Morphological Segmentation for English-to-Tigrinya Statistical Machine Translation

Yemane Tedla and Kazuhide Yamamoto
Natural Language Processing Lab
Nagaoka University of Technology
Nagaoka city, Niigata 940-2188, Japan
yemane@jnlp.org, yamamoto@jnlp.org

Abstract

We investigate the effect of morphological segmentation schemes on the performance of English-to-Tigrinya statistical machine translation. Tigrinya is a highly inflected Semitic language spoken in Eritrea and Ethiopia. Translation involving morphologically complex and low-resource languages is challenged by a number of factors including data sparseness, word alignment and language model. We try addressing these problems through morphological segmentation of Tigrinya words. As a result of segmentation, out-of-vocabulary and perplexity of the language models were greatly reduced. We analyzed phrase-based translation with unsegmented, stemmed, and morphologically segmented corpus to examine their impact on translation quality. Our results from a relatively small parallel corpus show improvement of 1.4 BLEU or 2.4 METEOR points between the raw text model and the morphologically segmented models suggesting that segmentation affects performance of English-to-Tigrinya machine translation significantly.

Keywords

Tigrinya language; statistical machine translation; low-resource; morphological segmentation

1. Introduction

Machine translation systems translate one natural language, the *source language*, to another, the *target language*, automatically. The accuracy of statistical machine translation (SMT) systems may not be consistently perfect but often produces a sufficient comprehension of

the information in the source language. The research presented here investigates English-to-Tigrinya translation system, using the Christian Holy Bible ("the Bible" hereafter) as a parallel corpus.

Tigrinya belongs to the Semitic language branch together with Arabic, Hebrew, Amharic and Tigre. Semitic languages have distinct and complex morphology characterized by "root-and-pattern template" morphology. The writing system of Tigrinya is called Ge'ez script. Each Ge'ez alphabet embeds a consonant and vowel in a single syllable. When Ge'ez is transliterated to Latin script, the consonants form the "root" of the word and the vowels constitute the "template". For example, for the word *sebere* 'he broke', the root is *s-b-r* (the consonants) and the template is *-e-e-e-* (the vowel pattern). Inflection and morphological derivation are performed by morpho-tactics including vowel alteration, morpheme affixation, germination as well as reduplication of consonants.

Tigrinya verb roots consist of mostly tri-literal consonants. Tigrinya words are inflected for tense-aspect-mood, gender, number, person, case, voice and so on. Additionally, there are clitics of prepositions and conjunctions that are affixed to words. The high rate of inflection in Tigrinya generates a large number of word forms which cause data sparsity problem in Tigrinya language processing. This poses a challenge for machine translation systems aggravating the out-of-vocabulary (OOV) problem caused by insufficient data. Therefore, most Semitic SMT research has experimented with different schemes of morphological segmentation in the preprocessing phase in order to alleviate OOV growth rate and improve token alignment.

In this research, we investigate the effect of morphological segmentation of Tigrinya words for English-Tigrinya SMT. We analyze phrase-based translation system and report the performance gain achieved as a result of segmenting Tigrinya words.

2. Related Work

The nature of SMT may differ depending on the inflection complexity of the involved languages. Nevertheless, most studies show the use of segmentation schemes helping SMT translation quality to some extent. Popovi´c and Ney (2004) applied stemming which resulted in reduction of translation errors from Spanish, Catalan and Serbian to English. Our first approach is similar to this work, whereby we also try partially segmenting or stemming words rather than performing full analysis of morphological affixes. Some researchers suggest that simple segmentation can also perform as much as complex approaches (Haj and Lavie, 2010). Earlier, Habash and Sadat (2006) investigated the impact of different segmentation schemes on Arabic SMT. They reported that segmentation schemes can vary between just proclitics pruning to sophisticated morphological analysis

