Towards a parallel corpus of Portuguese and the Bantu language Emakhuwa of Mozambique

Felermo D. M. A. Ali¹, Andrew Caines², Jaimito L. A. Malavi¹
¹Department of Computer Engineering, Lurio University, Mozambique
²Computer Laboratory, University of Cambridge, United Kingdom
{felermo.ali,jmalave}@unilurio.ac.mz,
andrew.caines@cl.cam.ac.uk

Abstract

Major advancement in the performance of machine translation models has been made possible in part thanks to the availability of large-scale parallel corpora. But for most languages in the world, the existence of such corpora is rare. Emakhuwa, a language spoken in Mozambique, is like most African languages low-resource in NLP terms. It lacks both computational and linguistic resources and, to the best of our knowledge, few parallel corpora including Emakhuwa already exist. In this paper we describe the creation of the Emakhuwa-Portuguese parallel corpus, which is a collection of texts from the Jehovah’s Witness website and a variety of other sources including the Universal Declaration of Human Rights and Mozambican legal documents. The dataset contains 47,415 sentence pairs, amounting to 699,976 word tokens of Emakhuwa and 877,595 word tokens in Portuguese. After normalization processes which remain to be completed, the corpus will be made freely available for research use.

1 Introduction

Machine translation (MT) is the process by which computational models are trained to transform a source language text into a target language text, and is a technology which in recent years has seen great improvements in performance thanks to the development of MT with neural networks (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015). NMT typically depends on supervised machine learning from large parallel corpora of texts in the source and target language pair. The creation of such corpora usually depends on extrinsic motivation for the existence of abundant parallel text: for instance, government activity in multilingual settings such as the European and Canadian Parliaments. In the context of NMT, this fact has led to a dichotomous situation of high and low-resource language pairings. For the academic community, English-German or English-French are prototypical high-resource translation pairs whereas most other language pairs are low-resource in comparison.

We present a corpus for a low-resource language pair: Portuguese and Emakhuwa (alternatively, ‘Makhuwa’ or ‘Makua’), the official and the most widely spoken languages of Mozambique, respectively. Emakhuwa is spoken in all the provinces of northern Mozambique, namely Niassa, Cabo Delgado and Nampula and also in Zambezia in central northern Mozambique. It is estimated that approximately 25% of the country’s population of 30 million people make use of the language on a daily basis as an alternative to Portuguese (de Paula and Duarte, 2016). Emakhuwa is also spoken in some of the neighbouring countries to the north of Mozambique – namely, Tanzania and Malawi (Ngunga, 2012) but speaker populations in these countries are relatively small.

There are 8 variants of Emakhuwa (Ngunga and Faquir, 2012): Elomwe (ISO-639 code: ngl), Esankaci, Esakaka (ISO-639 code: xsq), Echirima (ISO-639 code: mhm), Emarevoni (ISO-639 code: xmc), Emeeto (ISO-639 code: mgh), Enahara and Central (ISO-639 code: vmw). The variants are distributed across different districts of northern and central Mozambique (see Figure 1) and differ slightly in terms of accents and lexicon.

Like many languages spoken on the African continent, Emakhuwa has limited resources for computational linguistics compared with English, French, Portuguese, and so on. Thus we are collecting a dataset for MT in which Portuguese texts are paired with Emakhuwa translations. The 47,415 sentence pairs we have collected contain 699,976 word tokens of Emakhuwa and 877,595 word tokens in Portuguese. The corpus will be made available for...
research use when final normalization and data collection has been completed\(^1\). In addition we will seek out new data sources and continue to expand the corpus.

2 Related Work

In general, African languages have been relatively little explored by the NLP research community. However, lately this scenario seems to be improving as more data are being made available for research use, including translation work. OPUS (Tiedemann, 2012), the 1000Langs corpus crawled from Bible corpora (Asgari and Schütze, 2017), and the JW300 project (Agić and Vulić, 2019) are examples of this trend, as they made available parallel corpora of over 300 languages including Emakhuwa-Portuguese. The Crúbadán Project\(^2\) also used Jehovah’s Witness resources to create a corpus of Emakhuwa. However, the corpus is monolingual and relatively small with around 44,071 tokens extracted from 4 documents.

The Masakhane project works mainly on low-resource MT for African languages, and is an excellent source of African language parallel corpora (Nekoto et al., 2020). Currently, they have a growing online community of researchers who collaborate, contribute and share advancements in NLP for African languages. They have an open repository\(^3\) for sharing data as well as code and tools for building and facilitating NLP research in African languages. Nevertheless, the Emakhuwa language has not been explored yet by the community.

3 Data

The dataset contains Portuguese-Emakhuwa parallel sentences, collected from crawling the Jehovah’s Witness (JW) website\(^4\) and the African Story Book website\(^5\). The African Story Book contains 17 short stories in Emakhuwa with Portuguese translations. The Jehovah’s Witness website contains The Watchtower—Study Edition magazine as well as The Watchtower and Awaque! from 2016 to 2021. Also, contains 27 books of the New Testament Bible.

