Sustainable language technology for African languages

Arvi Hurskainen
University of Helsinki

(words: 8060)

Introduction

Africa has a peculiar history of language development. Before colonial times the area south of the Sahara was speaking exclusively African languages. The colonial period changed the situation so that mainly three European languages, Portuguese, French and English, became the languages of the elite, while the local population continued to use their local languages. Now after the end of the colonial times, the situation continues to be much the same. One would have expected that the role of imported foreign languages would have diminished and local language policies would have developed communication systems based on local languages. However, this has not taken place. Those three foreign languages continue to dominate in official matters across Africa, although most people are hardly able to communicate using these languages. The elites in each country employ for government business a language that the ordinary people do not understand. Therefore, the majority of the population are marginalised and excluded from power politics. Although there is much talk about the importance of local languages, very little concrete actions are made to improve the situation (Myers-Scotton 1990; Bamgbose 1991, 2000).

The digital age potentially offers new possibilities for developing the status of indigenous languages, such development would be in line with UN resolution on rights of indigenous people, “that control by indigenous peoples over developments affecting them and their lands, territories and resources will enable them to maintain and strengthen their institutions, cultures and traditions, and to promote their development in accordance with their aspirations and needs”.[1] The UN specifically emphasize the rights to develop, maintain and transfer own language. The enhancement of local languages might aid proper intercultural communications across vastly different language divide.. Systems for translating from speech to speech have already been constructed and tested (Nakamura 2009). The subject of debate is still the way how one language should be translated to the other. For example, Google Translate (GT) has adopted the method of using English as interlingua, through which translation is carried out between two languages. Furthermore, GT uses statistical machine translation (SMT) methods (Och 2006), which it recently has enhanced with neural machine translation (NMT) methods (Turovsky 2016). The use of interlingua in translation process is a viable solution in the construction of global communication system between all languages.

Currently, the main trend in translation technology is to develop translation systems based on SMT and NMT. These methods are suitable for closely related languages that, in addition, have large masses of human-translated texts for training the translation system. According to Och (2005), a solid base for developing a usable statistical machine translation system for a new pair of languages from scratch would consist of a bilingual text corpus (or parallel collection) of more than 150-200 million words, and two monolingual corpora each of more than a billion words. Statistical models from these data are then used to translate between those languages. Unfortunately, these requirements are not satisfied with most African languages.

Translation systems

[1] 61/295. United Nations Declaration on the Rights of Indigenous Peoples

There are two main approaches for developing machine translation. One is a method, which makes use of detailed analysis of the source language, and converts the message into target language by making use of linguistic information and lexicon. The other method makes use of statistical likelihood of correct translation by comparing translation alternatives found in parallel corpus text.

Statistical machine translation (SML)

The increasing calculation power of computers has tempted researchers to test and develop also such approaches to translation, which were totally out of reach during the pre-computer time. An example is machine translation using statistical methods (Koehn 2010). Because statistical methods use calculation in the translation process without human input, these methods have become very popular in translating between major languages. For instance Google Translate and Microsoft Translator use statistical approaches in their applications. The largest translation consortium, European Union (EU), also develops translation applications using mostly statistical methods.

Given that statistical machine translation has received wide support and application, it is reasonable to ask if the same methods be applied to African languages. In order to answer this question, it is necessary to investigate whether African languages have sufficient preconditions for developing successful statistical translation applications.

For statistical translation methods to succeed, two major preconditions must be satisfied, (a) the relative similarity between language pairs, and (b) sufficiently large masses of parallel texts. By parallel texts we mean texts that have been translated from one language to another by humans. The parallel texts form the basis for the translation system to search for translation examples. These texts are used for training the system as well as in the actual translation process of new texts.

Let us see first the availability of parallel texts for various language pairs. Globally, the translated fiction books form the single largest source of parallel texts. There are millions of books translated by humans into several languages. Only part of this source, if made available, would form a huge source for training statistical MT systems. Copyright restrictions are an obstacle in using these resources effectively. European Union’s policy is to translate official documents into all EU languages. Also United Nations translates its official documents into six official languages. Over the years this work has been done by humans. These carefully translated documents from various domains form a huge database of parallel texts for machine translation. Since all types of new texts have been written using computers for several years, new parallel texts become available increasingly.

If we look at the availability of parallel texts in African languages, the situation is quite different. Apart from Bible translations there are very few parallel texts between African languages. Between former colonial languages and African languages the situation is better, but still far from the amounts needed. In some countries, such as Tanzania and Kenya, some government documents are in official languages, English and Swahili. These countries use English and Swahili as their official language. On the part of fiction, there are very few translated books, and even fewer of them are available in computer form. We can conclude that African languages do not have sufficient amounts of parallel texts for statistical machine translation systems to develop.

The second condition for statistical machine translation to succeed is the similarity of source language and target language. Africa has large amounts of closely related languages with very similar morphological and syntactic structure. For example, Bantu languages form such a cluster, the joint noun class system and similarity in word order form a fruitful basis for statistical machine translation. However, because parallel texts are not available, this second necessary precondition is not satisfied. Another problem is that the major need for text-based machine translation is between African and European languages and not between two African languages. Therefore, even the second requirement, that is, language similarity, is not satisfied. On the basis of above one is tempted to make a conclusion that statistical machine translation is not a viable solution in Africa.