based on the availability of data. Sarikaya and Deng (2007) proposed joint morphological-lexical language model for translation involving a morphologically complex language. In their experiments concerning English-to-Arabic translation, they reported improved translation quality over trigram baseline. Badr et. al., (2008) advanced the work to Arabic translation by making use of context information along with segmentation. Similar studies on Hebrew-English SMT show an improvement on BLEU score (Singh and Habash, 2012). In the research, it was shown that linguistically motivated morphological analyzer performed better when compared to unsupervised analyzer. Amharic is another Semitic language closely related to Tigrinya in morphology and syntax. Amharic has relatively better support of resources and natural language processing (NLP) research compared to Tigrinya. Mulu and Besacier (2012), experimented on phrase-based English-Amharic SMT with 18,434 sentences parallel corpus, achieving a baseline score of 35.32%. They applied morphological segmentation to the Amharic data and were able to improve the BLEU score by 0.92%. Furthermore, a cloud platform from *ethiocloud.com* (*translator.abyssinica.com*) also provides Amharic-English translation. Additionally, Amharic is among the languages supported by Google translate.

Regarding Tigrinya, we find a very recent promising project which released a web application of English-to-Tigrinya translation (*tigrinyatranslate.com*). Tigrinya is not yet supported by Google translate and has very few entries (often empty) on Wikipedia. Apart from the Bible translation, we could not find other parallel corpus open for the public. However, some on line dictionaries and mobile applications are being developed. For example, *memhir.org* [1] maintains a dictionary of over 15,000 entries. Recently, geezexperience.com[2] is compiling a multilingual dictionary for Tigrinya and a number of other languages including English, German, Dutch, Italian, and Swedish. *Hidri publishers*[3] also provide a mobile version of their printed dictionary "Advanced English-Tigrinya Dictionary", with over 62,000 entries.

In this research, we use the Bible as our parallel corpus. Resnik et. al., (1999) discuss the usefulness of the Bible for language processing. They mention that the Bible has been translated to over 2000 languages, making it the most translated book in the world. The Bible text is carefully translated and organized on verse-level. According to Resnik et. al., (1999), the Bible has about 85% coverage of modern-day vocabulary and variations of writing styles. SMT requires large parallel data for high-quality translations. Therefore, the

---

1 *http://www.memhr.org/dic/*

2 http://www.geezexperience.com/dictionary/?dr=0

3 https://play.google.com/store/apps/details?id=org.fgaim.android.aetd2&hl=en

Bible alone will not be sufficient for building high-quality translation. However, it is easily accessible and may be tailored to build experimental models and investigate certain behaviors. Phillips (2001) used the Bible as bootstrapping text to set the parameters of a stochastic translation system and noted the prospects of enabling translation of thousands of languages using the Bible as a basis.

3. Methodology

3.1. Preprocessing

SMT systems are built from a large volume of source-to-target aligned sentences. The Tigrinya Bible is available in different formats on a number of websites and mobile applications[4]. There are also plenty of sources for several editions of the English Bible over the Internet. A version of Tigrinya-English Bible is available on *geezexperience.com*. However, the translations are not strictly aligned verse-to-verse. A verse in the Bible may contain one or more sentences. On the translation's Tigrinya side, there is frequent combination of one or more consecutive verses into a single verse. It is difficult identifying the boundary of combined verses automatically. Therefore, we corrected the verse alignment by joining the English counterparts as well. Following this, the corpus was cleaned and tokenized. During the cleaning phase, we retained a few types of punctuation in the Tigrinya corpus and transliterated Ge'ez script to Latin script for better manipulation during segmentation[5]. The English text was also tokenized and lowercased to minimize data sparseness. However, the Tigrinya corpus was not lowercased since lower and uppercases represent distinct syllables in the Tigrinya transliteration. After the preprocessing phase, the parallel corpus contained 31,277 aligned verses. We divided this corpus into training, tuning and test sets by extracting verses at random for use with the Moses translation system.

