Using Neural Machine Translation to Create African Languages Corpora

Kathleen Siminyu, Africa's Talking.

Context

Not only is Africa the second most populous continent in the world with over one billion people, it is also home to the highest linguistic diversity in the world, with over 1500 different languages. Due to the scramble and partition of Africa - which was the occupation, division and colonisation of African territories by European powers - many African countries, after gaining independence, selected one language, in search of unity, to be the official language. Generally this language was the former colonial language. The dominant ones among these are English, French, Portuguese and Spanish.

There are more than a dozen African countries where English, in particular, is an official language. This means that it is the dominant language of business, education and government. These include Zimbabwe, Uganda, Zambia, Botswana, Namibia, Kenya, Sierra Leone, Liberia, South Africa and Nigeria. But even among these nations, Nigerian English is as unique as Ghanaian English and Cameroonian English. There's an accent, slang and a multitude of local words and phrases thrown into the mix in different contexts making the language unique in each instance.

Rationale

Having English as the official language for education means it is the medium via which knowledge is imparted as well as evaluated. In the African context, depending on where a school is situated, whether in urban or rural areas, learners can broadly be classified into 3 different categories; those for whom English is a native language, those for whom English is a second language and those with very limited English proficiency.

At the end of the day, where English is the official language for education, regardless of one's mother-tongue or English language proficiency, all learners find themselves in English-medium schools where their understanding and proficiency in the language directly affects their ability to learn and perform.[1]

In addition, research done on the importance of language proficiency in the labour market has highlighted the difficulties faced by immigrants whose mother tongue differs from the main language of the domestic population. In countries that experience a large influx of immigrants, there are often severe labour market inequalities between natives and migrant populations in terms of employment prospects and earnings. These labour market discrepancies are partly attributed to differences in language ability. Specifically, the relationship between language ability and employment has been explored: good language skills can assist in job search activities and may signal an individual's productive capacity to prospective employers thereby increasing the likelihood of employment.[2]

Project Goals and Objectives

There is a need for people all over the world to be able to use their own language to learn and especially when using computers or accessing information on the Internet. This requires the existence of a variety of applications including local language spell-checkers, word processors, machine translation systems, search engines, etc. At the same time, the amount of work required to develop all aspects of natural language processing for a new language is huge.

This project proposes working on a variety of language processing tools for African languages, starting with the building of machine translation models that will use statistical translation to create bilingual (or multilingual) parallel corpora for African languages.

This project proposes working on a variety of language processing tools for African languages, starting with creating bilingual (or multilingual) parallel corpora. Thankfully, whereas in the past this would have to be achieved via the manual collection and annotation of enough text in each African language that we choose to include, machine translation provides us with a timely and cost effective alternative.

Machine translation is the automatic translations of text from one language to another. It has recently achieved impressive performance thanks to recent advances in deep learning and the availability of large-scale parallel corpora. As in this case parallel corpora data is scarce, we shall instead use techniques that have been proposed for building Machine Translation models for low resource languages.

Lample et al, 2018, propose a model that takes sentences from monolingual corpora in two different languages and then maps them into the same latent space. By learning to reconstruct in both languages from this shared feature space, the model effectively learns to translate without using any labeled data.[3]

Zhang and Zong, 2016, propose a model that transforms bilingual dictionaries into adequate sentence pairs so that a Neural Machine Translation model can distil latent bilingual mappings from the repetitive phenomena.[4]

Gu et al, 2018, propose an approach that utilizes transfer-learning to share lexical and sentence level representations across multiple source languages into one target language.[5]

Sennrich et al, 2015, propose pairing monolingual training data with an automatic back-translation, which can be treated as additional training data.[6]

Expected Results

The project will entail creating multilingual corpora and then lexical corpora for major African languages, beginning with Chichewa and Kiswahili. This will then open up opportunity to work on a variety of other NLP tools such as for preprocessing (noise removal, lexical normalization) and feature engineering (syntactic parsing, entity extraction, word embedding) for each respective language.

References

1) Vic Webb, “English as a Second Language in South Africa's Tertiary Institutions: A Case Study at the University of Pretoria”, 2002.
2) Daniela Casale, Dorrit Posel, “English language proficiency and earnings in a developing country: The case of South Africa” in Journal of Socio-Economics, 2011.
3) Guillaume Lample, Alexis Conneau, Ludovic Denoyer, Marc'Aurelio Ranzato, “Unsupervised Machine Translation Using Monolingual Corpora Only” at International Conference on Learning Representations, 2018.
4) Jiajun Zhang, Chengqing Zong, “Bridging Neural Machine Translation and Bilingual Dictionaries”, 2016.
5) Jiatao Gu, Hany Hassan, Jacob Devlin, Victor O.K. Li, “Universal Neural Machine Translation for Extremely Low Resource Languages”, 2018.
6) Rico Sennrich, Barry Haddow, Alexandra Birch, “Improving Neural Machine Translation Models with Monolingual Data”, 2015.