Rule-based machine translation (RBMT)

Fortunately, in addition to statistical approaches, there is also the so-called rule-based, or symbolic, approach to machine translation. This approach in fact was common until the 1990s, and only when large masses of parallel texts became gradually available, statistical approaches started to dominate. The rule-based approach is radically different from the statistical approach. While the latter has nothing to do with linguistics or language theory, the former is entirely based on linguistic knowledge. In statistical approach, sequences of characters are compared in source text and target text, and the correct translation is selected on the basis of statistical likelihood. Because no language analysis is included in statistical approaches, no generalisations can be made. Therefore, model translations should be found for each surface word sequence. This leads to the requirement of ever increasing masses of parallel texts.

In rule-based approaches no parallel texts are needed. The basic components are (a) the comprehensive morphological analyser of the source language, (b) the morphological and semantic disambiguator of the source language, (c) the syntactic analyser of the source language, (d) a system for isolating multi-word expressions, (e) the bilingual dictionary for transferring the lexical information of the source language into target language, (f) a rule system for converting the lexical forms into surface forms, (g) a rule system for controlling the correct word order in target language. These are just the basic components, and additional computing for correcting and adjusting the process is needed[2].

In order to be able to construct the basic components of a rule-based system, one needs (a) a good dictionary of the language, and (b) a thorough knowledge of the grammar of that language. This means that for a non-linguist it is very hard to construct a rule-based system. It is like working out a comprehensive linguistic description of the language. In fact, a comprehensive language analysis system contains more grammatical information and is more accurate than any of the traditional grammar books. The construction process forces the developer to include into the system also such features that are not described in ordinary grammars.

When the language is described on grammatical level, it is possible to make generalisations. Instead of writing rules for surface phenomena, it is possible to write rules using linguistic tags and thus reduce the need of rule writing drastically. The system will be compact and it can be installed in most environments.

One major weakness of statistical methods is that it performs poorly in languages with complex morphology. In African languages, especially the Bantu verbs are particularly complex. A verb may have a large number of different forms, at least theoretically. Statistical methods can hardly cover all of them. In fact, when looking at the output of statistical translation systems, one finds failures especially in their handling of verbs. With rule-based methods it is possible to construct a full-coverage analyser that never fails to analyse a verb form, or any other form, if it is grammatically correct.

[2] https://en.wikipedia.org/wiki/Rule-based_machine_translation, retrieved 15.4. 2017.

Distrust between SMT and RBMT developers

A research team at Montreal University has suggested a new approach to machine translation (Cho et al. 2014). At the centre of the system is a neural network called RNN Encoder-Decoder that consists of two recurrent neural networks (RNN). In this system, one RNN encodes a sequence of symbols (usually words) into a vector representation, and the other RNN decodes the representation into another sequence of symbols. Neural machine translation is in fact an extension of statistical machine translation, and it is not yet known how successful it will be. In any case it requires large masses of parallel corpora and extensive computer power. And, above all, it ignores the explicit linguistic knowledge. The problems encountered in this method include its poor performance with rare words and long sentences, and methods are being sought for solving the problem (Luong et al 2015).

There is discussion on hybrid approaches, where statistical machine translation is enhanced by introducing linguistic knowledge into the system. What this precisely means is not known. And if the need of linguistic knowledge is acknowledged, why not introduce this knowledge at the start?

Hybrid approach

The research shows that neither SMT nor RBMT produces fully correct text. Both have weaknesses, although of very different kind. This has led to the idea of hybridization, enhancing the existing system with features of the other. Most often the approach has been to have a statistical approach as a base and the translation result is enhanced with rule-based components (Zbib et al 2012; Nielssen and Ney 2004). Also rule-based systems can be enhanced with statistical components (Habash and Monz 2009). It is claimed that better translation results can be achieved using a hybrid combination of these approaches (Labaka et al 2014). Nevertheless, the quality of translation continues to be relatively low and research continues for finding better quality testing methods (Felice and Specia 2013; Birch et al 2010).

Although SMT is the dominant method in MT, it is not self-evident that it is the best method in most cases. When the source language (SL) and target language (TL) are structurally very different, SML seems to encounter serious problems (Habash et al 2009). We see the same in examples below.

Comparing Google Translate and Salama

In the following we will make tests with two different systems in translating between structurally very different languages, Swahili and English. Currently the only SML system for Swahili is Google Translate (GT). And the only fully rule-based system available for translating between Swahili and English is Swahili Language Manager (SALAMA). Because both systems are ‘pure’ systems in that they do not include hybrid components, it is possible to compare the performance of both approaches. The advantage of choosing these languages is that GT will not be put to a disadvantaged position, because English is the core language of the GT system. Therefore, GT does not need to translate via English to a third language, as it normally does. Examples will be given on both translation directions.

Structure of tested translation systems

According to the information available, GT is based on the statistical approach. Apparently it is to some extent enhanced with morphological analysis, although this component seems elementary. No proper documentation on its structure has been available to us. Documents of Tanzania government have very likely been used in training the system. This conclusion can be made on the basis of fairly good translation results of these documents on the web.

SALAMA has a totally different approach. It has no statistical element. It makes heavy use of the grammar and lexicon of the language. In semantic selection it uses default translation. That is, each lexeme has a default gloss in the target language. This is selected if no rule selects another interpretation. The structure of SALAMA resembles the processing method of OpenLogos (Scott and Barreiro 2009; Barreiro et al 2011) including modular pipeline structure.