3.2. Segmenting Tigrinya words

Tigrinya verbs, nouns, and adjectives are highly inflected. As a result, a single "token" in Tigrinya may embed a number of grammatical information that are expressed by using many tokens in English. For example, the token *InItezeyItesebIrenI* can be translated as "and if it did not break" using six words in English. Hence, there is considerable

---

4 *http://bible.geezexperience.com/tigrigna/, https://www.betezion.com/bible.php*

5 We adopt the SERA transliteration scheme (f*tp://ftp.geez.org/pub/sera-docs/sera-faq.txt).* In this paper, the letter "I" is added to mark the epenthetic vowel, known as "sadIsI".

unevenness in the count of tokens between these two languages. One solution to introduce better correspondence of words is decomposing the inflected Tigrinya word into its constituent morphemes. In the previous example, the Tigrinya word can be segmented into six morphemes; *InIte* zeyI* te* sebIr +e +nI*,[6] matching the total number of words of its English translation. This method is useful in reducing data sparseness and creating a better token-to-token correspondence. As can be inferred from Table 3, the number of tokens in the Tigrinya corpus grows by over 30% after segmentation. This increase minimizes the difference in token count between English and morphologically segmented Tigrinya corpus from 37.7% to only 0.02%. This research is conducted to evaluate whether translation to Tigrinya benefits from the effect of this segmentation.

We employ two methods of segmentation described as follows:

3.2.1. Affix-based segmentation

We performed affix based segmentation or shallow segmentation of Tigrinya words based on longest affix pruning. A list of Tigrinya prefixes and suffixes was compiled from several Tigrinya corpora, based on character n-gram frequency. The corpora include a 9.1 million news text from *Haddas Ertra* newspaper, the Bible and a Tigrinya lexicon crawled from the Internet[7]. The shallow segmentation produces three segments (sub-word units) as "*longest-prefix Stem Longest-suffix*". Note that we use "prefix, stem, suffix" for simplicity; however, the sub-words here are not necessarily valid linguistic units. To reduce over-stemming words, the minimum stem threshold was set to five characters. This threshold is selected because most Tigrinya words, especially verbs contain tri-literal roots or about six characters when transliterated.

3.2.2. Morphological segmentation

For deeper segmentation, we use an in-house morphological segmentation model based on conditional random fields (CRF). The model detects morphological boundaries using character based IOB tagging scheme[8]. This model performs almost full morphological analysis of the input word. For example, "*zeyItesebIrunI*" (word-3 in Table 1) can be stemmed to "*zeyIte* sebIr +unI*" and our model would further segment the composite prefix "*zeyIte**" into "*zeyI* te**" and the suffix "+*unI*" into "+*u +nI*". We note that the prefix "*zeyI*" is a fused form of "z*I*" relativizor and "*ayI*" negative circumfix; which have

---

6 The asterisk (*) and cross (+) signs are attached to the segments to mark the prefix and suffix morphemes respectively.

7 *http://www.cs.ru.nl/biniam/geez/crawl.php*

8 IOB – Inside-Outside-Begin tagging scheme

gone vowel alteration making their boundary obscure. We did not segment this case in this study. The model segments morphological boundaries with state-of-the art accuracy of 97%. Table 1 shows some examples of segmented words with raw text, stemmed version (prefix-stem-suffix) and the morphologically segmented version of the same word.

	Word 1	Word 2	Word 3
Word	zIsebere	InIteseberetI	zeyItesebIrunI
Stemmed	zI* seber +e	InIte* seber +etI	zeyIte* sebIr +unI
Segmented	zI* seber +e	InIte* seber +etI	zeyI* te* sebIr +u +nI
Gloss	he who broke	if she broke	that were not broken and
Grammar	REL STEM singular-3rd-masc.	CONJ STEM singular-3rd-femin.	REL-NEG PASSIVE STEM plural-3rd-masc. CONJ

Table 1: Example of segmented words and grammatical functions of the segments

3.3. Phrase-based translation system

In this work we employ phrase-based statistical machine translation. In phrase-based translation the source sentence is segmented into phrases and then each phrase is translated into target phrase. Finally, the translated phrases are combined (reordered) to form the target sentence. For example, consider the following sentence pairs as training sentences:

English source sentence: "how do you translate this?"

