USING ML TO CLOSE THE VOCABULARY GAP IN THE CONTEXT OF ENVIRONMENT AND CLIMATE CHANGE IN CHICHEWA

Amelia Taylor
Department of Computer Science
The Polytechnic, University of Malawi
Malawi
{ataylor}@poly.ac.mw

ABSTRACT

In the west, alienation from nature and deteriorating opportunities to experience it, have led educators to incorporate programs in schools, to bring pupils in contact with nature and to enhance their understanding of issues related to the environment and its protection. In Africa, and in Malawi, where most people engage in agriculture, and spend most of their time in the 'outdoors', alienation from nature is happening too, although in different ways. Large portion of the indigenous vocabulary and knowledge remains unknown or is slowly disappearing. There is a need to build a glossary of terms regarding environment and climate change in the vernacular to improve the dialog regarding climate change and environmental protection. We believe that ML has a role to play in closing the 'vocabulary gap' of terms and concepts regarding the environment and climate change that exists in Chichewa and other Malawian languages by helping to creating a visual dictionary of key terms used to describe the environment and explain the issues involved in climate change and their meaning. Chichewa is a descriptive language, one English term may be translated using several words. Thus, the task is not to detect just literal translations, but also translations by means of 'descriptions' and illustrations and thus extract correspondence between terms and definitions and to measure how appropriate a term is to convey the meaning intended. As part of this project, the identification of 'loanword patterns' into Chichewa from other languages such as English, may be useful in understanding the transmission of cultural items.

1 SOME RELEVANT LANGUAGE ISSUES IN MALAWI

Chichewa, with its geographical dialects , is the language of about half of Malawians, spoken both in rural and urban places. Other notable languages are Chiyao and Tumbuka, Chilomwe and Chisena. From 1907 to 1964, Malawi was a British protectorate and the education in the country was in the hands of the Christian missionaries with small governmental input. Missionaries taught English, but at the same time they engaged in learning the local language, conducting significant linguistic work (e.g., writing the first dictionaries and grammatical rules) and engaging in the translation of the Bible in the main local languages of Chichewa, Yao and Tumbuka. Over the years and more significantly, after Malawi became a Republic in 1964, the government assumed a more active role in education by opening schools, and offering universal primary education. Malawi's primary school education shifted towards using Chichewa as the language of instruction. Chichewa was designated a national language alongside English, the latter was used for official and business matters. In secondary schools and higher education, English is used in teaching and learning. The language policy in Malawi education has been an issue of debate, with some arguing that the use of the native vernacular will improve children's learning of new concepts, while others arguing that the need for 'language switching' (mainly between English and Chichewa) in classrooms brings tensions and pressures especially in the teaching of abstract and science subjects (Kaphesi (2003)). Others brought special attention to a worrying 'vocabulary gap' that exists between Chichewa and other official indigenous languages of Malawi (Chiuye & Moyo (2008)). The lack of adequate teaching and learning materials in local languages creates 'linguistically impoverished and deprived' learners (Kamwendo (2016)) and 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)). Studies on the impact of language switching on the teaching of mathematics in schools in Malawi showed that, although teachers had problems in translating and finding the terminology that best describes some of the mathematical concepts, they did not receive systematic training in the use of language (Kaphesi (2003)). Vernaculars and glossary of terms will help children understand concepts that they would otherwise find difficult to understand if taught only in English (Kayambazinthu (1998)).

2 A GROWING VOCABULARY GAP FOR ENVIRONMENT AND CLIMATE CHANGE

Malawi is a beautiful country with varied ecosystems (e.g., the well known Lake Malawi), grasslands areas and forests. These ecosystems have seen massive degradation over the years. Efforts to restore the habitat, wild animals and protect the forests have been made by international organisations working with the Malawi government (UNESCO (2017)). It is not the purpose of this proposal to discuss all of these, but to emphasise that all these players recognise the need to increase knowledge and awareness about climate change and make it available to school children and their educators. Large sums of money have been spent by international supported campaigns in Malawi through posters, community discussions and recently, videos , to encourage the planting of trees, public and personal sanitation, and avoid littering and charcoal burning. A recent comprehensive study on the charcoal market in Malawi, noted that there is a scarcity of accurate and good quality information on the state of things in Malawi, and that contributes to the maintaining of the status quo and a continuing degradation of the environmen (Kambewa et al. (2007)).

