ISSN 2278-3091
Volume 8, No.5, September - October 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse08852019.pdf
https://doi.org/10.30534/ijatcse/2019/08852019

English to Luganda SMT: Ganda Noun Class Prefix Segmentation for Enriched Machine Translation

Abdul Male Ssentumbwe[1], ByeongMan[2] Kim and HyunAh Lee[3]
[1]Department of Software Engineering, Kumoh National Institute of Technology, South Korea, ssentu@kumoh.ac.kr
[2]Department of Software Engineering, Kumoh National Institute of Technology, South Korea, bmkim@kumoh.ac.kr
[3]Department of Software Engineering, Kumoh National Institute of Technology, South Korea, halee@kumoh.ac.kr

ABSTRACT

Luganda or Ganda is a morphologically rich and low-resource language from Uganda. The morphological richness of Luganda sentences has an impact on the quality of translation and this work looks at improving machine translation (MT) for English to Luganda. Luganda sentence formation bases on 10 noun classes with a prefix for singular and plural. In various aspects, the interaction of these class prefixes in sentences usually enforces different words such as nouns, verbs, adverbs and adjectives to be in agreement with the subject in a given sentence. Such scenarios have resulted in various Luganda word inflectional and derivational tendencies because each noun, verb, adverb or adjective finds itself having to combine with a prefix from a class it belongs and thus creating new word forms. Using 6 statistical machine translation (SMT) models divided equally into base and morphology models, we propose a procedure to segment Luganda sentences from our English-Luganda parallel Bible corpus. Our morphological segmentation approach bases on Ganda Noun Class (GNC) prefixes and we design a tool we call GandaKIT to segment Luganda sentences at pre-processing stage and desegement them after translation at post processing stage. In experiments, we compare translation performance of SMT base models against systems trained with morphological segmentation at pre-processing stage. Our results show an improvement in MT performance over base models ranging by a difference of 1.58 BLEU points and 0.2257 NIST score for our best system.

Key words: English to Luganda statistical machine translation, Ganda noun class morphology, SMT morphological segmentation, low-resource languages.

1. INTRODUCTION

Machine Translation (MT) in general, has had tremendous growth given the current trends. This is evident in terms of; increased translation app usage, intensive research activity in the area and sprout of new MT technologies. Of these MT

technologies, Statistical Machine Translation (SMT) and its varieties, contributed greatly to concerned language translation communities [1]. In this paper, we propose English to Luganda MT. Why MT for English to Luganda pair?

Luganda or Ganda is a language spoken by over 5 million people (Baganda) from Buganda tribe in Central Uganda [2]. In Uganda, English and Swahili are the official languages for communication and, Luganda is the second widely used language in the country, notably in trade and at school for young children in the region. Ugandan natives also belong to various tribes and thus about 43 as many languages exist [3]. Swahili in Uganda is not so widely used in comparison to Luganda, though Swahili language is the main language for countries like Tanzania and Kenya in East African region. Since English is highly resourced and Luganda is fairly well documented, we believe translation technology for English to Luganda pair can contribute to promote education development for schoolchildren and other communities using both languages.

However, Luganda is a very low-resource language and like many other languages in this category, lacks computer software tools when compared to the dominant world of English resources available offline and online. In an effort to bridge some of these gaps, we thought translation from English to Luganda, will create an impact especially to the ever growing number of Uganda online users for both languages [4]. Upon this premise, we derive our English to Luganda base models against which we perform further experiments and evaluation of our proposed idea.

Luganda Original (a)	*Omuwala atambula.*			
Possible breakdown:	*O*	*muwala*	*a*	*tambula.*
English meaning:	The	girl	is	walking.
Luganda Original (b)	*Abawala batambula.*			
Possible breakdown:	*A*	*bawala*	*ba*	*tambula.*
English meaning:	The	girls	are	walking.

Luganda as a low-resource language is likely to suffer from low MT performance. These limitations in terms of experimental data normally leaves little chance for some words or phrases to be appropriately scored or translated; such a phenomena is termed as a data sparsity problem as discussed in [5]. The language itself is agglutinative, leading to rich morphological properties that raise to inflectional and derivational word tendencies. In short, our base model translations are likely to suffer from some low performance in scores. How is this a challenge? Let us consider the two

Luganda sentence examples in Figure 1 (a) and (b). The base model looks at original Luganda sentence or word forms, without morphological breakdown and thus translation result performance may be low for our English to Luganda base model. This calls for a shift away from the base model if we want to improve translation performance in some way for this pair.

2. RELATED WORKS

Researchers involved in works for high-resource and low-resource languages with such morphological richness, use or propose various techniques, some generalized and others language specific in terms of implementation. The key underlying mitigating technique is to segment these highly agglutinating terminologies into constitute parts. For example in [6] word segmentation is described for Chinese to Vietnamese language pair as part of MT pre-processing stage because characters are written with no spacing. The authors propose a word re-segmentation model that improves existing word segmentation approaches resulting in better performance for phrase-based SMT.

