InterlinguaPlus Machine Translation Approach for Under-Resourced Languages: Ekegusii & Swahili

Edward O. Ombui[1, 2], Peter W. Wagacha[2] and Wanjiku Ng'ang'a[2]
[1] Computer Science Dept. Africa Nazarene University (Kenya)
eombui@anu.ac.ke
[2] School of Computing and Informatics, University of Nairobi (Kenya)
waiganjo@uonbi.ac.ke, wanjiku.nganga@uonbi.ac.ke

Abstract

This paper elucidates the InterlinguaPlus design and its application in bi-directional text translations between Ekegusii and Kiswahili languages unlike the traditional translation pairs, one-by-one. Therefore, any of the languages can be the source or target language. The first section is an overview of the project, which is followed by a brief review of Machine Translation. The next section discusses the implementation of the system using Carabao's open machine translation framework and the results obtained. So far, the translation results have been plausible particularly for the resource-scarce local languages and clearly affirm morphological similarities inherent in Bantu languages.

Keywords: Machine Translation, InterlinguaPlus, Ekegusii

1. Introduction

Development of language applications for local languages in Africa requires innovative approaches since many of these languages are resource scarce. By this we mean that electronic language resources such as digital corpora, electronic dictionaries, spell checkers, annotators, and parsers are hardly available. These languages are also predominately spoken rather than written. Moreover, they are generally used in environments where there are other competing languages like English and French which have been well documented over the years with properly defined grammars, unlike the local languages with poorly defined grammars and dictionaries. This has been a major setback in the development of technologies for African languages. The presence of diacritics in most of these languages has also contributed to the complexity involved in the development of language technology applications. (Ombui & Wagacha, 2007).

Nevertheless, there is pioneering work with the South African languages, which includes the definition of proper language grammars and development of a national language policy framework to encourage the utilization of the indigenous languages as official languages (NLPF, 2003).

In this paper, we consider two Bantu languages in Kenya namely Ekegusii and Swahili. There are approximately two million Ekegusii language speakers (KNBS, 2009). Swahili is widely spoken in East and Central Africa and is one of the official languages of the African Union with lots of printed resources.

For the work that we are reporting, we have adopted the InterlinguaPlus approach using the Carabao open machine translation framework (Berman, 2012). In this approach, all similar meaning words, synonyms, from each language and across the languages existing in the system are stored under the same category and assigned an identical family number. These words are also tagged with numbered lexical information[1]. For example, *Egetabu* (a book) [1=N; 2=SG; 5=No]. Tag1 stands for the part of speech (1-POS), Noun, tag2 for number (2-No.), Singular, and tag5 indicates whether the noun is animate or inanimate etc. An amalgamation of the word's family identification number and tag numbers form a unique ID for the word. In addition, a novel way of only storing the base forms of each word and having a different table containing affixes that inflect the word drastically reduces the lexical database size and development time in general. This approach is implemented through the manual encoding of the sequence rules for the two languages.

Preliminary results are encouraging and clearly reveal similarities in the language structure of Ekegusii and Swahili. The advantage of this approach is that the translation is bidirectional and maintains the semantic approach to translation just as a human translator. In addition, it is suitable for rapid generation of domain specific translations for under-resourced languages.

[1] Grammatical, Stylistic and Semantic tags

Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 68–72,
Baltimore, Maryland, USA, 26 June 2014.

2. Machine Translation

Over the history of MT, several techniques and approaches have continued to be developed despite previous discouraging reports (ALPAC, 1966). The major approaches and methodologies include: Rule-based and Corpus-based, Direct translation and indirect translation (i.e. transfer-based and Interlingua-based) (Hutchins, 1993 & Hutchins, 1994). With the introduction of Artificial Intelligence technology in MT, more recent approaches have been proposed including alignment template approach to Statistical MT (Och & Ney, 2004), Knowledge-based approach (Nirenburg et al., 1992), Human in loop, and Hybrid methods (Groves & Way, 2006).

One of the strengths of the InterlinguaPlus approach (Berman, 2012) is that it preserves semantic information of the lexicon. Therefore, translation is primarily based on semantic equivalents between the lexicons of these languages.

As a result, the traditional language pair-based translation is replaced by bidirectional translations between the languages existing in the system. Any language can be the source or a target language.