There seems to be a paradox between the fact that Africans are now seen to be environmentally unfriendly and cannot be expected to make substantive contributions to the world's environmental problem (Ikuenobe (2014)), and the African conception of the world, ubuntu, in which man is seen to exist in harmony with nature and thus articulates a moral attitude towards the environment. While these are complex issues, it is evident that communication in the local languages about these issues plays an essential role. In a study on the causes of deforestation in Mwazisi (near Vwaza Marsh Game Reserve), the low levels of awareness among the local population regarding forest use and management was identified as one of the factors contributing to forestry cover reduction (Ngwira & Watanabe (2019)). 'Traditional ecological knowledge that is, local people's classification, knowledge, and use of the natural world, their ecological concepts, and their resource management institutions and practices' is at an especially high risk of disappearing if not adequately documented (Maffi (2001)).

In the west, alienation from nature and a growing incapacity to experience it, have led educators to incorporate long term educational programs in schools, to bring pupils in contact with nature and to enhance their understanding of issues related to the environment and its protection. Special focus is given to pupils using 'talk' to organise, and express their ideas, opinions and feelings about their environment imaginatively (The UK National Association for Environmental Education) . There is a clear emphasis on learning, on using scientific language and the correct use of vocabulary in context but also to explain meaning.

Alienation from nature is happening in Africa too, although in different ways. For example, some have pointed out that a large portion of the indigenous vocabulary and knowledge remains unknown or is slowly disappearing (Ikuenobe (2014) and Cloete (2011)). From a list of more than 20 categorisations of landscape types in the language Xitsonga, compiled by Wolmer very few are now recognised by the younger generation (Duffy (2008)). The list demonstrates centuries of keen observation and experience of the local population, with terms ranging from words such as Kuthuma denoting thicket (hiding places of hyenas, leopards and lions) to Patsa denoting open areas where buffaloes graze, to Chawunga /mananga, denoting a remote, quiet and fearful area where only birds, and wild animals are found. Each of these terms encapsulates a rich meaning about what it defines. Many of these terms are now lost to urbanized Xitsonga speakers(Cloete (2011)).

There is a need therefore to bridge the vocabulary gap by creating glossaries that allow learners to name the environment around them and thus recognise issues of climate change by 'maintaining as much control over meanings as possible' because 'by naming the world people name their realities (Hall & Smith (2000)).

3 The role of Machine Learning

We believe that ML has a role to play in closing the 'vocabulary gap' of terms and concepts regarding the environment and climate change that exists in Chichewa and other Malawian languages by helping to creating a visual dictionary of key terms used to describe the environment and explain the issues involved in climate change and their meaning. Chichewa is a descriptive language, one English term may be translated using several words. For example, "pollution", without specifying what kind of pollution it refers to, does not have a direct counterpart in Chichewa. "Air pollution" may be translated as "kuwonongeka kwa mpwenya wa chilengedwe", where 'chilengedwe' may mean environment but usually is used to mean 'creation', and is also used to refer to "natural resources", "luso lachilengedwe'.

There are several practical steps in which ML can be used. We propose the task of building of a glossary for the environmental science (similar to the one on wikipedia for English), in Chichewa and other local languages used in Malawi using text available on the internet. Some of this text is obtained by machine-generated translations but some is written by native speakers. The interesting thing is not to detect just literal translations (where these are possible), but also translations by means of 'descriptions' and illustrations and thus extract correspondence between terms used and perhaps a measure of how appropriate a term is to convey the meaning intended. As part of this project, we will identify 'loanword patterns', which may be useful in understanding the transmission of cultural items. In many Bantu languages (such as Chichewa), lexical borrowings may be distinguished from the inherited vocabulary on the basis of phonological irregularities.

