Swahili is a Bantu language widely used in Eastern and Central Africa, spoken by over 100 million people in countries including Kenya, Uganda, Tanzania, Democratic Republic of Congo, Zambia, Mozambique, Malawi, Rwanda and Burundi, Somalia, and the Comoro Islands. To a limited degree, Swahili mirrors some conventions found in English. For example, adverbs often follow the verb they modify. Generally, Swahili is a subject-verb-object (SVO) language, also like English. Despite these similarities to English, Swahili has core characteristics that are radically different from English. We focus on the following main characteristics of Swahili:

1. Agglutinating Nature of Swahili Verbs

A single Swahili word can easily be a complete sentence. This one word sentence will contain the subject, the tense of the action, the verb, the object (all contained in the prefix) and the suffix usually gives us the manner in which the subject and object(s) related via the verb/action. This *agglutinating* nature of Swahili is key to its complexity. This proved to be a major challenge in our machine translation effort.

For example, consider the sentence:

Walituchagulia

Wa li tu chagu lia

Wa=They li=⟨ past tense ⟩ tu=us/object of verb chagu=to choose lia=for us/mode

The regular expression below shows all the possible combinations of subject, tense and object:

$$(ni|u|a|tu|m|wa|i|li|ya|ki|vi|zi|ku|pa)(me|li|ta|na)?(ni|tu|ku|mu|wa|cho)?$$

2. Adjectives, pronouns, and determiners usually come after the nouns they modify

For example, the direct translation of "Alinunua magari mawili" is "He bought cars two."

3. Long range word dependencies can modify meaning of a complex verb

Swahili sentences sometimes have long-range dependencies, in that an earlier verb can modify the meaning of the current verb. This was exactly one of the problems that we had in trying to provide a smooth translation, since this dependency would mean further modification of the output from the Stemmer.
For example: "Watakuwa wamewaza tutakapofika." The whole sentence means, "They will have thought about it by the time we arrive". In trying to deconstruct the token "wamewaza", our stemmer (the component which *de-agglutinates* complex verb forms) returns, "they have thought" which is faithful and correct, but "watakuwa" refers to the future, which means that the context and meaning would modify our translation of "wamewaza". This is a major challenge in Swahili translation.

4. Swahili has 3 types of sentences

The first type is the simple sentence comprising a single clause. The 2nd, a complex sentence that has a main clause and one or more subordinate clauses (which usually follow the main clause) and a compound sentence that has 2 or more main clauses held together by a coordinating conjunction

5. Implicit determinants

The Swahili verb kind of incorporates the notion that there is an action, and this action is directed toward the object that comes after the verb in the sentence. e.g. [the] above.

1 Corpus

The corpus was gathered from the BBC Swahili website http://www.bbc.co.uk/swahili.

Dev Sentences

1. Baadhi ya waandamanaji wanaoipinga serikali walirusha mawe na mabomu ya petroli kuwashambulia polisi waliokuwa wanawazuia kuandamana hadi makao ya bunge.
2. Wachimbaji migodi hao 11 walitumia ngazi iliyotupwa ndani ya mugodi huo kujinasua siku ya jumatatu.
3. Mwongoza filamu wa Uingereza,Steve McQueen alipata tuzo ya filamu bora ambayo 12 Years a Slave ilipata ushindi
4. Katika taarifa kwa vyombo vya habari shirika hilo linasema kuwa abiria wote 200 wako salama .
5. Kamati hiyo, itajumuisha watu kutoka katika sekta mbali mbali ikiwemo ya habari na utumbuizaji.
6. Majeshi ya serikali, yamepambana vikali na waasi hao katika maeneo tofauti ya mji huo.
7. Ni wakati mgumu kwa Bitcoin ambayo inaarifiwa imeshambuliwa na wahuni
8. Kocha wa Man City Manuel Pellegrini amesema Man City itatamba dhidi ya Man U katika CL
9. Mwaka 1 baada ya Oscar Pistorius kumuua mpenzi wake Reeva Steenkamp, amezungumzia majuto na masikiti yake kuhusu kifo cha Steenkamp
10. Mwanawe Rais wa Afrika Kusini Jacob Zuma, Duduzane Zuma huenda akashitakiwa kwa kuendesha vibaya gari na kumuua mwanamke mmoja

Test Sentences

1. Ili kuingia robo fainali itahithitajika kuichapa Barcelona mabao matatu kwa bila
2. Filamu ya Great Gatsby ilishinda tuzo mbili, ya utengenezaji mzuri wa filamu na ubunifu wa mavazi
3. Kongamano hilo lilidhamiria kutafuta mbinu ya kulinda wanyamapori walio katika hatari ya kuangamia
4. Operesheni ya kuokoa wachimba madini waliokwama kwenye shimo la mgodi yasitishwa
5. Hayati Nelson Mandela aliacha mali yenye thamani ya dola milioni nne ambayo alisema igawanywe kati ya familia, shule na chama cha ANC