Other majority of approaches are generally on internal structure of words (morphology). For example [5] using factored translation models, investigated the importance of adding morphology in SMT process for translating from English to Hindi and Marathi, which are both Indian languages. The approach improved quality of their MT system for the two Indian languages in terms of adequacy and fluency. Additionally, researchers in [7] attempt to tackle the data sparsity problem in Arabic-Chinese MT by applying various tokenization means coupled with adding linguistic features (such as lemma, POS morphology in terms of affixes, stems, and clitics) to Arabic (source side) corpora. By combining these approaches, they train different factored SMT models resulting in better MT for the pair.

Some researchers for example opt to leverage machine learning based techniques; for instance [8] designed a morphological segmentation tool based on conditional random fields (CRF) in their works for English to Tigrinya SMT. They investigate morphological segmentation variants for Tigrinya a low-resource language (from Ethiopia and Eritrea) which is highly inflectional. The segmented models authors propose improved the quality of SMT for the language pair. On the other hand, similar approaches have resulted in generalized morphological segmentation toolkits such as Morfessor 2.0 and Morfessor FlatCat described in [9] and [10] respectively, and use machine learning techniques based on probabilities to morphologically segment text corpora.

It is important to note that MT works involving Luganda or Ganda language are generally scarce to find. Therefore, some previous MT works involving Luganda that were readily available at the time of this writing, is a technical report [11]. The report experimented Swahili to Luganda machine translation and it describes an MT rule-based system using Bible as their corpus. These rule-based systems depend entirely on linguistic knowledge of a language pair to translate from each language involved. Such approaches are generally costly and time consuming in nature because of the need to code many linguistic rules for both the source and target language in order to support MT experiments. In addition, the authors adopt a previous system called SALAMA a Swahili Language Manager described in [12] that was used to translate Swahili to English. Similarly, [13] describe a knowledge light approach to translate English to Luo, using a Trilingual Corpus (English-Swahili-Luo). The corpus included New Testament of the Holy Bible and other resources. Their focus was adding morphosyntactic information (e.g. factoring POS tags) into their corpus, an approach that enabled some better word alignment and thus improving results for their English to Luo, and Swahili to Luo Moses SMT systems.

In other works, [14] describe development of a SAWA Corpus used to perform translation with Moses SMT for English to Swahili. The corpus included New Testament of the Holy Bible, Holy Quran and other data resources that together were part of speech tagged, lemmatized and manually sentence aligned. From reported results, their system underperformed for English-Swahili when compared with Google translation system at the time while for Swahili-English, their system performed better.

Another low-resource study from [15] focused on extracting bilingual multi word expressions (MWE) to improve translation performance for Chinese to Mongolian SMT system while [16] using Quechua an agglutinating low-resource language, with help of morphemes, attempt to improve alignment and thus improve Quechua to Finnish SMT system.

The highlighted MT works, involving some high-resource and low-resource language pairs, all try in some way to solve problems related to data sparseness, addressing matters of highly agglutinating languages (morphologically rich) and in some cases limitations in terms of parallel corpora.

In this paper, we propose an English to Luganda SMT system with Moses Toolkit based on an English-Luganda Bible parallel corpus. We propose and investigate the impact on translation performance of our SMT models, where we enrich the target side (Luganda) with morphological segments inspired by Ganda Noun Classes (GNC). To aid this, we design a tool we call GandaKIT, responsible for injecting noun class prefix based morphological forms as part of the pre-processing stage. The tool design also handles post processing translated sentences (with segmentation) and prepare them for final evaluation.

3. OVERVIEW OF LUGANDA

Ganda or Luganda writing is based on the latin alphabet similar to that of English except that letters ‘x’ and ‘q’ are excluded, *‘h’* is rarely used while *‘ny’* and *‘ŋ’* are added [17], [18]. The language has 5 vowels *(a, e, i, o, u),* 18 consonants *(b, p, v, f, m, d, t, l, r, n, z, s, j, c, g, k, ny, ŋ)* and 2 semi-vowels

(w, y). We need to note that, *'ny'* is a consonant pronounced as one but written with two letters. The basic sentences in Luganda follow 'Subject + Verb + Object' order in general though can transform into other patterns [19]. For example, *'Omuwala agenda ku somero'* meaning 'The girl is going to school.' In this case, *Omuwala* (The girl) [**Subject**] + *agenda*(is going)[**Verb**] + *ku*(to)[**location particle**] + *somero*(school)[**Object**].

3.1 Ganda Noun Class

The noun classes in Luganda orthography guide agreement of different parts in a sentence. These noun classes are responsible for the modifications of various Luganda word forms [17], [20] in sentence. Table 1 demonstrates these classes and except for Class X only, we can observe that the rest of the classes each has a prefix for singular and plural word form [21].

The English definite article 'the' and indefinite article 'a' do not have direct equivalences in Luganda but instead *'a-'*, *'e-'*, and *'o-'* are attached to a noun class and then appended to a noun or a verb to create required transformation. For example from Table 1, Class I will become *(omuntu, omuwala, omwana, abantu, abawala, abaana)*. We can notice that for this scenario singular nouns used *'o-'* while plural forms used *'a-'* [22].