In the Swahili to English component, the core engine is the morphological analyzer based on finite state technology (Koskenniemi 1983; Hurskainen 1992). In disambiguation and syntactic mapping, as well as in isolating multiword expressions, SALAMA uses Connexor’s Constraint Grammar parser (Karlsson 1995; Tapanainen 1996; Hurskainen 1996; 2004).

The English to Swahili translator uses a Dependency Parser for English (Järvinen and Tapanainen 1997) for morphological analysis, disambiguation and syntactic mapping. The result is further modified with several sets of rules written in the Constraint Grammar environment (Tapanainen 1996) and with rewriting rules. The operation of SALAMA is described elsewhere (Hurskainen 2007; 2012).

We demonstrate the differences in performance between Google Translate and SALAMA using example sentences from the news media. There are examples of both translation directions. The purpose of this comparison is to show what are the strengths and weaknesses of each approach in translating between very different languages such as Swahili and English. The aim is not to prove the supremacy of one system over the other.

Examples from Swahili to English translation

(1) Tunaomba Watanzania wavute subira, tunaelewa kuwa watu wana shauku. Kuhusu jeshi hilo kushindwa kuwakamata watuhumiwa katika matukio ya milipuko ya awali iliyotokea Arusha, Jumapili alisema hata wahalifu na walioko nyuma ya matukio haya ni werevu na wana akili za kukwepa mkono wa dola.

GT
We Tanzanians NO patience, we understand that people are interested. About the army fail to arrest suspects in cases of initial explosions occurred Arusha on Sunday said that criminals and those behind these events are smart and do not mind the dodgy hand of dollars.

SALAMA
We ask the Tanzanians to be patient, we understand that the people have desire. Concerning this troop to fail to catch the suspects in the events of the first explosions which happened in Arusha, Jumapili said even the lawbreakers and who are there behind these events are wise and have intelligence of avoiding the hand of the state.

Comments:
GT apparently does not have means for identifying multiword expressions. The string *wavute subira* is a form of the multiword expression *vuta* ‘pull’ *subira* ‘patience’. Put together they mean 'be patient'. SALAMA has isolated it and translates it correctly.

GT does not translate the string *jeshi hilo* correctly, as a troop attempting to arrest the suspects. Instead it interprets that soldiers are accused. The proper name *Jumapili* is not recognised as a person name but translated as 'Sunday'. The string *wana akili* is translated as 'intelligent children' instead of 'have intelligence'. GT translated *dola* as 'dollar', although it means here 'state'.

Example (2) demonstrates problems encountered in translating complex verb forms.

(2) Mwenyekiti wa Bunge, Samuel Sitta alisema tayari amekwishawasiliana na Rais Kikwete kuhusu vikao vya Bunge analoliongoza.

GT
Chairman of the National Assembly, Samuel Sitta said already kwishawasiliana and President Kikwete about the sessions he loliongoza.

SALAMA
The chairman of Parliament, Samuel Sitta said already he has communicated with President Kikwete concerning the sessions of the Parliament which he leads.

Comments:
The sentence has two common verbs, which GT fails to translate, obviously because of their complicated structure. It seems that in the current development phase GT tries to figure out the subject prefix and some tenses of the verb only. However, here it fails to identify the extended tense marker *mekwisha*. If the verb has other prefixes, such as relative and object prefix, the system fails. The verbs *wasiliana* and *ongoza* are frequent verbs and they should be recognized by the system. The word *Bunge* disappears mysteriously.

Concordance and word order are important features of Swahili. In translating from Swahili to English, concordance is not a problem, but word order of long noun phrases might be (3)

(3) Vitabu vyangu vizuri vile vitatu vilivyowapendeza wanafunzi vimekutwa baada ya kutafutwa.

GT
My books vilivyowapendeza well as three students vimekutwa after searchable.

SALAMA
Those my three good books which pleased the students have been found after searching.

Comments:
Here is a test sentence to see how the two systems translate long noun phrases. All words belong to the core vocabulary. However, the verbs *vilivyowapendeza* and *vimekutwa* are unknown to GT. The words *vitabu vyangu vizuri vile vitatu* belong to the same noun phrase meaning ‘those my three good books’, but GT recognizes only first two of them and then messes up the rest. Also the string *baada ya kutafutwa* is interpreted in a strange way. There are two overlapping multiword expression candidates, *baada ya* (after) and *ya kutafutwa* (searchable). GT chooses the latter one and translates it as ‘searchable’, but by so doing it loses the latter part of *baada ya*. Yet GT translates it happily as ‘after’, although *baada* without a referent cannot have a sensible meaning.

One of the most difficult problems in machine translation is how to handle multiword expressions so that correct translation can be generated. There are several types of multiword expressions. The example in (4) contains perhaps the most difficult case to implement. The

MWE has four components, two of its words inflect, and the structure is arbitrarily discontinuous. This means that the head word of the MWE *nyumba* is detached from the rest *za kulala wageni*. Yet the structure has the meaning ‘guest house’ that inflects in singular and plural.