2 Our Translation

Dev Sentences

1. some of protesters who are opposing government they did throw stones and petrol bombs to them attack police who were they is prevent protesting towards parliament building
2. excavators mines those 11 they did used ladder abandoned inside of mine that save oneself from Monday day
3. director film of UK Steve McQueen he did get film award best which 12 Years a Slave it did get victory
4. in news for vessels news organization that it is sema that passengers ali 200 ako safe
5. committee that will involve people from in sector various various including of news and entertainment
6. government armies dealt with thoroughly and rebels those in regions different of city that
7. it is season hard for Bitcoin which informed imeshambuliwa and criminals
8. coach of Man City Manuel Pellegrini he has sema Man City it will shine against of Man U in CL
9. one year 1 after of Oscar Pistorius to him kill lover her Reeva Steenkamp he has talk regrets and sorrow his about death of Steenkamp
10. child of president of Africa South Jacob Zuma Duduzane Zuma will possibly to be charged for to drive dangerously car and to him kill woman one

Test Sentences

1. in order to to enter quarter finals it is required to beat Barcelona bao three for bila
2. Great film Gatsby ilishinda award two of production good of film and creativity of costumes
3. forum that it did intend look for means of protect wild animals who have in extinction danger
4. kuokoa operation wachimba minerals waliokwama in the shimo la mgodi temporarily suspended
5. late Nelson Mandela he did leave property which has dollar worth million four which he did sema divided family kati school and party of ANC

3 Translation Strategies

1. Language Model Scoring of Candidate sentences

   Our translation approach generated multiple candidate English sentences for each Swahili sentence. We scored each candidate sentence using an English language model and picked the sentence with the minimum negative log likelihood. The language model incorporated trigrams, bigrams, and unigrams with backoff and laplacian smoothing for unigrams unknown to the model. We trained the model on the Brown corpus.

   This strategy affects the selected translation for every sentence, when any word in the sentence is found to have multiple possible translations.

2. Swahlili verb stemmer

We implemented a stemmer that parses complex (*agglutinated*) verb forms and provides candidate English translations for inclusion in the candidate sentence. As described previously, this process of parsing agglutinated words for their constituent parts is central to any effective translation from Swahili. Our implementation is very limited due to the complexity of the problem, and the quality of our translations is affected.

This strategy affects the selected translation for every sentence, because virtually every verb instance is agglutinated and needs to be parsed.

3. Word Reordering (Noun-Adjective to Adjective-Noun)

   Using straightforward regular expression matching of POS tagged sentences, we reorder instances of nouns followed by adjectives.

4. Phrase Rewriting (ya)

   A common Swahili pattern is Noun-Preposition-Noun, where the preposition is specifically the word *ya*. An example of the direct translation from Swahili is *bombs of petrol*, and this rule rewrites the phrase as *petrol bombs*.

4 Error Analysis

Swahili is considered a hard language to translate, as noted on a lecture slide. We didn't fully appreciate the difficulty until we started building the stemmer to parse agglutinated verb forms. As we discussed the properties of Swahili, we mentioned the importance of verbs and how these verbs pack so much information. For this reason, there is an absolute need for a Stemmer.

1. Stemming

   For a few of the word forms, we had good translations to English, but for some of them, the overall translation was in the wrong tense, and this requires a good tool to change verb tense. As mentioned above, the long range dependencies of verbs were a part of this problem, and this means that the Stemmer we tried to build was not able to generalize well to all these corner cases in order to return perfect phrases all the time.

   For example, in considering the verb "wanawazuia" (to prevent) a correct stemmer has to distinguish between "zui"+"a", "zu"+"ia". We over-simplify, and interpret this only as the latter case. We strip the trailing "ia", but the remaining "zu" is a very different verb (to fabricate).

   We are not sure if there is a correct deterministic to the problems posed by agglutination, but we expect that a rich parallel corpus and phrase alignment would greatly improve the quality of translation.

   Our system naively translates adjacent words, leading to incorrect English e.g. "to to". "Ili" directly translates to "in order to" and "kuingia" translates to "to enter". Our stemmer doesn't help with this case because "ku" in "kuingia" would isolate the stem "ingia" as "to enter". To improve this, a much more sophisticated phrase alignment strategy would be needed.

2. Phrases in Swahili

   Our system does not recognize two-word Swahili phrases "baadhi ya", "mbali mbali" that should be translated to a single English word. Our system would need some sort of language model / phrase alignment capability to determine the preferred translation.