Table 1: Ganda/Luganda Noun Classes and prefixes for singular and plural forms

Class (Prefix)	Singular Form			Plural Form		
I (*mu-ba*)	***mu**-ntu* (person)	***mu**-wala* (girl)	***mw**-ana* (child)	***ba**-ntu* (people)	***ba**-wala* (girls)	***ba**-ana* (children)
II (*mu-mi*)	***mu**-sege* (wolf)	***mu**-ti* (tree)	***mw**-aka* (year)	***mi**-sege* (wolves)	***mi**-ti* (trees)	***my**-aka* (years)
III (*li-ma*)	***li**-iso* (eye)	***li**-nnyo* (tooth)	***li**-nnya (name)*	***ma**-aso* (eyes)	***ma**-nnyo* (teeth)	***Ma**-nnya* (names)
IV (*ki-bi*)	***ki**-ntu* (thing)	***ki**-tabo* (book)	***ky**-angwe* (sponge)	***bi**-ntu* (things)	***bi**-tabo* (books)	***by**-angwe* (sponges)
V (*ka-bu*)	***ka**-ti* (stick)	***ka**-tale* (market)	***ka**-timba* (net)	***bu**-ti* (sticks)	***bu**-tale* (markets)	***bu**-timba* (nets)
VI (*ku-ma*)	***ku**-gulu* (leg)	***ku**-tu* (ear)	-	***ma**-gulu* (legs)	***ma**-tu* (ears)	-
VII (*gu-ga*)	***gu**-sajja* (big man)	***gu**-solo* (beast)	***gu**-*yanja (ocean)	***ga**-sajja* (big men)	***ga**-solo* (beasts)	***ga**-*yanja (oceans)
VIII (*lu-n/m*)	***lu**-goye* (cloth)	***lu**-papula* (paper)	***lu**-yimba* (song)	***n**-goye* (clothes)	***m**-papula* (papers)	***n**-nyimba* (songs)
IX (*n/m-n/m*)	***n**-koko* (hen)	***n**-te* (cow)	***m**-bwa* (dog)	***n**-koko* (hens)	***n**-te* (cows)	***m**-bwa* (dogs)
X (*tu*)	***tu**-zzi*, (water),	***tu**-ta*, (milk),	***tw**-enge* (alcohol)	-- no singular/plural distinction -- in the sense 'precious little of'		

Irregular prefix forms *"mw", "ky"* for singular and "my", *"by"* for plural and *"tw"* from class X. Class VII normally adds a meaning of something being so big or ugly depending on context. Class VI has only 2 known nouns [17].

3.2 Transformation with adjective prefixes

GNC also affects adjectives. Let us consider the adjective *'–lungi'* (beautiful) in Luganda with some GNC from Table 1. In Table 2, we observe adjective *'-lungi*' transformed for singular and plural forms for each noun class involved. In this case, nouns transfer their GNC prefixes onto adjective *(-lungi)* but the meaning for the Luganda adjective involved stays intact. Some example sentences from Table 2 can be: *'Omuwala omulungi'* (beautiful girl), *'Abawala abalungi'* (beautiful girls) from class I, *'Omuti Omulungi'* (beautiful tree), and *'Emiti emirungi'* (beautiful trees) from class II. This kind of behaviour (modification effect of GNC) continues to take centre stage along general Luganda sentences and in view of MT; it is something we need to pay close attention to going forward for better translation performance.

Table 2: Transformation of the adjective *'-lungi'* (beautiful / good)

Class	Noun Singular	Adj. Pfx Sn.Form	Noun Plural	Adj. Pfx Pl.Form
I	***Om**uwala (girl)*	***omu**-lungi*	***Aba**wala (girls)*	***aba**-lungi*
II	***Omu**ti* (tree)	***omu**-lungi*	***Emi**ti* (trees)	***emi**-rungi*
III	***Eri**iso* (eye)	***ed**-dungi*	***Ama**aso* (eyes)	***ama**-lungi*
IV	***Eki**tabo* (book)	***eki**-rungi*	***Ebi**tabo* (books)	***ebi**-rungi*
V	***Aka**ti* (stick)	***aka**-lungi*	***Obu**ti* (sticks)	***obu**-lungi*
VI	***Oku**gulu* (leg)	***oku**-lungi*	***Ama**gulu* (legs)	***ama**-lungi*
VII	***Ogu**sajja* (big man)	***ogu**-lungi*	***Aga**sajja* (big men)	***aga**-lungi*
VIII	***Olu**goye* (cloth)	***olu**-lungi*	***En**goye* (clothes)	***en**-nungi*
IX	***En**koko* (chicken)	***en**-nungi*	***En**koko* (chickens)	***en**-nungi*
X	***Otu**ta* (precious milk)	***otu**-lungi*	-	***otu**-lungi*

Adj. - Adjective, Pfx - prefix, Sn.Form - singular form, Pl.Form - plural form. Irregular prefix forms; *'r'* replaces *'l'* after *'e' or 'i'* e.g. Class II (plural), Class III (singular) and Class IV. Other changes in spelling are Class III (singular) *'l'* in stem *'-lungi'* was replaced with *'dd'* while Class VIII and IX *'l'* was replaced with *'n'* [17].