Consequently, the lexical database size is drastically reduced and the task of building multiple dictionaries is concentrated in constructing just one Interlingua lexical database. This kind of approach is evidently advantageous when building machine translation applications for under-resourced African languages because it expedites the process of adding a new language with minimal effort especially when adding languages of similar grammatical makeup, which could reuse some of the existing grammar rules.

3. Implementation

Figure 1 below illustrates the translation process in the Ekegusii Machine Translation (EMT) system. The user inputs a sentence, which is parsed into its constituent tokens. These tokens are then matched and mapped to their equivalent target-language tokens using the Family and mapping Identification numbers respectively. In addition, the sequence[2] e.g. Subject+Verb+Object is parsed into elements (lexical units) and authenticated against the elements of the analyzed sentence. If it is valid, the elements are mapped according to the sequence and modified by the corresponding sequence in the target language. Some of the features that can be modified include deleting or adding a new element. E.g. He ate a mango.[eng:SVO]. *A+li+kula Embe* . Note that Swahili and generally the local African languages do not have determiners. Therefore, when translating from Eng-Swa, the English determiner is dropped. However, it is added if the translation is vice-versa. This is made possible by assigning a locally unique identity number, preserved across languages in the database, to each lexical unit of a sequence. The sequence manager in the system uses these identity numbers to appropriately handle lexical holes and the source/target of each transformation.

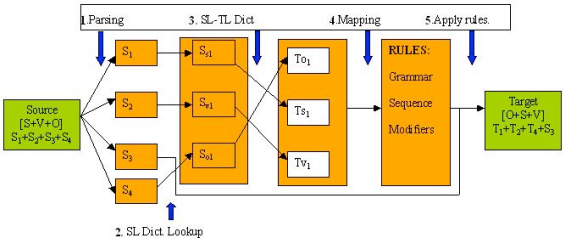

Figure1: EMT's MR-PDF

Subject (S_1); Verb (V_1); Object (O_1); Determinant (det); delimiter (del).

The above process, MR-PDF[3], is an acronym for the five translation stages (explained below) with the last two stages shifted at the beginning so as to give it an easy-to-remember name. We will use example 1, English to Ekegusii SVO phrase to elucidate the process.

Example 1

He ate a mango.

Stage 1: Parsing

The sentence is analyzed syntactically according to its constituent structures i.e. tokens including syntax delimiters like question marks, exclamation marks etc.

He + ate + a + mango.

S_{s1}:[He] S_{v1}:[ate] Det:[a] S_{o1}:[Mango] del:[.]

It is worth noting that at this stage, the parts of speech have not yet been identified.

Stage 2: Source Language Dictionary Lookup

Each token from stage 1 is looked up in the respective source language dictionary to check whether it exists in that language. In case it is not

[2] Set of elements, which refer to tokens that have specified features e.g. grammatical data, style, word-order, etc.

[3] Mapping (M), Rules (R), Parsing (P), Dictionary look-up (D), Family word-match (F).

found, the word is left untagged and passed-on as it is to the next stages up to the output.

Stage 3: Family word-match
Every morpheme is examined considering all possible combination of affixes to it and each configuration stored. These are then aligned with the corresponding target language dictionary entities.
[He]= [*Ere*]
[ate]= [*ariete*] Past form of eat=*karia*
[a]= [a] yields the same token if an equivalent is not found in the target language.
[Mango]= [*Riembe*] Singular, noun.

All other delimiters, e.g. question marks (?), comas (,) are presented as they appeared in the source string. From the above example, all possible modifiers of the verb "to eat" are generated i.e. eat, ate, eaten, eats, eating, and matched with the corresponding verb in Ekegusii dictionary i.e. *Karia, ariete, nkoria,* etc.

The tricky part of it is that one may not always have an equivalent number of modified verbs in the target or source dictionaries. To resolve this ambiguity, the program picks the modified verb with the best match in the target language dictionary i.e. in terms of matching lexical or style information e.g. the type of tense, number, animation, gender etc.

If we refer to the same example above, the following is examined as shown in Table 1and Table 2.