   The way forward on this is to store the known phrases in Swahili, use bigram features (phrases are two words in most cases), obtain the translation for this token and return result.

3. Identifying correct verbs

   If the stemmer mis-parses an agglutinated form, it chooses the wrong verb. Of course in these cases, the resulting translation is particularly bad.

4. Failure to add correct determiner

   Explicit determiners are uncommon in Swahili. Our system does not detect instances that require insertion of determiners like "the" e.g. "the forum" or "the film Great Gatsby". A phrase alignment model would identify these cases.

5. Named Entity Resolution

   Our system has no NER, so it makes the mistake of thinking "Great Gatsby" is two words instead of a single compound noun. Then we naively apply the rule that reorders NOUNS followed by ADJECTIVES. We need a Named Entity Recognition capability to know that "Great Gatsby" is a single noun. Then, we wouldn't swap "film" and "Great".

5 Google Translation

Dev Sentences

1. Some protesters who oppose the government they threw stones and petrol bombs attacked police who prevented them from marching to the home of parliament
2. 11 Miners those used in mugodi iliyotupwa level the ride out on Monday
3. Led British film, Steve McQueen received the award for the best film which won 12 Years a Slave
4. In a statement to the media that the organization says all 200 passengers are safe.
5. Committee, will include people from various sectors including information and concerts.
6. Government forces, yamepambana strong and rebels in different parts of the city.
7. Bitcoin is a difficult time for which inaarifiwa been attacked by thugs
8. Man City coach Manuel Pellegrini said itatamba Man City against Man U in CL
9. 1 year after killing her boyfriend Oscar Pistorius Kireeva Steenkamp, and mosques raised his regret about the death of Steenkamp
10. Son, South African President Jacob Zuma, Duduzane Zuma goes and charged with driving the wrong car and killed one woman

Test Sentences

1. To enter the quarter-finals itahithitajika three goals to beat Barcelona without
2. Great Gatsby film won two awards, the fine craftsmanship and innovative film costumes
3. The symposium lilidhamiria find out how to protect vulnerable wildlife species,
4. Operation to save miners were stuck in the mine pit yasitishwa
5. The late Nelson Mandela had left property worth four million dollars which is said to be scattered among family, school and ANC

6 Comparative Analysis

Google	To enter the quarter-finals itahithitajika three goals to beat Barcelona without
Our System	In order to to enter quarter finals it is required to beat Barcelona bao three for bila

Google fails to translate "itahitajika", which could be a result of lack of parsing, or using a dictionary that does not contain the word. The translation of the word "bila" not faithful is roughly correct, because "bila" could mean without or nil. Therefore, this is a context problem. In this case, nil is correct.

Our system apparently lacks the word "bila" in our lexicon hence we pass the Swahili through untranslated.

Google	Great Gatsby film won two awards, the fine craftsmanship and innovative film costumes
Our System	Great film Gatsby ilishinda award two of production good of film and creativity of costumes

Our system was unable to identify that "Great Gatsby" is a compound noun representing a film.

Our stemmer was unable to identify and parse the verb "ilishinda". This is another symptom of the limitations of our stemming. Google recognizes this word form correctly. Google also correctly adds the determiner.

Google	The symposium lilidhamiria find out how to protect vulnerable wildlife species
Our System	Forum that it did intend look for means of protect wild animals who have in extinction danger

Google fails to correctly interpret the core verb phrase, "to intend" as a verb. Our system identifies the correct form but does not get the word order correct. As we discussed previously, Google is superior to our system at adding determiners "The symposium".

Google	Operation to save miners were stuck in the mine pit yasitishwa
Our System	kuokoa operation wachimba minerals waliokwama in the shimo la mgodi temporarily suspended

Google translate is unable to translate "yasitishwa". Our system fails to identify the core verb, which results from the same limitations of our stemmer as we've already discussed.

Google	The late Nelson Mandela had left property worth four million dollars which is said to be scattered among family, school and ANC
Our System	Late Nelson Mandela he did leave property which has dollar worth million four which he did sema divided family kati school and party of ANC

Google does a faithful translation, but fails to convey the right meaning. "Igawanywe" is best interpreted to mean "divided" as opposed to scattered. We see another instance of our system failing to identify a multiword phrase "kati ya". We see a subtle new problem in this translation. Our stemming includes a pronoun in "he did leave" because that is a faithful parse of the agglutinated verb form. But given the prior instance of a proper name (Nelson Mandela) we should have known to drop the pronoun in this case. To improve this, we need both a NER system and enhanced rules for adjusting the faithful gloss of agglutinated verb forms.