The vocabulary thus created using ML, can be cleaned by a Chichewa linguist/ speaker. The following examples of translations were done by Paul Kazembe, a senior teacher who teaches Chichewa as a Secondary school subject. Paul's translation can also inform the algorithms used to extract various definitions. We are using in this proposal some of his translations. I have asked that he translated a list of terms first using no technical books or dictionaries, purely by using his own understanding (Appendix A).

Words such as 'arable land' have an established meaning in Chichewa 'malo olima' - which literally means 'the land of the farmer' or 'agricultural land'. There is also possibly a borrowed term that is used and Paul translated 'arable land' as 'minda'.

Terms such as 'acid rain' are hard to translate. The Chichewa for rain is 'mvula'. Hence a translation by description is used. Notice that the word 'acid' is a loan word from English.

"MVULLA KAPENANSO MADZI OGWA KUCHOKERA MLENGALENGA OMWE AMAKHALA NDI ASIDI."

Similarly in translating 'manure' the loan word 'manyowa' may be used or the expression 'zinyalala zowolerana'.

Another good example is the word 'aquaculture' which was translated by Paul as 'Ulimi wa za mmadzi' which literally means 'farming on water' and will need a contextual description in order to be fully understood.

The term "adaptation (to environment)" was translated as "kuyanjana ndi nyango" which literraly means "reconciliation with the climate" (the word kuyanjana means reconciliation) or 'Kugwirizana ndi malo' where 'malo' literally means place and kugwirizana means agreement, relationship or union.

The same for 'carbon footprint', which Paul translated loosely as 'kuyeza', but a definition by description is more appropriate:

"KUCHULUKA KWA MPWEYA WOIPA OMWE WATUMIZIDWA MLENGALENGA NAWONONGA CHILENGEDWE KWA NTHAWI YONSE WOMWE: MACHINIWO AKHALA AKUGWIRITSIDWA NTCHITO. KAPENA CHIPANGIRENI CHINTHU CHINA CHAKE."

Words such as backflow, carbon neutral, cell, condensation, consumer, drainage, fossil fuel, groundwater, habitat, landfill are hard to translate in Chichewa and need a translation by context or illustration, hence are harder to translate immediately even by language proficient like Paul who has a rich English and Chichewa vocabulary.

4 THE PROPOSAL

What we propose is as follows:

(1) To start from a list of 'seeds' (see Appendix A), which are translation of environmental and climate change terms by language experts such as Paul, and dictionary definitions from Chichewa-English dictionaries. We use these seeds in searching over the internet for usage. Of interest would be to detect which of the results retrieved are in fact machine translations.

(2) Using these seeds, to gather content from the internet in Chichewa on the topics of environment, descriptions of nature and wildlife, climate change. Some text will be original writings in Chichewa, some would be human translation of articles written in English or other languages (or based on these articles), and some will be text generated using machine translators (for example Google translate). ML techniques can be used to aid in detecting the type of document and de-coding the translation to identify key terms (Dzmitry et al. (2014) and Baroni & Bernardini (2006)).

(3) To generate a glossary of environmental terms together with meaning, examples of usage, and a measure of how appropriate a term is to convey the meaning intended e.g., based on 'selective concept extractions' (Riloff (1993)).

(4) To analyze similarities between Chichewa texts and English text on climate change to detect loan words and the presence of code-switching (Ehara & Tanaka-Ishii (2008)).

(5) From the searches which we get when searching with 'seed terms', we want to extract images which appear inline text and, by using both the content surrounding them and actual picture captions, to tag them and add them as a pictorial representation of the terms of the glossary (Devlin (2015) and Bai & An (2018)).

ACKNOWLEDGMENTS

We thank Paul Kazembe for checking and helping with the Chichewa translations.

REFERENCES

Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311: 291–304, 2018.

Marco Baroni and Silvia Bernardini. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3), 2006. ISSN 02681145. doi: 10.1093/llc/fqi039.