3.3 Transformation with verb prefixes

In the case of verbs, GNC has its particular verb prefixes for both singular and plural situations. For example using the verb *'tambula'* meaning to walk. Table 3 shows some morphology transformation based on the noun's class [22].

Table 3: Transformation of the verb *'tambula'* to walk

Class (Verb Prefix)	Singular Form	Plural Form
I (a-ba)	***a**-tambula*	***ba**-tambula*
II (gu-gi)	***gu**-tambula*	***gi**-tambula*
III (li-ga)	***li**-tambula*	***ga**-tambula*
IV (ki-bi)	***ki**-tambula*	***bi**-tambula*
V (ka-bu)	***ka**-tambula*	***bu**-tambula*
VI (kuga)	***ku**-tambula*	***ga**-tambula*
VII (gu-ga)	***gu**-tambula*	***ga**-tambula*
VIII (lu-zi)	***lu**-tambula*	***zi**-tambula*
IX (e-zi)	***e**-tambula*	***zi**-tambula*
X (tu)	***tu**-tambula*	***tu**-tambula*

Class X has no distinction between singular and plural form while some plural classes have similar verb prefix forms.

Using Table 3, we can form the following sentences: *'Omuwala atambula'* (The girl is walking), *'Abawala batambula'* (The girls are walking) from class I, *'Ogusolo gutambula'* (The beast is walking), and *'Agasolo gatambula'* (The beasts are walking) from class VII. In these examples, the verb *(-tambula)* is influenced by GNC with its corresponding verb prefix.

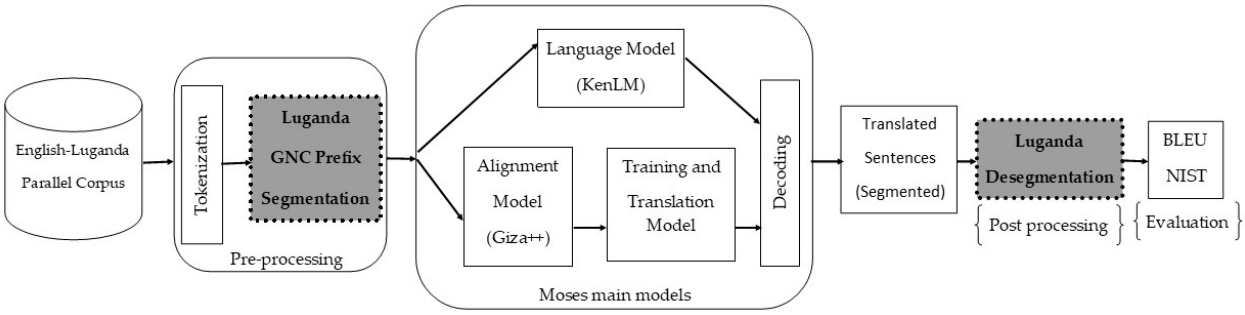

Figure 2: Approach architectural representation: GandaKIT

3.4 Transformation with negation prefixes

Depending on the case at hand, Luganda has different negation forms. Our main interest is with negation prefixes, which we attach to words. For example negating verbs in Table 3, we add *'t'* for verbs starting with vowel and *'te'* for those verbs starting with a consonant. For example: *'Omuwala tatambula'* (The girl is not walking), *'Abawala tebatambula'* (The girls are not walking) from class I, *'Ogusolo tegutambula'* (The beast is not walking), and *'Agasolo tegatambula'* (The beasts are not walking) from class VII. In these examples, we can see the influence of GNC when negating the verbs used in the sentences.

These few examples demonstrate the inherent power of GNC that forces modifications of other parts of a sentence in order to agree with it [17]. We show this using prefixes for GNC on given nouns, verbs, and adjectives for singular and plural cases including irregular prefix and stem spelling modifications. Other prefixes for first, second and third person singular and plural cases were also collected for this study since they are important in morphology formation. Luganda morphology also has suffixes and infixes that contribute to the inflectional and derivation tendencies in word formation. However, because noun class prefixes influence majority of word transformations, our proposed morphological segmentation approach is indeed too generalized on the same premise. On addition, because we consulted different sources when manually compiling these prefixes, we may not mention all of them here and there is also a possibility that we could have missed out some prefixes due to the documentation we used during compilation.

4. PROPOSED GANDAKIT APPROACH

In attempt to improve performance or quality of MT output, different researchers adopted various ways for morphology segmentation depending on the nature of languages involved. As highlighted previously, Luganda morphology formation has prefix, infix and suffix usage where GNC prefixes exercise greater power in modification of Luganda sentence components in most cases. Our approach therefore, focuses on designing a generalized Ganda morphology segmentation tool based mainly on Ganda Noun Class prefixes, and at pre-processing stage, we segment Luganda sentences and desegment translated Luganda sentences at post processing stage. The overall architecture of our GandaKIT approach is illustrated in Figure 2.