This sentence can be translated into Tigrinya as:

Tigrinya target sentence: "*Izi kemeyI gErIka tItIrIgWImo ?* "

The following phrase pairs can be formed from the given training sentences.

Source sentence	Target sentence
how do you	*kemeyI gErIka*
translate	*tItIrIgWImo*
this	*Izi*
?	*?*

Table 2: Example of phrase-based translation pairs

Based on such type of large bilingual text, phrase-based systems learn models that can predict the most probable translation outputs. Phrase translation based on noisy channel

model is defined as follows using the Bayes rule:

$$argmax_t\ p(t|s) = argmax_t\ p(s|t)p(t)$$

Where t = target sentence, s = source sentence and p = probability distribution;
The probability $p(t)$ forms the language mode while $p(s|t)$ models the phrase translation.
During decoding, the source sentence s is segmented into sequence of n phrases s_1^n.
Each source sentence s_i in s_1^n is translated into a target phrase t_i. Then a possible reordering of the target sequences into a sentence is performed using a distortion probability model. We use Moses translation system for our phrase-based experiments as described in the next section.

4. Experiments and Results

4.1. Moses setup

The tools we used for building phrase-based translation model, language model (LM) and evaluations are from Moses SMT toolkit (Koehn et al., 2007). Word alignment was performed with MGIZA++, an extended and optimized multi-threaded version of GIZA++. While there is a large variation of individual verse length in the corpus, the average verse length of the Tigrinya unsegmented corpus is 19.9 and grows to 31.4 after morphological segmentation (Table 3). Therefore, for the cleaning step, the maximum length of sentences is set to 60. We train language models using KenLM tool which has the advantage of being fast and uses low memory. The order of n-grams is set to five to account for words split by segmentation. Six language models are built based on the segmentation schemes and two datasets. The reordering limit for the distortion model is set to the default six. We explain our settings for the datasets, evaluation tests and the baseline system as follows.

1) Data: The training, tuning, and test data are all extracted randomly from the Bible parallel corpus. Table 3 lists the size of the verse-aligned parallel corpus (*dataset-1*), whereas Table 4 lists the size of sentence-aligned parallel corpus (*dataset-2*) extracted from *dataset-1*. Note that in *dataset-1*, there are verses combined for strict alignment as explained in the preprocessing section. In this way, the verse-aligned corpus consisted of 31,277 verses; with 29,307 verses used for training, 970 verses for tuning and 1000 verses held out for testing. However, the verse alignment process also introduces lengthy sentences, possibly making word alignment more difficult due to too different sentence alignment in the verses. Given the small size of the corpus, this may affect the overall quality of alignments. In order to investigate its effect on translation quality we constructed *dataset-2*, the sentence-aligned parallel corpus by extracting only the single sentence verses

based on the Tigrinya corpus. Sentence identification was performed using the sentence-end marker of Tigrinya. Therefore, all verses with a single sentence-end marker were extracted. Consequently, the extracted corpus comprises a total of 20,578 parallel sentences, which is about 65.8% of the original verse-aligned corpus. We notice that the morphologically segmented Tigrinya corpus (*dataset-1*) is the closest match to the number of tokens in the English corpus. The average verse length of the English tokens is 32.0 and that of the morphologically segmented Tigrinya corpus is 31.4 (Table 3). It is interesting to see whether this match would be more useful in creating effective word alignments for the machine translation (MT) models. Our experimental results using both corpora are summarized in Tables 5,6,7 and 8.