(4)
nyumba zangu nzuri na ghali hizo tatu za kulala wageni

GT
my beautiful and expensive houses three guest

SALAMA
these my three good and expensive guest houses

Comments:
The example demonstrates the ability of the systems to control word order as well as to isolate discontinuous multiword expressions. *Nymba zangu nzuri na ghali hizo tatu* would be a noun phrase in itself. However, in this case also the words *za kulala wageni* are part of the noun phrase, because the words *nyumba* and *za kulala wageni* together constitute a multiword expression meaning ‘guest houses’. GT loses the words *hizo* and *za kulala*. The word ‘three’ is in the wrong place, and if *wageni* is translated as ‘guest’, it should be in plural. As a whole, the translation does not make sense. SALAMA masters even cases of complicated word order. Particularly noteworthy is that it also handles correctly discontinuous multiword expressions, even such ones, where parts of structure have an unknown distance from each other.

Examples from English to Swahili translation

In this section we test how Google Translate and SALAMA translate some complex word and sentence structures. We pay attention particularly to word order, concordance, and correct word formation in the target language. The examples below were extracted from Kenyan and Tanzanian English newspapers, and from the Internet.

(5) Post-editing, or the editing done to improve machine-translated content to a publishable quality, has long been part of the translation repertoire in one form or another.

GT
Post-editing, au editing kufanyika ili kuboresha mashine-kutafsiriwa maudhui ya publishable ubora, kwa muda mrefu imekuwa sehemu ya tafsiri ya Répertoire katika namna moja au nyingine.

SALAMA
Uhariri wa kufanyika tena, au uhariri kuboresha maudhui iliyotafsiriwa kwa mashine kwa ubora tayari kwa kuchapisha, kwa muda mrefu umekuwa sehemu ya mkusanyiko wa maonyesho wa tafasiri katika umbo moja au jingine.

Comments:
GT does not recognize words 'post-editing' or 'editing'. The word 'done' is translated with stative infinitive form of the verb *fanya* ‘do’. In the word 'machine-translated' the latter part is

not recognized. The word ‘publishable’ is also unknown. In the verb *imekuwa* there is the wrong subject marker 'i'. It should be 'u' to refer to the subject *uhariri* of the noun class 11. The word ‘repertoire’ is strangely translated as *Répertoire*!

(6) This now commonplace tool has brought with it gains in productivity, more efficient resource management, and incredible value in research and development of MT itself - popular data-driven methods like Google Translate are largely reliant on human translations!

GT
Hii chombo sasa ni kawaida umeleta faida katika tija, usimamizi wa rasilimali na ufanisi zaidi, na thamani ya ajabu katika utafiti na maendeleo ya MT yenyewe - mbinu maarufu data inayotokana kama Google Tafsiri kwa kiasi kikubwa ni kujitegemea juu ya tafsiri ya binadamu!

SALAMA
Hii ala ya kawaida sasa imeleta nayo manufaa katika tija, menejimenti fanisi zaidi ya viingizia, na thamani isiyoaminika katika utafiti na maendeleo ya MT yenyewe - mbinu zenye msingi katika data zinazopendwa kama Google Translate ni zenye kutegemea kwa kiasi kikubwa tafsiri za kibinadamu!

Comments:
It is impossible to know where the subject prefix 'u' comes from in *umeleta*, because neither *chombo* nor *kawaida* belong to noun classes that agree with 'u'. Also the translation of the first part is wrong and the words 'with it' are without translation. The sequence ‘more efficient resource management’ is translated as *usimamizi wa rasilmali na ufanisi zaidi* , although it should be *menejimenti fanisi zaidi ya viingizia*. The adjective ‘data-driven’ is translated as *data inayotokana*, although the more correct translation would be *inayotokana na data*. Also the adjective ‘reliant’, translated as *ni kujitegemea*, is not grammatically correct, and the verb should not have the reflexive prefix *–ji-*. The string 'human translations' is plural. ‘Google Translate’ is a named entity and should not be translated.

In the example in (7) we test how the systems handle long noun phrases and concordance.

(7) Those my three good and expensive books that pleased students have been found.

GT
Wale yangu vitabu vitatu nzuri na gharama kubwa kuwa radhi wanafunzi zimepatikana.

SALAMA
Vitabu vyangu vizuri na ghali vile vitatu ambavyo viliwapendeza wanafunzi vimepatikana .

Comments:
This example demonstrates how the systems handle long noun phrases. The words ‘those my three good and expensive books’ constitute a noun phrase. GT gives a translation for each word, but it is not able to control the word order or concordance. The failure is understandable, because GT does not control word order or concordance with grammatical rules. On the other hand, SALAMA has translated the sentence correctly. For a rule-based system, even long noun phrases are not a problem, because the concordance rules as well as word order rules control the translation.

(8) The Nation established that the Cord team had received crucial information from Safaricom which lawyers were using to analyse the results released by the IEBC.

GT
Taifa imara kwamba Cord timu walipokea habari muhimu kutoka Safaricom ambayo wanasheria walikuwa kutumia kuchambua matokeo iliyotolewa na IEBC.

SALAMA
The Nation lilihakikisha kwamba timu ya Cord ilikuwa imepokea taarifa nyeti kutoka Safaricom wanasheria gani wakikuwa wakitumia kuainisha matokeo yaliyotolewa na IEBC.