4.1 Luganda GNC prefix segmentation process

This process as in Figure 2 comes after tokenization of Luganda sentences in our dataset at pre-processing stage. Having appreciated the discussion on Ganda Noun Classes (GNC), we now lay our segmentation process:

Step (1), we identify and compile most Ganda noun prefixes, from the ten classes with guidance from known Luganda sources [17], [18]. The same prefixes cover for adjectives as in Table 2 and consideration is taken into account for cases when Luganda article prefixes *('a', 'e', 'o')* are used or not.

Step (2), identification of GNC verb prefixes for example for present, past, future and other tenses follows. Not all verb prefixes directly resemble their noun class prefixes and therefore, we continue to seek reference to Luganda documentation sources as we collect them. Verbs in Luganda sentences normally have to agree with the subject's GNC in sentence and this helps in proper verb prefix identification.

Step (3), we collect negation prefixes also for present, past, future and other tenses. These too may not necessarily resemble GNC prefixes exactly but agreement with subject in question assists in negation prefix identification.

Step (4), we lay an assumption for some words or tokens to be ignored. The word ignore list includes Luganda possessives, pronouns, adverbs, conjunctions, and prepositions. These words normally, are written standing alone in a Luganda sentence, which is similar to their English counterpart. During alignment, we assume these words can easily be aligned and there is no need to be affected by the segmentation process. If segmented in some way, we believe this may cause some mismatch in alignment leading to low MT performance. The ignore list also includes all punctuation marks.

Finally, in our generalization we merge prefixes from first to third step without repeating any similar ones. This final combined prefix list (dictionary or table) is the basis for our first level morphological segmentation upon which we base our experiments and subsequent evaluations. From this list, we determine length of the longest prefix that guides our final generalization in the segmentation process. Longer prefix if confirmed during prefix search for a word, takes priority in the segmentation process down to the one with smallest length. Segmented sentences are then passed on to the next levels of the MT process i.e. language model training, alignment model and translation model training. We now show an example of

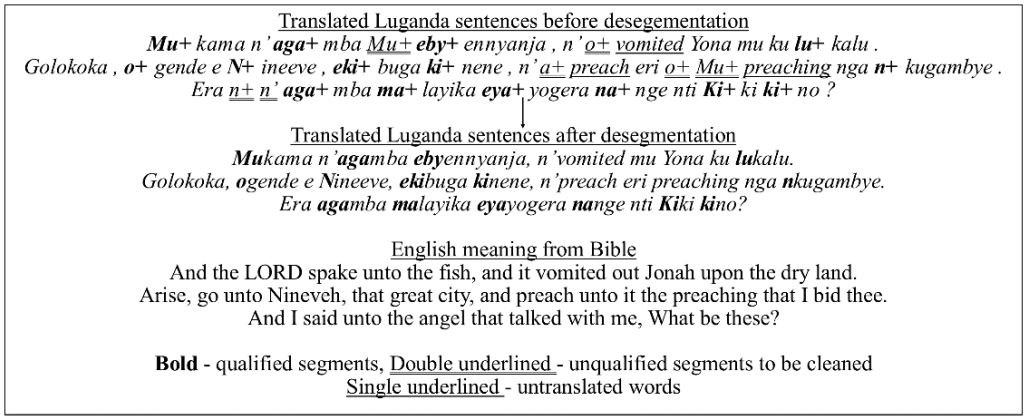

Figure 3: Examples of desegmenting translated Luganda sentences.

segmented Luganda Bible sentence before and after segmentation where segmented prefixes are denoted with a plus sign attached onto a word or token at the front (+).

Luganda (**before**): *N' abagamba nti Ndi Mwebbulaniya ; ntya Mukama , Katonda ow' omu ggulu eyakola ennyanja n' olukalu .*

Luganda (**after**): *N' **aba+** gamba nti **N+** di **Mw+** ebbulaniya ; **n+** tya **Mu+** kama , **Ka+** tonda ow' omu ggulu **eya+** kola **enn+** yanja n' **olu+** kalu .*

English (meaning): And he said unto them, I am a Hebrew; and I fear the LORD, the God of heaven, which hath made the sea and the dry land.

4.2 Luganda desegmentation process

We add this as a post processing stage after translating English sentences to Luganda (Figure 2). It is a necessary step because our translated Luganda sentences come out in morphologically segmented form and the process is supposed to revert these sentences to their normal form.

The first step in desegmentation, is to clean segments in translated sentences to leave only qualified segments. The segment cleaning process is based on what we term as generalized rules for qualified segments.

For example:

- Delete the first segment occurrence(s) to the left if two or more prefix segments are next to each other,
- Remove prefix segments next to untranslated words,
- If a prefix segment is before a character with an apostrophe, we remove them both.

Cleaning segments is therefore essential because during the SMT process, some morphological segments deposited in translated sentences are out of order and therefore unnecessary for normal Luganda word forms in a sentence. Without clean qualified segments in translated sentences, we can end up with wrongly formed words and hence leading to poor MT performance. We then reunite all qualified segments to form proper final words in translated sentences used for final evaluation. Figure 3 illustrates three examples of translated Luganda sentences before and after desegmentation stage. It shows some unqualified segments (double underlined) that the process removes before reuniting all qualified segments (in bold) with stems or root forms in a sentence. Our GandaKIT therefore comprises of both morphological segmentation and desegmentation processes.