Language	**Morpheme**	**Part Of Speech**	**"Modified Morphemes"**
English	Eat	Verb	Ate; eaten; eating, eats, etc.
Ekegusii	Ria	Verb	Karia, ariete, nkoriare, etc.

Table 1: *Lexical information*

"Modified Morphemes"	**Tense**	**Number**
Ate	Past	Singular or Plural
Eating	Present continuous	Singular or plural
Mbariete	Past	Plural
Ariete	Past	Singular

Table 2: *Style information*

Language: English
Ate [tense-past; number-any]

It is apparent that both dictionaries are used to provide grammatical information, semantic data and potential equivalents in the target language during this stage.

Stage 4: Mapping
At the mapping stage, the Source text is validated against all existing sequences trees in the language. Only the most complete and detailed tree is picked. From example 1 above, the most appropriate sequence tree will be as follows and illustrated in figure 2.

He ate a mango	*Ri-embe a-rie-te*
[PN] + [V] + [Det] + [N]	[N] + [V]

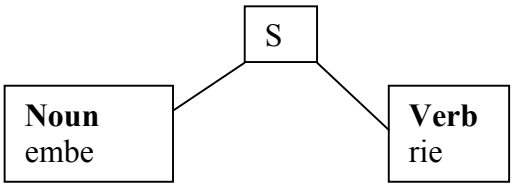

Figure 2: Sequence tree

The elements in the source sequence will map exactly into the [N] + [V] sequence. At this point all the redundant guesses are eliminated and disambiguation occurs. There are more comparisons and checks - like subject and style checks, etc.

Stage 5: Apply Rules.
The elements in the source sequence are modified by the corresponding sequence in the target language. The affixes are attached, or some new elements added or others completely deleted. Each element's unique identity is used to map the source sequence to the equivalent target sequence identities. Remember that Ekegusii does not have determiners and therefore it is dropped.

From the example above, the noun is then modified by adding the singular prefix- *ri*, (noun class 13) while the verb is modified by concatenating the subject- *a* (singular pronoun) to the verb- *rie* and finally adding the suffix- *te* (Past tense). The final sentence then becomes as shown below
Riembe ariete -> ri-embe a-rie-te

In case it is converted to plural, the noun prefix will change to- *ama* (noun class 6) and the pronoun to- *ba* while maintaining the past tense suffix- *te*

Amaembe bariete -> Ama-embe ba-rie-te

Finally, the sentence word order is rearranged according to the best fitting sequence tree in the target language sequence table.

4. Results

The results gotten so far are plausible. The word order is correct as per the programmed sequence rules for each language e.g. English: *This is a book*; Ekegusii. *Eke n'egetabu*; Kiswahili: *Hiki ni kitabu*. In addition, the bidirectional functionality is often more than 50% accurate on the wider domains and about 90% accurate on specific domains, in our case the obituary's domain. This evaluation is based on phrase level. Besides, once a phrase text has been translated, it can also be used as the source text and the translator will yield the exact translation as the initial source text. This therefore makes a strong case for the high intelligibility of the system.

The idea of storing only the word base forms and having a separate table for the affixes has drastically reduced the lexical database size as well as the building time. It was also noted that there is need for careful configuration of the rule units[4] for the affixes and lexicon otherwise the translation will be inaccurate. If we are to use the example above, the canonical[5] form will be as follows: English: FID-144 Book [POS: N; Number: SG; Animation: No]. However, for Ekegusii, there is need for additional rules units to indicate the noun class[6] because the nouns inflection is dependent on the noun class, otherwise the machine translator might concatenate the wrong prefix. Therefore, the English example above will be matched as follows. Ekegusii: FID-144 ***tabu*** [POS: N; Animation: No, EkeNC[7]: 8/9].

Consequently, the translator compares the rule units of the word with the rule units of the modifiers[8] in the affixes table and picks the most matching affix, in this case the prefix "*ege"* [POS: N; Number: SG; Animation: No, EkeNC[9]:8/9], ensuing n accurate translated word "ege***tabu***". On the contrary, if the Ekegusii rule units were not added or wrongly configured, the translation will be bizarre e.g. "Omo***tabu***" which is an invalid Ekegusii name. In fact, the prefix "omo" [EkeNC: 1] is often reserved for singular human[10] nouns.