Comments:
The name of the newspaper is 'The Nation' and it should not be translated. The verb 'established' is translated with the adjective *imara*. The past perfective form 'had received' is translated with past tense *walipokea*, but with the subject prefix of the wrong class. The correct translation is *ilikuwa* imepokea. Again, 'were using' is translated so that the first verb is inflected correctly, but the second one is in infinitive *kutumia*. It should be *walikuwa wakitumia.* Again, GT does not get the relative structure correctly. There is an attempt to form the relative verb structure *iliyotolewa*, but the concordance is wrong. SALAMA misinterpreted the word ‘which’ as an interrogative pronoun and translated it as *gani*.

Discussion

The examples above demonstrate the types of problems, which a SMT has in translating between languages such as Swahili and English. The same text translated with the RBMT system shows the main differences in performance between these two systems. Below we sum up the findings of the translation tests.

Assessment of Swahili to English translation

Many kinds of problems can be found in translating from Swahili to English. Here we discuss some of them.

Identification of word forms

There occur frequently such word forms in Swahili which GT does not recognize. The reason is probably not the small number of the so called 'words' included into the system. Many such lexical words that are not recognized are probably listed in the system. Problems come from the vast number of different forms that the word may take. Particularly Swahili verbs are a nightmare for a statistically operating translation system, because each Swahili verb may have millions of surface forms, at least in theory. Also in practice, some commonly occurring verbs were found to have more that 2000 different forms each in a small corpus of 2 million words.

SALAMA is based on the analysis of each word form. While it is using finite-state methods, it is possible to describe even very complicated word structures, such as Swahili verbs. Therefore, if SALAMA encounters an unknown word, it is probably a typo or a word of a different language. Such unknown words do not affect the basic translation process.

Handling multiword expressions (MWE)

GT does not have proper means for handling MWEs. Names of companies, government agencies and titles seem to have been handled to a large extent in GT. However, these are the simplest types of MWEs, because they are 'frozen'. They do not inflect and they do not allow other words within the word cluster. The majority of Swahili MWEs inflect and some of them allow other words within the MWE word cluster.

It is also very important to note that a word cluster that is a MWE in one context is not necessarily that in another context. Therefore, MWEs must be defined in an environment, where text context can be made use of for concluding whether the word cluster is a MWE or not in that particular context.

Furthermore, there must be a mechanism for identifying the members of the MWE in non-continuous MWEs for producing the correct translation, including the correct word order.

SALAMA has addressed the problems of treating the MWEs. They are handled in the environment, where rules can be written for constraining the application of rules. Therefore, each type of constraint can be handled. Such constraints include the context in sentence, the need to inflect the MWE cluster, the need to allow other words within the MWE, and the need to (re)define the part of speech of the MWE. The correct definition is important in producing correct surface text.

Handling complex noun phrases

The performance of the two translation systems in handling noun phrases was tested in examples (3), (4) and (7) above. The example sentences were constructed according to grammatical rules. GT apparently knows that a numeral follows the noun, and that adjective follows the noun. The way how the GT system combines word pairs is incomprehensible. How is it possible that from the string *vitatu vilivyowapendeza wanafunzi* one can get a translation 'three students'? This even violates the basic rule that the noun precedes the numeral.

SALAMA translates long noun phrases according to grammatical rules. This is possible, because the system first makes a detailed analysis of each word form and then makes use of the tags in constructing disambiguation and mapping rules. Even complicated grammatical phrases are not a problem for SALAMA.

Handling proper names

Correct treatment of proper names in MT is very problematic. Various types of proper names require different treatment. Some proper names, such as person names, are transferred to TL as such. However, many person names have a form that could also be an ordinary word. If such a word is inside the sentence and begins with capital initial, it is likely a proper name. Example (1) demonstrates this. Yet GT interprets *Jumapili* as ordinary word and translates it as ‘on Sunday’. SALAMA identifies it as a person name. If such a word begins the sentence, the problem is even bigger, because all sentence initial words are capital initial.

The only safe solution for handling ambiguous proper names is to give to such words two interpretation, one for ordinary word and one for proper name. The selection is then made using context sensitive rules, which sometimes are very complex.

Such proper names that are multiword expressions are easier to handle, because they do not get easily mixed with ordinary words. Names of ministries and organizations are examples of easy cases.

Comparison of GT and SALAMA in Swahili to English MT

The methods used by these systems to translate are very different. Because SALAMA is based on language analysis, it does not normally encounter unknown words, except new proper names and words of other languages. SALAMA analyses and translates the most complicated verb structures, including verb compounds.

GT produces sometimes excellent translation. This is the case especially when the source text is close to what was used in training the system. Then sometimes it messes up the text completely, leaves out words, changes the part-of-speech category etc.

SALAMA is weak in producing correct articles in English, because Swahili does not use articles. Only the presence or absence of the definite article is implemented. The indefinite article is not produced in the current system. Due to its approach, GT is sometimes able to produce also a correct indefinite article.

Assessment of English to Swahili translation

Coverage of vocabulary

It could be expected that because the SL English is an isolating language, GT would have equivalents for common words. Example (5) above has four unknown English words. SALAMA has identified all the words, because it has an extensive analysis system of English and a large dictionary for mapping the Swahili equivalents.

Complex verb forms

The correct production of Swahili verb forms is a nightmare for GT. Some common and simple verb forms are produced correctly, but it does not even make an attempt to translate forms with relative and object prefixes. SALAMA translates even the most complicated verb structures. This is possible, because the system first produces the appropriate tags and then converts them to surface form. An example of a compound verb is in (8) above. GT translates the words ‘the Cord team had received’ as *Cord timu walipokea*. SALAMA translates it *timu ya Cord ilikuwa imepokea*. The form in English is past perfective, and it should be translated using the auxiliary verb *kuwa*.