5. LUGANDA SMT EXPERIMENTS

5.1 Corpus and domain scope

In our Luganda SMT experiments, religious material is the main domain used. This is because for most low-resource languages, normally this resource is the best parallel corpora easily available. In this work, the scope of religious material used is the Old and New Testament books of the Holy Bible (King James Version-KJV) in both English and Luganda [23]. After collection of Bible data, we thoroughly check each book within to ensure all verses are complete. We then manually merge the books at document level in three groups (Training, Development, Testing datasets) as in Table 4, and manually ensure sentences (i.e. every Bible verse was taken as a whole to be a sentence) are aligned well for every group. Detailed summary about token and vocabulary size for English and Luganda is also provided.

Finally, we merged Old and New Testament datasets in respect to training, development and test sets to form Combined dataset respectively.

Table 4: Summary of Old Testament and New Testament datasets

KVJ Bible	Training		Development		Test Set	
Old	En	Lg	En	Lg	En	Lg
Sentences	22,598		301		261	
Token Size	709,508	550,742	9,528	7,374	9,212	6,964
Vocabulary	11,446	43,147	1,302	2,029	1,193	1,843
New	En	Lg	En	Lg	En	Lg
Sentences	7,627		151		186	
Token Size	210,032	160,182	3,740	2,840	4,493	3,576
Vocabulary	6,459	17,365	730	954	973	1,196

En - English, Lg – Luganda

5.2 Experimental configurations

We installed Moses SMT on Ubuntu Server configured with 32Gb RAM, Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz

core processor with GTX 1080 Ti 10Gb GPU support.

5.3 Luganda SMT brief setup

SMT comes in various flavours and in our approach, we adopt phrase based SMT with popular MT research tool Moses [1].
Assumptions;
If we are translating English (en) to Luganda (lg) then, adopting Bayes rule [24], our translation is represented as in equation 1:

$$\text{argmax}_{lg} p(lg|en) = \text{argmax}_{lg} p(en|lg)p(\text{lg}) \quad (1)$$

Where $\text{argmax}_{lg} p(lg|en)$ is the probability maximization function, $p(en|lg)$ is the translation model and $p(\text{lg})$ is the language model.

We use Moses default language model KenLM [25], Giza++ tool for word alignment training [26], and we represent phrase based SMT with unsegmented sentences (Baseline) as → **Base**, and represent phrase based SMT with morphological segmentation at pre-processing stage as → **Morpho**.

5.4 Luganda SMT models trained

We train 6 SMT systems where the Base vs. Morpho models, using Old Testament, New Testament and Combined Testaments' datasets are prepared. For example we prepare, Base + Old Testament dataset → **(Base+Old)**, Mopho + Old Testament dataset → **(Morpho+Old)**, Base + New Testament dataset → **(Base+New)**, Morpho + New Testament dataset → **(Morpho+New)**, Base + Combined Old and New Testament datasets → **(Base+Combined)**, and finally, Morpho + Combined Old and New Testament datasets → **(Morpho+Combined)** models for training.
All systems undergo standard pre-processing stages described in [27]. The only difference, Luganda dataset materials used for Morpho SMT systems, first undergo morphological segmentation based on compiled GNC prefixes, after tokenization during standard pre-processing [28], [29]. Word alignment and language model training for the systems follows using Giza++ and KenLM respectively, setting our language model to 5 n-gram, sentence length limit at 90 throughout all the systems. Then followed by translation model training for all systems with model tuning carried out using each system's development dataset.

6. RESULTS AND DISCUSSION

6.1 Segmented Luganda dataset distribution

Table 5 shows token and vocabulary info after segmentation of our Luganda (target side) in datasets used for Morpho+Old, Morpho+New and Morpho+Combined SMT models. Overall, after morphological segmentation, token count increased while vocabulary count decreased. This phenomenon is in line with our expectations. For example in Table 3, many would be verbs are disintegrated after segmentation because hypothetically they are supposed to be represented by one word *'tambula'* (to walk). Additionally, this behaviour indirectly increases translation probability for such a disintegrated word than having many verbs fighting for the same spot and little chance appearing in actual translation.

Table 5: Token distribution between unsegmented and segmented Luganda Bible sentences

Training Dataset	Tokens		Vocabulary		Sentences
	Unsegmented	Segmented	Unsegment	Segmented	
Combined	710,924	1,033,255	52,516	27,527	30,225
Old	550,742	801,782	43,147	23,317	22,598
New	160,182	231,473	17,365	9,757	7,627

Old and New Testaments, Combined - merged Old and New Testament datasets.

