On English to Kinyarwanda machine translation

Emmanuel HAVUGIMANA

June 24, 2017

1 Statistical Machine Translation or Rule Based Machine Translation

Currently, I have not encountered an machine translation from another language to Kinyarwanda that is robust or have a tolerable accuracy. By tolerable accuracy, I mean good noun class consistence as they are like what singular and plural are to English and additional feminine and masculine to French..

Sentence either sounds awkward and may confuses read.

1.1 Specifics on Kinyarwanda

Kinyarwanda is a bantu language; like all bantu language it is noun class based(with about 16). For classes, a place have a different marker than a person . for example

- Kigali is a good place → Kigali ni ahantu heza
- Emmanuel is a good person → Emmanuel ni umuntu mwiza

there are parts of adjectives and mostly verbs that depends on noun-classes that has no such dependence in corresponding English texts

Hence, trying to roll out SMT tools on it will capture less repeatability verb changes a lot, and adjectives change a lot. But also trying to do only rule based system neccessites the manually updating lists of roots, rules, and a good databases

1.2 On what I am trying to look at

A rule based core and a statistical search for roots, patterns that confirms to kinyarwanda constraints of Uturemajambo(word construction) and orthography. orthography is a bit constrained in a manner that I can look at a word and have a high chance of telling if it is kinyarwanda or than saying so in English and English words have a lot of consonants combination

1.2.1 On orthography

- Every consonant must be followed by vowels or is a part of allowed consonant combinations (ibihekane) that has also to be followed by a vowel (in most case) Here are list of allowed combinations as by 2014 law that can be found at

https://goo.gl/t7TWZG. It constrains the combinations to the following combinations

(["ny","sh","pf","ts", "mb","mf","mp","nd","ng","nk","ns","nsh","nshy","nt","nt","nz",
"bw","bg","cw","dw","fw","gw","hw","jw","kw","mw","nw","nyw","pfw","pw","rw","shw","shyw",
"by","cy","jy","my","nny","pfy","py","ry","sy","ty","vy", "byw","myw","pfyw","ryw","vyw",
"mbw","mfw","mpw","mvw","ndw","ngw", "njw","nkw","nkw","nshw","nshyw","nsw","ntw","nzw",
"mby","mpy","mvy","ncy","ndy","njy","nsy","nty", "mbyw","mvyw","njyw"])
in addition to consonants(Same as English consonants)

- There additional constraints that restrict some vowels following some consonants combination which can work as additional constrain that are described in the text
- Vowels can not follow each other in a word — "food" violate and every word should end with a vowel or "", but also there is a constrain on what can be followed by "", (can be inferred from constraint that consonants should be followed by a vowel) (kinyarwanda syntax),
- (Thinking of adding number) – finding an approximate number of allowed consonant combinations in English texts –as compared to Kinyarwanda text –probably higher

1.2.2 On Uturemajambo

In kinyarwanda Verbs are conjugated by attaching tense markers on verb root, in additional subject, object and other sort of markers. With all markers on on the root, there are rules that dictates how to go on recovering the word from its constituents

I am implementing uturemajambo by replacements

2 Rule Based Machine Translation(RBMT)

2.1 Kinyarwanda test

2.1.1 why Kinyarwanda testing

Being able to test the Kinyarwanda level of test can be great in analyzing the translated text before being returned to test if it is Kinyarwanda.

2.1.2 Using syntax

With constraints on what constitute Kinyarwanda. I can use rules to select texts that can be entered into Kinyarwanda corpora (as one layer in checking process) or syntax checking

This test can allow testing new words conformity and can is faster than word search and can perform better than word search as it can address strange type of writing like poetry (even when it is first encountered) — but it can also include false positives(Like detecting Kirundi as kinyarwanda or any language with almost similar syntax constraints)

```
def kinyarwanda_level(texts):
""" Kinyarwanda_level takes in a text returns the  a number between 0 and 1 level of  kiny
for example kinyarwanda-level(["inyarwanda ni indirimbo nyarwanda"]) returns 1.0"""
count=0
total=0
for i in (consonants_in(texts)):
if len(i)>1:
total+=1
if is_it_kinya_consonant_combo(i):
count+=1
else:
continue
return count/total
```

2.1.3 using vocabulary test

I have not implemented this approach as there are many Kinyarwanda words out there to be formed and also it can computation intensive than just parsing the text... One implementation that I can consider including is have an updated list of Kinyarwanda words(say sorted) and do search for presence of each word as I go around text and if most of words in text happen to be Kinyarwanda words , call it Kinyarwanda, or use a number of most frequent words for less search space and time.

2.2 verb centered Sentence construction

As verb conjugated include subject, objects ,tense marker and other sort of parts, I am going to focus to constructing a sentence from this central parts.

Currently, thinking for using kinyarwanda and English parallel sentences from Bible, rwandan consitution and other laws to to extract most important roots(manually by me or other persons who are into the success of the system) and also using a statistical approach to getting roots by say finding two sentences which have same verb (in english, but have different other words) –like 5 and the do analysis for word that may be repeating (I may choose to restructure Kinyarwanda First, or not depending on the added benefits and cons of doing it that way –like capture character changes that may have resulted from uturemajambo after adding word parts together)

- Root finder, or root manually put together with their corresponding English verbs
- Find the order of parts on verb
- add the parts (like tense marker), subject markers

3 Statistical analysis

3.1 Frequency tables

I can get word frequencies and other related frequencies like following consonants or words that can help me predict which noun class the noun belong and auto detect verb tenses – in cross translated texts —Hopefully root extractor

For example it is more likely that if "ba" mostly follows abana, then abana is in "ba" group and can the way to try to experiment with it is to find each word and retrieve the next vowel or if consonant(two letters) with uturemajambo reversing for example by —ma be retrieved as bi, and bw as bu for having consistent representation.

examples

- Imana ibishyirira mu isanzure ry'ijuru kugira ngo bivire isi,
- Imana ibaha umugisha

There is a relationship between Imana and the I that mostly comes after Imana while in sentence.

(Can I capture some of this code and have the system help me and update some of my features) Like classify for me the nouns in noun classes