Concordance

Perhaps the worst nightmare for statistical MT is the production of concordance in noun phrases and verb phrases. A revealing example is (7). It has seven words that have to agree with *vitabu* 'books'. It is a noun of class 8. In GT translation, none of the seven words has the correct agreement marker. The last words ‘have been found’ was translated otherwise correctly, but the subject prefix concordance was wrong. This example suggests that GT has no method for controlling class agreement. In verbs, the default seems to be class 9 for singular and class 10 for plural, if the subject is not animate.

Mapping of words and word order

English and Swahili have very different word orders. It is also common that a word is mapped to a cluster of words, and a cluster of words is mapped to a single word. Furthermore, Swahili uses verb prefixes, such as subject, relative and object prefixes, to mark expressions that are represented by separate words in English. All this complicates translation. SALAMA has a system for handling these phenomena on the basis of grammar. GT seems to have problems with MWEs, verb prefixes, and with word order in noun and verb phrases.

Problems in detecting translation errors

Developers of SMT systems are well aware that it is often very difficult to know why translation failed. Therefore, research has been done for finding better ways for finding the sources of failure (Wisniewski et al 2013). The solution often used is to accumulate more parallel texts in the belief that when there is more material, the likelihood of finding correct matches improves. Larger corpora also produce better statistics of alternative translations.

In working with RBMT systems, especially if the structure of the system is modular, it is always possible to detect the source of translation error. To know what is the best way to correct the error requires deep knowledge of the system and experience in trying various methods. At which point in processing sequence should the correction be made? How general should the new rule be so that it has a maximum effect without causing wrong translation elsewhere? Such considerations are necessary for optimizing the translation system.

Applications derivable from the rule-based approach

In addition to MT, the rule-based approach makes it possible to develop a number of other high quality applications. Below we discuss some of them.

Spelling checkers and linguistic taggers

A comprehensive and accurate morphological tagger alone, without disambiguation, is an excellent resource for such applications as spelling checkers and dictionary search. Spelling checkers help to identify typing errors and are therefore useful for anybody who types text. The morphological analyzer can be used for helping in the use of electronic dictionaries, because the analyzer finds the base form of inflected words. The dictionary user may type any form of the word, and the system leads the user to the dictionary entry of the base form. We will discuss more on dictionaries below in the section ‘Dictionary compilation’.

If the morphological analyzer is enhanced with a disambiguating module, this tool can be used for tagging text corpora. The so-called POS-tagged corpora are produced using such tools. In addition to this basic POS-tagging, many kinds of features can be added to the analyzer. These include syntactic mapping, morphological and syntactic features, and even lexical glosses in another language. For example, Helsinki Corpus of Swahili 2.0[3] with 25 million words was tagged using such an extensive tagger.

Dictionary compilation

If the language analysis system is carefully constructed so that it has high quality components, it is possible to use the system for converting a corpus into dictionary (Hurskainen 2008). Not only must the analysis system have high recall and precision scores, but also the disambiguation system must have high quality, so that ambiguity can be resolved reliably. This is particularly important for finding the correct examples of use in text, as well as for producing correct frequency counts. The isolation of multiword expressions adds to the value of the result, because also constructions of more than one word can be searched.

If we consider such languages that have no proper dictionary but which have a sizeable amount of written texts, these texts could be collected as a corpus and made use of for compiling a dictionary.

[3] http://urn.fi/urn:nbn:fi:lb-2014032624

Advantages of the system include: (a) the word to be searched can be typed in any inflected form, (b) no lexical words are omitted, (c) word frequencies in the corpus will be available, (d) cross-references will be produced, (e) also multiword expressions can be searched, (f) example sentences will be retrieved - together with translation if such an application is provided.

Even a large corpus contains only part of the words used in communication. The dictionary system can be enlarged by adding also such words that do not occur in the corpus. In this case the system produces only the default lexical information of the word without examples of use.

Perhaps the most fascinating feature in such modern dictionaries is that any form of the word can be entered, and the system finds the base form of the word and all information attached to it. This requires that a morphological analyzer first processes the entered word for finding out the base form and possibly other morphological information, which is then used in searching matches.

If the dictionary is initially compiled from a Swahili corpus, the result is a Swahili-English dictionary. Such a dictionary can, with little effort, be converted into an English-Swahili dictionary. When the user enters an English word, the system produces all glosses of that word in Swahili, together with use examples on the basis of Swahili equivalents. In this application, however, the word must be typed in base form. Inflected forms cannot be used, because the system does not include an analyzer of English.

Dictionary testing

With the help of a comprehensive morphological analyzer it is possible to test the quality of dictionaries (Hurskainen 2002). The existing master lexicon is first reduced to match the lexical entries of the dictionary to be tested. Then this reduced lexicon is used for analyzing the text corpus. For example, if the dictionary was intended for normal text, such as used in news media, a large corpus of news texts is suitable for testing. The test result shows two things. First, it shows such words that were not recognized. This means that these words used in text were not included in the dictionary. Second, it also shows such words in the dictionary that were never used in text. In other words, the system reveals how well the dictionary covers the language used in news texts.