6.2 BLEU and NIST metric analysis

Using six held out test datasets, for each system, we perform automatic evaluation. We use renown BLEU [30] and NIST [31] for scoring our systems represented in Table 6. In particular, we use 'mteval-v14.pl' script used normally in MT workshops [32].
From Table 6 all our 3 Morpho+Models performed better than the Baseline models with (Morpho+New) scoring above other systems with a gain of 1.58 BLEU points and 0.2257 NIST score against the Baseline. However, we notice that for systems trained on New Testament of the Bible, evaluation results are a bit low when compared to systems trained on Old and Combined datasets. We believe this may be due to the low-resource nature of the New Testament dataset in terms of tokens in comparison to the Old Testament as observed in Table 5. Interestingly, the Combined models for Baseline and Morphological segmented models, do not perform better than the models trained on only the Old Testament. In this scenario, we believe inherently Old and New Testament datasets possess some distinctions that make them slightly different domains which in turn impacts on performance of Combined dataset [33], [34]. We also observe the same trend in terms of out of vocabulary (OOVs) results for each system in Table 7. The Morpho models registered a reduction in untranslated words (OOVs) with (Morpho+New) having highest reduction by 53 OOVs in comparison to (Base+New) model.

Table 6: Final BLEU and NIST results for system models trained

System Model	BLEU			NIST		
	Old	New	Combined	Old	New	Combined
Baseline	29.80	20.96	27.84	5.9371	4.5303	5.9268
Morpho+Model*	**31.20**	**22.54**	**28.68**	**5.9663**	**4.7560**	**5.9579**

Old and New Testaments, Combined - merged Old and New Testament datasets. *BLEU and NIST automatic evaluation for Morpho+Models is after post-processing of translated sentences.

Table 7: Comparison of OOV results

System Model	OOV 1-gram tokens			OOV 1-gram tokens %		
	Old	New	Combined	Old	New	Combined
Baseline	106	253	234	1.5%	7.1%	2.2%
Morpho+Model*	**86**	**200**	**194**	**0.8%**	**3.9%**	**1.2%**

6.3 Comparison analysis

To the best of our knowledge, no known BLEU or NIST scores were available for English to Luganda translation pair at the time of our writing. We therefore compare our results to some closely related works on low-resource SMT.

If we compare to [14], the authors using Moses reported (BLEU 20.0 and NIST 2.92) for their best system with SAWA English to Swahili corpus where Swahili is a low-resource language. Using a small trilingual corpus including New Testament of the Holy Bible data, [13] reported BLEU scores of 18.0, 22.0 and NIST scores of 4.12, and 5.31 respectively for English to Luo Moses MT. More reported results for low resource translation pairs include that of Chinese-Mongolian Moses MT with (BLEU 19.34 and 23.71) by [15] and also English-Hindi, Hindi-Marathi Moses MT with (BLEU 35.49 and 27.58) by [5].

Our results in Table 6 are closely within the ranges of scores similar to those in related works' scores for other low-resource language MT tasks. We believe, as more data for English-Luganda pair becomes readily available, and with more MT research, the performance is likely to improve greatly.

7. CONCLUSION

In this paper, we introduce for the first time English to Luganda statistical machine translation with Moses, and proposed a general approach for using Ganda Noun Class (GNC) prefixes to segment Luganda sentences (target side) at pre-processing stage and desegmenting them after translation at post processing stage. Our investigations so far, from evaluation of trained systems depict that the approach based on GNC prefixes contributes to enriched performance of MT systems on English to Luganda translation pair. In our immediate future works, we intend to gather more datasets, and investigate combined use of suffixes and infixes in relation to this work.

ACKNOWLEDGMENTS

This paper was supported by Kumoh National Institute of Technology.

REFERENCES

1. P. Koehn , R. Zens, C. Dyer, O. Bojar, A. Constantin, A., and *et al.*, "**Moses: Open Source Toolkit for Statistical Machine Translation**," in *Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07*, 2007, p. 177. https://doi.org/10.3115/1557769.1557821
2. P. M. Lewis, G. F. Simon, and C. D. Fennig, "**Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International**," 2015. Available at: https://www.ethnologue.com/18/language/lug
3. D. M. Eberhard, G. F. Simon, and C. D. Fennig, "**Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International**," 2019. Available at: https://www.ethnologue.com/country/ug
4. Uganda Communications Commision, "**Post, Broadcasting and Telecommunications Market & Industry Q3 Report, July-September 2017**," Kampala, Uganda, 2017.
5. S. Sreelekha and P. Bhattacharyya, "**Role of Morphology Injection in SMT**," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 1, pp. 1–31, Sep. 2017. https://doi.org/10.1145/3129208
6. P. Tran, D. Dinh, and L. H. B. Nguyen, "**Word Re-Segmentation in Chinese-Vietnamese Machine Translation**," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 16, no. 2, pp. 1–22, Nov. 2016. https://doi.org/10.1145/2988237
7. F. Aqlan, X. Fan, A. Alqwbani, and A. Al-Mansoub, "**Improved Arabic–Chinese Machine Translation with Linguistic Input Features**," *Futur. Internet*, vol. 11, no. 1, p. 22, Jan. 2019. https://doi.org/10.3390/fi11010022
8. Y. Tedla and K. Yamamoto, "**Morphological Segmentation for English-to-Tigrinya Statistical Machine Translation**," *Int. J. Asian Lang. Process.*, vol. 27, no. 2, pp. 95–110, 2017.
9. P. Smit, S. Virpioja, S. Grönroos, and M. Kurimo, "**Morfessor 2.0: Toolkit for statistical morphological segmentation**," in *Proc. of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2014, pp. 21–24.
10. S. Grönroos, S. Virpioja, P. Smit, and M. Kurimo, "**Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology**," in *Proc. of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1177–1185.
11. A. Hurskainen, "**Machine Translation of the Bible**," 2009.
12. A. Hurskainen, "**Swahili Language Manager: A Storehouse for Developing Multiple Computational Applications**," *Nord. J. African Stud.*, vol. 13, no. 3, pp. 363–397, 2004.
13. G. De Pauw, N. Maajabu, and P. W. Wagacha, "**A Knowledge-Light Approach to Luo Machine Translation and Part-of-Speech Tagging**," in *Proc. of the Second Workshop on African Language Technology - AfLaT 2010*, 2010, pp. 15–20.
14. G. De Pauw, P. W. Wagacha, and G. De Schryver, "**Towards English-Swahili Machine Translation**," in *Research Workshop of the Israel Science Foundation: Machine Translation and Morphologically-rich Languages*, 2011.
15. W. Linyu, L. Miao, C. Lei, Y. Zhenxin, S. Kai, and *et al.*, "**Extracting bilingual multi-word expressions for low-resource statistical machine translation**," in *2015 International Conference on Asian Language Processing (IALP)*, 2015, pp. 21–24. https://doi.org/10.1109/IALP.2015.7451522
16. J. E. Ortega and K. Pillaipakkamnatt, "**Using Morphemes from Agglutinative Languages like Quechua and**