Grace Chiuye and Themba Moyo. Mother-tongue education in primary schools in malawi: From policy to implementation. *South African Journal of African Languages*, 28(2), 2008. ISSN 23051159. doi: 10.1080/02572117.2008.10587309.

Elsie L. Cloete. Going to the bush: Language, power and the conserved environment in Southern Africa. *Environmental Education Research*, 17(1), 2011. ISSN 14695871. doi: 10.1080/13504621003625248.

Jacob et al Devlin. Exploring nearest neighbor approaches for image captioning. 2015.

Rosaleen Duffy. From Wilderness Vision to Farm Invasions: conservation and development in Zimbabwe's south-east lowveld by W. Wolmer Oxford: James Currey, 2007. Pp. 320. £17.95 (pb). *The Journal of Modern African Studies*, 46(4), 2008. ISSN 0022-278X. doi: 10.1017/s0022278x08003601.

Bahdanau Dzmitry, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*, 2014.

Yo Ehara and Kumiko Tanaka-Ishii. Multilingual Text Entry using Automatic Language Detection. *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, 2008.

Thomas D. Hall and Linda Tuhiwai Smith. Decolonizing Methodologies: Research and Indigenous Peoples. *Contemporary Sociology*, 29(3), 2000. ISSN 00943061. doi: 10.2307/2653993.

Polycarp A. Ikuenobe. Traditional African Environmental Ethics and Colonial Legacy. *International Journal of Philosophy and Theology (IJPT)*, 2(4), 2014. ISSN 23335750. doi: 10.15640/ijpt.v2n4a1.

Mataya Bennet Kambewa, Patrick and, Sichinga Killy, and Johnson Todd. *Charcoal: the reality*. The International Institute for Environment and Development (UK), 2007. ISBN 978-1-84369-678-0.

Gregory Hankoni Kamwendo. The new language of instruction policy in Malawi: A house standing on a shaky foundation. *International Review of Education*, 62(2), 2016. ISSN 15730638. doi: 10.1007/s11159-016-9557-6.

Elias Kaphesi. The influence of language policy in education on mathematics classroom discourse in malawi: The teachers' perspective. *Teacher Development*, 7(2), 2003. ISSN 17475120. doi: 10.1080/13664530300200190.

Edrinnie Kayambazinthu. The language planning situation in Malawi. *Journal of Multilingual and Multicultural Development*, 19(5), 1998. ISSN 01434632. doi: 10.1080/01434639808666363.

L. Maffi. *On biocultural diversity: Linking language, knowledge and the environment*. Washington: Smithsonian Institution Press, 2001.

Susan Ngwira and Teiji Watanabe. An Analysis of the Causes of Deforestation in Malawi: A Case of Mwazisi. *Land*, 8(3), 2019. ISSN 2073-445X. doi: 10.3390/land8030048.

Ellen Riloff. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the National Conference on Artificial Intelligence*, 1993.

UNESCO. *Climate Change Risk in Malawi: Country Risk Profile*. 2017.

A Appendix

This section contains Chichewa translations for climate change related terms (Paul Kazembe). When a translation is not known, there are empty columns.

English Term	Chichewa Tterm
Acid Rain	Mvula Mvula ya asidi.
Adaptation	Kuyanjana ndi nyengo / Kugwirizana ndi malo
Afforestation	Kudzala mitengo.
Agroforestry	Ulimi wamasomphenya
Air pollution	Kuwonongeka kwa mpweya
Carbon Footprint	Kuyeza . . . ... Muyezo wa . . . ...
Carbon Neutral	
Carbon Taxes	
Carnivores	Nyama
Catchment area	Malo
Cell	Tinthu ting'onoting'ono tamoyo timene timapanga thupi la chamoyo.
Climate Change	kusintha kwa nyengo
Cloud	Mtambo
Cold Air	Pempho yozizira Kapmphepo kayaziyazi.
Commercial and Industrial Waste	Kusamala zinyalala. Zotsalira
Compost	Manyowa Zinthu zowolera mnthaka
Composting	Kuwola Kusanduka manyowa