Intelligent language learning systems

Each language type has such learning areas that are particularly difficult to grasp. Bantu languages use a noun class system, which results in complicated agreement patterns across the sentence structure. Learning the use of all noun classes with their class markers, including exceptions, is a nontrivial task. For the learner it is often difficult to make sure that the structure is correct. Printed grammar books normally have serious gaps in advising the learner. And at least it is often very difficult to find the place where the problem is discussed.

With the help of the complete language analysis system it is possible to construct many types of modules for helping the student in learning (Dickinson and Herring 2008; Hurskainen 2009, 2010). The learner can test whether the expression is grammatically correct. The system responds, if the string has a mistake, the system informs about the type of error and gives advice how to proceed. Errors can be in typing the word, in concordance, in word order etc. The learner is interactively advised until the expression is correct.

In addition to testing the structures, several types of guided lessons can be constructed. Self-tutoring learning courses can be constructed. The learner does not necessarily need a human tutor; the system functions as a tutor. The system does not have limitations in regard to vocabulary or word forms. All several millions of word forms can be used in learning.

Vocabulary compiler

The language teacher as well as the learner is often in a situation, where the translation of a new text is cumbersome. Dictionaries do not always help, and at least their use takes unnecessarily much time. By using a sophisticated language analyzer it is possible to compile vocabularies for any text. The vocabulary list is produced so that first the whole text is analyzed. Then the most frequent words are removed and only less frequent words are retained. The length of the vocabulary list can be tuned according to the level of the learner. A beginner needs longer lists than an advanced learner.

Bootstrapping the rule-based approach to other languages

Because under-resourced languages are in constant shortage of financial resources for developing language technology, it would be wise to make use of existing technology. The rule-based approach does not require parallel corpora or any other types of massive data sources. The basic requirements include a dictionary and a grammar for modelling the morphological analyzer. More advanced applications require a good environment for writing rules for disambiguation and syntactic mapping, as well as for isolating MWEs.

A technology developed and thoroughly tested for one language can be applied to another language, especially if the languages have similar features. Therefore, a system developed for Swahili can be without major effort transferred to other Bantu languages.

It is often argued that rule-based language technology is expensive to develop compared with statistical approaches. It is true, because rule writing is human activity and as a task far from trivial. However, if the grammar is properly integrated into the system, it is done for future generations, and other applications can be developed on this basis.

For example, an intelligent language learning system, earlier developed for Swahili, was applied to learning Runyakitara spoken in Uganda (Katushemererwe and Hurskainen 2011). Mapping between disjoining and conjoining writing systems in Bantu languages was also developed for two Bantu languages, Kwanyama and Northern Sotho (Hurskainen and Halme 2001; Hurskainen et al 2005; 2006).

Conclusion

The chapter demonstrates the strengths and weaknesses of statistical and rule-based machine translation systems. The content of the chapter is somewhat provocative. It reminds us that if statistical methods continue to predominate, most world’s languages will be marginalised. The gap between dominant and local languages is going to increase. With the help of tests made on the performance of both kinds of translation systems, we hope to have demonstrated that rule-based translation systems are suitable for machine translation of highly inflectional languages. Computers offer huge potential for global human communication, and only a fraction of their possibilities has been used. However, it is dangerous to believe that mathematical calculations and algorithms alone can solve translation problems. Human knowledge of how languages behave is a huge resource. That knowledge should be fed into the translation process, just from the beginning and not only when statistical methods fail.

It is possible to develop high quality language applications, such as MT, for an under-resourced language, without extensive language resources. This suggests that no technical barrier exists for extending viable language technology to less resourced languages, including languages with complex inflection and derivation. The tests made in this paper even suggest that SMT between such languages as Bantu languages and English encounters insurmountable problems. The development of language resources needed for rule-based technology requires a covering dictionary of the language and knowledge of its grammar. Several applications can be developed on the basis of these resources.

References

Bamgbose A. (1991) *Language and the Nation. The Language Question in Sub-Saharan Africa*. Edinburgh: Edinburgh UP for the International African Institute.

Bamgbose A. (2000) *Language and exclusion: The Consequences of Language Policies in Africa.*

Barreiro A., Scott B., Kasper W. and Kiefer B. (2011) OpenLogos machine translation: philosophy, model, resources and customization. *Machine Translation*, 25(2), 107-126.

Birch A., Osborne M., Blunsom P. (2010) Metrics for MT evaluation: evaluating reordering. *Machine Translation*, 24(1), 15-26.

Cho Kyunghyun, van Merrienboer Bart, Bahdanau Dzmitry and Bengio Yoshua (2014) *On the Properties of Neural Machine Translation: Encoder--Decoder Approaches*, Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.

Dickinson M. and Herring J. (2008) Russian Morphological Processing for ICALL. *The Fifth Midwest Computational Linguistics Colloquium (MCLC-5)*. East Lansing, MI

Felice M. and Specia L. (2013) Investigating the contribution of linguistic information to quality estimation. *Machine Translation,* 27(3-4), 193-212.

Habash N., Dorr B. and Monz C. (2009) Symbolic-to-statistical hybridization: extending generation-heavy machine translation. *Machine Translation* ,23(1), 23-63.

Hurskainen A. (1992) A Two-Level Computer Formalism for the Analysis of Bantu Morphology. An Application to Swahili. *Nordic Journal of African Studies*, 1(1), 87-122.

