Using ML to close the vocabulary gap in the context of environment and climate change in Chichewa

Amelia Taylor
University of Malawi
tNyasa Data Labs

ICLR 2020 Workshop on Tackling Climate Change using ML

Chichewa

- Chichewa (also known as Nyanja) is one of the main Bantu languages
- It is spoken in Malawi, Zambia, Zimbabwe and Mozambique
- In Malawi it is the national language alongside English. Other notable languages are Chiyao and Chitumbuka

- Between Chichewa and Nyanja there are some dialect differences although speakers of one language understand well the other
- Most of the automatic translation tools are based on Nyanja texts.

Chichewa in Education

- During the time Malawi was a British protectorate (until 1964) the education in schools was done in English

- After the independence, Chichewa was introduced in primary schooling and the stringent entry requirements into Civil Service (i.e., a very good command of English) were relaxed.

- Secondary school and University education are still using English as the main classroom language

Problem Statement - 1

- There has been a deterioration in the proficiency of English among the young generation

  (Kapesi, 2011): tensions and pressures especially in the teaching of abstract and science subjects

  (Chiuye and Moyo, 2008): worrying 'vocabulary gap' that exists between Chichewa and other official indigenous languages of Malawi

  (Kamwendo, 2016): availability of materials in Chichewa creates 'linguistically impoverished and deprived' learners

Climate Change and Environment - 2

- At the same time there has been an increase in the availability of information.
- And an even increasing need for the dissemination of knowledge and awareness about climate change that is available to school children and their educators.

- (Kambewa and Mataya, 2007): the scarcity of accurate and good quality information on the state of things in Malawi contributes to the maintaining of the status quo and a continuing degradation of the environment

The Need for Developing Terminologies

- (Ikuenobe, 2014) and (Cloete, 2011): large portion of the indigenous vocabulary and knowledge remains unknown or is slowly disappearing.

- From a list of more than 20 categorisations of landscape types in the language Xitsonga, compiled by Wolmer very few were now recognised by the younger generation

- (Kamwendo, 2016): there is a need to '*develop terminologies and a broad lexical base*'. This will help in '*diffusing and refuting stereotype notions that indigenous African languages lack a conceptual framework to express scientific notions with appropriate scientific vocabulary*' (Chiuye, Moyo 2008).

Aim

- We propose to build a corpus of parallel English-Chichewa texts containing terms for the environmental science in Chichewa using resources available on the Internet.

I) Start from a seed corpus containing manual translations of terms and phrases that are grammatically correct and constitute 'good examples'

II) Use this seed corpus to gather usages from the Internet involving this terminology:
   - Search using Chichewa words/ phrases
   - Establish the relevance of the search results (e.g., if the results are show new usages of the terms, and detect if the translations were machine generated)
   - Establish new usage in context of terms and add these to the corpus
   - Establish usage of the same terms but in contexts unrelated to environmental/climate issues.

III. Use the corpus to study the understanding and attitudes to environment as expressed online in social media and newspaper articles.

Search Using Chichewa Words/ Phrases

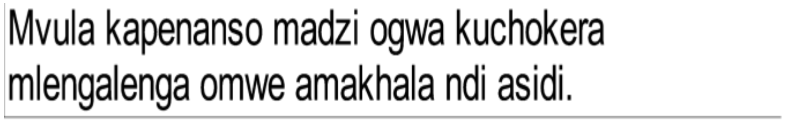
- Arable land = ‘Malo olima’ or ‘minda’

Google Translation: malo olima

- Manure = ‘manyowa’ or ‘zinyalala zowolerana’

Google Translation: manyowa (loan word from English)

- Acid rain:



Mvula kapenanso madzi ogwa kuchokera mlengalenga omwe amakhala ndi asidi.

Google Translation: *Mvula ya asidi*

(A literal translation but not necessarily meaningful.)

Search Using Chichewa Words/ Phrases

- Adaptation to environment: “kuyanjana ndi nyango” which literally means “reconciliation with the climate”

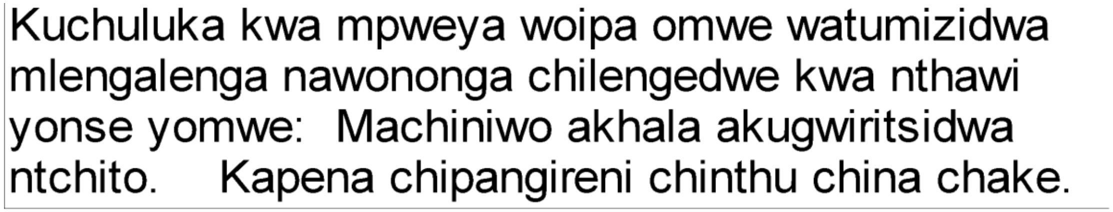
Google Translation: kutengera chilengedwe

- Alternative fuel = *Njira zina zamakono*

Google Translation: mafuta ena

(ena means another, whereas ‘njira zina zamakono’ literally means to find alternative ways)

- Carbon footprint:



Kuchuluka kwa mpweya woipa omwe watumizidwa mlengalenga nawononga chilengedwe kwa nthawi yonse yomwe: Machiniwo akhala akugwiritsidwa ntchito. Kapena chipangireni chinthu china chake.

Google translate: mawonekedwe a kaboni

What data is available online?

- The type of results that one finds when searching with Chichewa phrases or words are:
  - Questionnaires used by NGO's to gather data in Malawi/ Zambia
  - Manuals published by environmental organisations and translated in Chichewa
  - Newspapers articles
  - Articles written by companies / agencies such as Mining companies
  - Christian literature
  - Social media (mixed Chichewa and English) – to a small extent
  - Machine translations of Articles which were first written in English

Questions

- Can we detect the results that were machine generated?
- Do these favour specific translations?
- Are these different than those used by writers which are native Chichewa speakers?
- Is the young generation interested in issued to do with the environment and what language will they use in discussing these issues?

How can ML techniques be used?

- Aligning the English and Chichewa translations
- Establishing similarities between texts
- Authorship detection: in our scenario we have two translators, a machine translator like the GT and a human translator.
- Determining synonyms (or different ways to express the same concept) based on the answers to our queries.

Thank you!

- We are still in the concept phase.
- We welcome your suggestions and or ideas for collaborations or insights.

- Contact: ataylor@poly.ac.mw