The results obtained also expound the diversity of Ekegusii language linguistic rules[11] as compared to English. Most Indo-European languages, specifically English, espouse the SVO[12] sentence structure rule. However, in Ekegusii both SVO and VOS rules are valid sentence structure rules. For example, English: Mum ate mangoes [SVO]. Ekegusii: 1.*Omog'ina nariete amaembe [SVO]*. *2. Nariete amaembe Omong'ina [VOS]*. Interestingly, the Ekegusii sequence and grammar rules that were copied and pasted to Swahili with minimal alteration resulted in almost precise translations between the two languages. This inevitably affirms the similarity in the language structure of the two languages and the ease in defining, constructing and translating between local languages as compared to/or from English.

The project demonstrations made so far to peers and students have generated a lot of enthusiasm in African languages research and given a good indication of the reception of technology in a familiar language platform.

5. Conclusion

The InterlinguaPlus approach is good particularly for under-resourced languages in terms of generating rapid translations that give a good gist of the meaning in the second language. Although it takes some time to write the grammar rules for a new language at the beginning, it however takes a relatively shorter time when adding languages of similar grammatical makeup. Therefore, the approach is very feasible especially when considering under-resourced languages which may not be afforded the appropriate finances and sufficient political will to have technological resources built for them.

The lexical database building methodology, whereby words and their grammatical data are stored in respective families and assigned a unique identification, provides an excellent way of reducing the chances of ambiguity that may exist in the phonetic disparities inherent in these local languages.

The InterlinguaPlus approach employed in the Carabao Open MT framework forms a good foundation to scale existing language resources

[4] A tag bearing any piece of grammatical data: part of speech, number contrast, gender, conjugation pattern, etc.
[5] Base form of the word before any inflection
[6] There are about 17 Ekegusii noun classes
[7] Ekegusii Noun Class
[8] In this case, Prefixes
[9] Ekegusii Noun Class
[10] Professions, etc.
[11] Sequence and grammar rules
[12] Subject, Verb, Object

to many other under-resourced languages using minimal effort i.e. the number of rules written for a language and consequently the time taken to develop a new language.

6. References

Automatic Language Processing Advisory Committee (ALPAC). 1966. Languages and Machines: Computers in Translation and *Linguistics*. National Academy of Sciences, National Research Council, 1966. (Publication 1416).

Declan Groves, and Andy Way. 2006. Hybrid Data-Driven Model of MT. ttp://citeseerx.ist.psu.edu/showciting?cid=5495125 (Retrieved March 15, 2014)

Declan Groves. Bringing Humans into the Loop: Localization with Machine Translation at Traslán http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.210.2867 (Retrieved March 15, 2014)

Edward Ombui, and Peter Wagacha. 2007. Machine Translation for Kenyan Local Languages. In *Proceedings of COSCIT conference*. Nairobi, Kenya.

Franz J. Och, and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. http://acl.ldc.upenn.edu/J/J04/J04-4002.pdf (Retrieved January 25, 2014)

John W. Hutchins. 1993. Latest Developments in Machine Translation Technology: Beginning a New Era in MT Research. *MT Summit* (1993), pp. 11-34.

John W. Hutchins. 1994. *Research Methods and System Designs in Machine Translation: A Ten-Year Review, 1984-1994*. http://www.mt-archive.info/BCS-1994-Hutchins.pdf (Retrieved August 1, 2013)

Kenya National Bureau of Statistics (KNBS, 2009). Ethnic Affiliation http://www.knbs.or.ke/censusethnic.php. (Retrieved September 6, 2013)

National Language Policy Framework. 2003. Retrieved from the Department of Arts and Culture website of South Africa. https://www.dac.gov.za/sites/default/files/LPD_Language%20Policy%20Framework_English_0.pdf

Sergei Nirenburg, Jaime Carbonell, Masaru Tomita, and Kenneth Goodman. 1992. Machine Translation: A Knowledge-Based Approach. http://acl.ldc.upenn.edu/J/J93/J93-1013.pdf (Retrieved November 22, 2013)

Vadim Berman. 2012. Inside Carabao: Language Translation Software for XXI Century. Retrieved from the LinguaSys website http://www.linguasys.com/web_production/PDFs/InsideCarabaoWhitePaper.pdf.