̀ (honour), *ọ*wọ́ (hand), and *ọ̀*wọ́ (group) will be mapped to *owo* without diacritics.

3 Dataset Collection

4 Conclusion

Acknowledgements

We would like to thank Adebayo O. Adeojo, Babunde O. Popoola, Olumide Awokoya, Modupe Olaniyi, Princess Folasade, Tolulope Adelani, and Oluyemisi Olaose for their support in translating English sentences to Yorùbá and the verification of Yorùbá diacritics. This project was funded by the AI4D language dataset fellowship https://www.k4all.org/project/language-dataset-fellowship/

References

Tunde Adegbola and Lydia U. Odilinye. 2012. Quantifying the effect of corpus size on the quality of automatic diacritization of yoruba texts. In *Spoken Language Technologies for Under-Resourced Languages*.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July. Association for Computational Linguistics.

F. O. Asahiah, O. A. Odejobi, and E. R. Adagunodo. 2017. Restoring tone-marks in standard yoruba electronic text: Improved model. *Computer Science*, 18(3):301–315.

Franklin O Asahiah. 2014. Development of a standard yoruba digital text automatic diacritic restoration system. *Phd. Thesis, Obafemi Awolowo University, Ile-Ife, Nigeria*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the world. twenty-second edition.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, ..., Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, .., and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November. Association for Computational Linguistics.

P. Resnik, M. Olsen, and Mona T. Diab. 1999. The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33:129–153.