Hurskainen A. (1996) Disambiguation of morphological analysis in Bantu languages. *Proceedings of COLING-96*, pp. 568-573.

Hurskainen A. (2002) Tathmini ya Kamusi Tano ya Kiswahili (Computer Evaluation of Five Swahili Dictionaries). *Nordic Journal of African Studies,* 11(2), 283-300. ISSN 1235-4481.

Hurskainen A. (2004) Optimizing Disambiguation in Swahili. In *Proceedings of COLING-04*, The 20th International Conference on Computational Linguistics, Geneva 23-27.8. 2004. Pp. 254-260.

Hurskainen A. (2007) A rule-based environment for Swahili development. *MultiLingual,* 18(8), 53-58. ISSN 1523-0309.

Hurskainen A. (2008) SALAMA Dictionary Compiler - A Method for Corpus-Based Dictionary Compilation. *Technical Reports in Language Technology*. *Report No 2, 2008* http://www.njas.helsinki.fi/salama

Hurskainen A. (2009) Intelligent Computer-Assisted Language Learning: Implementation to Swahili. *Technical Reports in Language Technology*, Report No 3, 2009, http://www.njas.helsinki.fi/salama

Hurskainen A. (2010) Language learning system using language analysis and disambiguation. *Technical Reports in Language Technology, Report No 9, 2010* http://www.njas.helsinki.fi/salama

Hurskainen A. (2012) Quality Swahili machine translation. *MultiLingual*, 23(7), 39-42. ISSN 1523-0309.

Hurskainen A. and Halme R. (2001) Mapping between disjoining and conjoining writing systems in Bantu languages: Implementation on Kwanyama. *Nordic Journal of African Studies,* 10(3), 399-414. ISSN 1235-4481.

Hurskainen A., Louwrens L. and Poulos G. (2005) Computational Description of Verbs in Disjoining Writing Systems. *Nordic Journal of African Studies*, 14(4), 438-451. ISSN 1235-4481.

Hurskainen A., Louwrens L. and Poulos G. (2006) Describing Verbs in Disjoining Writing Systems. In Finite-State Methods in Natural Language Processing. *Proceedings of the workshop on Finite State Methods and Natural Language Processing*, held Sept.1- 2 2005 in Helsinki. A. Yli-Jyrä, L. Karttunen, and J. Karhumäki (Eds) FSMNLP 2005, LNAI 4002, pp. 292-294. Springer Verlag Berlin Heidelberg. ISBN-10 3-540-35467-0, ISBN-13 978-3-540-35467-3.

Järvinen T. and Tapanainen P. (1997) A Dependency Parser for English. *Technical Reports*, No. TR-1. Department of General Linguistics, University of Helsinki.

Karlsson F. (1995) Designing a parser for unrestricted text. Karlsson F. et al (Eds), *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text,* 1-40. Berlin: Mouton de Gryuter.

Katushemererwe F. and Hurskainen A. (2011) Intelligent Computer Assisted Language Learning System: Implementation on Runyakitara. In M. Kizza (ed.) Vol. VII, *Special Topics in Computing and ICT Research.* Fountain Publishers, Kampala, Uganda.

Koehn P. (2010) *Statistical Machine Translation.* Cambridge: Cambridge University Press.

Koskenniemi K. (1983) *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Department of General Linguistics. University of Helsinki. Publication No. 11.

Labaka G, España-Bonet C., Màrquez L. and Sarasola K. (2014) A hybrid machine translation architecture guided by syntax. *Machine Translation*, 28(2), 99-125.

Luong Minh-Thang, Ilya Sutskever, Quoc V. Le, Oriol Vinyals and Wojciech Zaremba (2015) *Addressing the Rare Word Problem in Neural Machine Translation*. arXiv:1410.8206 [cs.CL].

Myers-Scotton C. (1990) Elite Closure as Boundary Maintenance: The Case of Africa. *International Journal of the Sociology of Language* 103, 149-163.

Nielssen S. and Ney H. (2004) Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, 30(2), 181-204.

Nakamura S. (2009) Overcoming the Language Barrier with Speech Translation Technology. *Science & Technology Trends - Quarterly Review No.31*.

Och F. (2005). *Statistical Machine Translation: Foundations and Recent Advances.* Google. Retrieved December 1, 2016.

Och F. (2006) *Statistical machine translation live.* Google Research Blog, April 28, 2006.

Scott B. and Barreiro A. (2009) "OpenLogos MT and the SAL representation language". In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation* / Edited by Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Francis M. Tyers. Alicante, Spain: Universidad de Alicante. Departamento de Lenguajes y Sistemas Informáticos. 2–3 November 2009, pp. 19–26.

Tapanainen P. (1996) *The Constraint Grammar Parser CG-2*. Publications No. 27. Department of General Linguistics, University of Helsinki.

Turovsky B. (2016). Found in translation: More accurate, fluent sentences in Google Translate. *The Keyword Google Blog. Google*. Retrieved March 23, 2017

Wisniewski G., Singh A.K. and Yvon F. (2013) Quality estimation for machine translation: some lessons learned. *Machine Translation,* 27(3-4), 213-238.

Zbib R., Kayser M., Matsoukas S., Makhoul J., Nader H., Soliman H. and Safadi R. (2012) Methods for integrating rule-based and statistical systems for Arabic to English machine translation. *Machine Translation,* 26(1-2), 67-83.