The corpus also contains text from various institutional sources. These were digitized by performing optical character recognition (OCR) on a set of PDF files from Centro Catequético Paulo VI (Paul VI Catechetical Centre), the Universal Declaration of Human Rights, and Mozambican Land Law. The majority of texts come from the Centro Catequético Paulo VI which has published a series

\(^{1}\text{placeholder.com}\)
\(^{2}\text{http://crubadan.org/languages/vmw}\)
\(^{3}\text{https://github.com/masakhane-io/masakhane-mt}\)
\(^{4}\text{https://www.jw.org}\)
\(^{5}\text{https://www.africanstorybook.org}\)
After that, I realized that it was the right choice.

Do I live on what I believe?

A man had two sons.

The law needs to be known by the whole community, because from the knowledge of the law we can defend and safeguard our rights.

in rural areas, local communities participate:

Table 1: Examples from the Portuguese-Emakhuwa parallel corpus, with English translation.

Table 2: Dataset counts where: JW = Jehovah’s Witness, CCA = Centro Catequético Paulo VI, ASB = African Book Story, LL = Land Law of Mozambique, HR = Universal Declaration of Human Rights; PT = Portuguese, VMW = Emakhuwa.

| Source                | Lang | Sentences | Tokens   |
|-----------------------|------|-----------|----------|
| JW                    | PT   | 42,840    | 798,371  |
|                       | VMW  |           | 638,365  |
| CCA                   | PT   | 4067      | 73,221   |
|                       | VMW  |           | 56,724   |
| ASB                   | PT   | 294       | 2746     |
|                       | VMW  |           | 1945     |
| LL                    | PT   | 128       | 1656     |
|                       | VMW  |           | 1832     |
| HR                    | PT   | 86        | 1601     |
|                       | VMW  |           | 1110     |
| Total                 | PT   | 47,415    | 877,595  |
|                       | VMW  |           | 699,976  |
account, since Emakhuwa takes many words from Portuguese, but without a consensus on spelling conventions. As an example, take the word “bible” which in Portuguese is “biblia”. In Emakhuwa it is sometimes written as “biblia” and at other times written as “bibiliya”. This inconsistency also exists within Emakhuwa language resources itself, as the alphabet and spelling standards went through several revisions – the latest in 2012 (Ngunga and Faquir, 2012). Therefore, some future processing is necessary to normalize the word standards in the dataset. The corpus will be available for research use, and updates may be obtained from our project website\(^6\). Table 2 summarizes the corpus in terms of sentences and the number of tokens coming from each data source.

4 Baseline translation model

We train an initial NMT model with the default OpenNMT configuration\(^7\), which consists of a 2-layer LSTM with 500 hidden units on both the encoder and decoder (Klein et al., 2017). Sentences from the corpus are randomly assigned to training and test sets at a ratio of 9:1. These data splits will be released with the corpus so that others may compare the performance of their models against this one. We evaluate with BLEU (Papineni et al., 2002) as shown in Table 3. This presents a baseline level of performance to improve upon in future work.

5 Conclusion

In this paper we describe the creation of a new parallel corpus of Emakhuwa and Portuguese. The preparation of the corpus is on-going as the texts require more processing and normalization, but it will be made freely available for research use, most probably for machine translation. Currently, the dataset is made up of mostly religious and legal content but in future our objective is to diversify the range of sources and topics covered. This is important in order that Portuguese-Emakhuwa translation models work well across various domains.

Acknowledgements

The authors wish to acknowledge the Centro Catequético Paulo VI for allowing to use their materials for research purpose. The second author is supported by Research England via the University of Cambridge Global Challenges Research Fund. He wishes to thank Paula Buttery, Tanya Hall, Carol Nightingale & Sara Serradas Duarte for their help and guidance.

References

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. Association for Computational Linguistics.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations (ICLR).

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations.

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, Ricky Macharm, Perez Ogayo, Oreevaoghenie Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Dégila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Gholah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitena...
Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, 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.

Armindo Ngunga. 2012. Interferências de línguas moçambicanas em português falado em moçambique. Revista Científica Da Universidade Eduardo Mondlane, Série: Letras e Ciências Sociais, 1:7–20.

Armindo Ngunga and Osvaldo Faquir. 2012. Padronização da Ortografia de Línguas Moçambicanas: Relatório do III Seminário. Coleção As Nossas Línguas III. Maputo: CEA.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Ronaldo Rodrigues de Paula and Fábio Bonfim Duarte. 2016. Diversidade linguística em moçambique. In Kadila: culturas e ambientes – Diálogos Brasil-Angola. São Paulo: Bluche.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12).