Finnish to Aid in Low-Resource Translation," in *Proc. of 13th AMTA 2018 Workshop: LoResMT 2018*, 2018, pp. 1–11.

17. A. C. Hamilton, P. N. Mukasa, D. Ssewanyana, C. Ssentoogo, and C. H. S. Kabuye, ***Luganda Dictionary and Grammar: Luganda-English and English-Luganda Dictionary with Notes on Luganda Grammar***. Alan Hamilton, 2016.
18. J. D. Chesswas, ***The Essentials of Luganda***, 4th ed. London: 0xford University Press, 1967.
19. K. B. Kiingi, "**A style manual for English-Luganda-English dictionary definers and Luganda terminological modernizers**," 2005.
20. E. Baertlein and M. Ssekitto, "**Luganda Nouns: Inflectional Morphology and Tests**," *Linguist. Portfolios*, vol. 3, no. 1, pp. 21–26, 2014.
21. R. S. Balagadde and P. Premchand, "**The Structured Compact Tag-Set for Luganda**," *Int. J. Nat. Lang. Comput.*, vol. 5, no. 4, pp. 01–21, Aug. 2016. https://doi.org/10.5121/ijnlc.2016.5401
22. M. E. Ssemakula, "**A Basic Grammar of Luganda**," *The Buganda Home Page*. Available at: http://www.buganda.com/ggulama.htm
23. WordProject, "**Luganda online Bible with MP3 Audio**," *International Biblical Association*. Available at: https://www.wordproject.org/bibles/lug
24. P. Koehn, F. J. Och, and D. Marcu, "**Statistical phrase-based translation**," in *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, 2003, vol. 1, pp. 48–54. https://doi.org/10.3115/1073445.1073462
25. K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "**Scalable Modified Kneser-Ney Language Model Estimation**," in *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 690–696.
26. F. J. Och and H. Ney, "**A Systematic Comparison of Various Statistical Alignment Models**," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, 2003. https://doi.org/10.1162/089120103321337421
27. P. Koehn, "**MOSES Statistical Machine Translation System User Manual and Code Guide**," 2019.
28. D. M. Mathews and S. Abraham, "**Human Annotation and Interpretation of Public Sentiments about Jio Coin marked in Social Networks using Machine Learning Algorithms**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 4, pp. 1162–1167, Aug. 2019. https://doi.org/10.30534/ijatcse/2019/25842019
29. I. S. Makki and F. Alqurashi, "**An Adaptive Model for Knowledge Mining in Databases 'EMO_MINE' for Tweets Emotions Classification**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 7, no. 3, pp. 52–60, Jun. 2018. https://doi.org/10.30534/ijatcse/2018/04732018
30. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "**BLEU**," in *Proc. of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001, no. July, p. 311. https://doi.org/10.3115/1073083.1073135
31. G. Doddington, "**Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics**," in *Proc. of the Second International Conference on Human Language Technology Research*, 2002, pp. 138–145.
32. S. Ding, H. Khayrallah, P. Koehn, M. Post, G. Kumar, and *et al.*, "**The JHU Machine Translation Systems for WMT 2017**," in *Proc. of the Conference on Machine Translation (WMT)*, 2017, vol. 2, pp. 276–282. https://doi.org/10.18653/v1/W17-4724
33. N. Durrani, H. Sajjad, S. Joty, A. Abdelali, and S. Vogel, "**Using Joint Models for Domain Adaptation in Statistical Machine Translation**," in *Proc. of MT Summit XV, MT researchers' track*, 2015, vol. 1, pp. 117–130.
34. M. Huck, A. Birch, and B. Haddow, "**Mixed-domain vs. multi-domain statistical machine translation**," in *Proc. of MT Summit XV, MT researchers' track*, 2015, vol. 1, pp. 240–255.