English Term	Chichewa Tterm
Condensation	
Consumer	Munthu Nyama Makampani
Consumption	Kudya Kugwiritsa ntchito
Controlled Burning	Kuwotcha zinyalala
Crop Rotation	Ulimi wa kasinthasintha
Cyclone	Mphepo
Decomposers (e.g., bacteria decomposing matter)	Tizilombo ta matenda
Deforestation	Kuwononga Chilengedwe
Desalinisation	Kuchotsa mchele
Desert	Chipalamba Malo a mchenga okhaokha
Dew	Mame
Drainage	Madzi osapindulitsa.
Drinking Water (potable water)	Madzi akumwa Madzi opanda matenda
Drought	Chilala
Dryland Salinity	Mchere wa mnthaka
Ecology	Maphunziro a pa zamoyo
Ecosystem	Kukhalira pamodzi modalirana.
Endangered Species	Chiwopsezo
Energy	Mphamvu
Energy Efficiency	Mphamvu
Environment	Zomwe wayandikana nazo Zomwe zakuzungulira
Erosion	Kukokoloka kwa nthaka
Evaporation	Nthunzi Kuuluka kwa madzi.
Feedback	Zotsatira
Fertilizers	Fetereza Mchere wa mnthaka
fog	Chifunga
Food Chain	Kudalirana kwa za moyo
Food Security	Kukhala ndi chakudya.
Forest	Nkhalango Dondo
Fossil Fuel	
Garden Organics	Zinthu zomera mmakomo
Genome	
Global Warming	Kusintha kwa nyengo.
Governance	Utsogoleri
Greenhouse effect	
Greenhouse Gas	Mpweya
Groundwater	Madzi a nthaka
Growth	Kukula
Habitat	Malo
Heat	Kutentha
Heavy Rain	Mvula yambiri Mvula yosakata
Herbicide	Mankhwala
Herbivores	Nyama zodya thengo
High pressure area	Malo osowa chakudya Malo otentha
Hydrosphere	Madzi
Incineration	Kuotcha zinyalala. Kuwotcha zinyansi.
Industrial Agriculture	Makampani pa za ulimi
Infiltration	Madzi akulowa pansi Kulowa pansi kwa madzi
Inorganic Matter	
Intercropping	Ulimi wa kasakaniza
Irrigation	Ulimi wamthirira Ulimi wa chilimwe
Land Use	Ubwino wa malo
Landfill	Zinyalala zomwe zakwiliridwa
Landfill gas	Mpweya woipa
Lightning	
Low pressure area	

English Term	Chichewa Tterm
Microorganism	Kanthu kamoyo
Mist	
Monoculture	Ulimi
Mortality Rate	Imfa
Natural	Chilengedwe
Natural resources	Zinthu zachilengedwe
Noise pollution	Phokoso
Nutrients	Chakudya
Pesticide	Mankhwala
Plastic	Pulasitiki
Pollution	Kuwononga
Power	Mphamvu
Precipitation	Mvula
Precipitation	
Producer	
Product	Chinthu
Productivity	Phindu
Rain	Mvula
Rainwater	Madzi a mvula
Rainwater harvesting	Kukolola madzi a mvula Madzi amkolola
Raw materials	
Recycling	Chibwereza
Reforestation	Kubwezeretsa chilengedwe Kudzalanso mitengo.
Reuse	Chibwereza Kugwiritsanso ntchito
Risk	Vuto
Septic sewage	Suweji
Sewage	Suweji
Soil	Dothi
Sun	Dzuwa
System	Chilinganizo
Topsoil	Dothi lapamwamba
Tropical Cyclone	Mphepo
Warm Air	Mpweya wotentherako
Waste	zinyalala
Waste Management	Masamalidwe a zinyalala
Water Cycle	Zokhudza madzi
Water Quality	Madzi abwino
Water Table	Madzi a pansi panthaka
Weather	Nyengo
Wetlands	Malo a chinyontho
Wind Energy	Makina oyendera mphepo
Wind Turbine	Mphepo
Work	Ntchito