Data	Verses	English tokens	Tigrinya Tokens		
			unsegmented	stemmed	morph-segmented
Training-1	29,307	938,837	584,318	837,675	918,719
Test-1	1,000	31,994	20,042	28,808	31,500
Tuning-1	970	31,383	19,624	28,254	30,889
Dataset-1	31,277	1,002,214	623,984	894,737	981,108
Average verse length		32.0	19.9	28.6	31.4

Table 3: Dataset-1, Verse-aligned parallel corpus

Data	Sents	English tokens	Tigrinya Tokens		
			unsegmented	stemmed	morph-segmented
Training-2	19,299	581,799	356,002	513,876	563,696
Test-2	651	19,980	12,292	17,884	19,545
Tuning-2	628	18,799	11,596	16,710	18,380
Dataset-2	20,578	620,578	379,890	548,470	601,621
Average sent. length		30.2	18.5	26.7	29.2

Table 4: Dataset-2: Sentence-aligned parallel corpus

2) Evaluation: We evaluate the performance of translation using BLEU, METEOR and TER metrics. We also analyze the perplexity and OOV statistics to investigate the LM improvement achieved by segmentation. OOVs are tokens in the test data that are not

present in the training data and perplexity measures complexity of LMs. Lower perplexity score indicates a better fit of the test set to the reference set.

We designed four sets of system evaluations based on the MT models and the test sets. The settings are given as follows:

1. Verse based models (*MT-verse*) : Tables 5, 6

   Dataset-1: the verse-aligned corpus

   - contains 31,277 verses (1 verse >= 1 sent)

   Test-1: 1000 verses

2. Sentence based models (*MT-sent*): Tables 5, 6

   Dataset-2: extracted only the single sentence verses from *Dataset-1*

   - contains 20,578 sentences

   Test-2: extracted only the single sentence verses from *Test-1*

   - contains 651 sentences

3. *Dataset-1 + Test-2*: Tables 7, 8

The result of *MT-verse* and *MT-sent* models may not be directly comparable since the test data used is different in both cases. Therefore for a better comparison of the two models, we evaluated both models using the sentence-based test data (*Test-2*).

4. Raw text models against segmented text models: Tables 5, 6

For models from the segmented corpus evaluation is straightforward; the translation output of the segmented MT model and the segmented reference are evaluated. However, the baseline model is built from the unsegmented parallel corpus. Therefore, the evaluation output of the unsegmented and segmented MT models are not directly comparable. Hence, for fair and easier comparison, we segmented the translation output of the baseline model and evaluated it against the segmented reference. Models *system-1b* and *system-1c* are examples of such models (Table 5).

Another comparison method is restoring the translated segments in the segmented models to their root words and evaluating against a raw text reference. This method conducts evaluation between words instead of morphemes and can be performed with a separate detokenization algorithm or input data representation, which is left for future research.

3) Baseline: The baseline system is built from the clean and tokenized (unsegmented) version of the Bible corpus. The performance in terms of BLEU score is 15.6 for the verse-aligned models (*MT-verse*) and 13.0 for the sentence-aligned models (*MT-sent).* All the models from the segmented corpus outperform these baseline scores.

4.2. The effect of segmentation

The experimental results of the baseline and segmented models are described in Tables 5 through 8. Following we discuss the the effect of segmentation according to the evaluation settings mentioned earlier.

4.2.1. MT-verse system Vs. MT-sent system

In general, the evaluation metrics in Table 5 show that the verse-aligned models score better results than the sentence-aligned models. However the performance drop in the *MT-sent* models is likely due to the larger proportion of OOVs resulting from restricting the corpus to single-sentence verses only. Table 6 shows the OOV ratio and model perplexity of the two systems. For example, the OOV ratio of *MT-verse* baseline is 6.7% while that of *MT-sent* is 8.5%. Similarly the perplexity of *MT-sent* is higher than *MT-verse*. Therefore verse aligned models scored better results. Nonetheless, the difference is rather small, suggesting that with proper data size, sentence based models may have performed better. For example, the BLEU score of *sys-segm* and *sys-segm-sent* is 20.7 and 19.8 respectively. The difference is only 0.9 BLEU points, although the *MT-sent* corpus is much smaller than the *MT-verse* corpus. Notice that the *MT-verse* models are tested on the 1000 test verses (*Test-1*). These verses include single-sentence verses as well as multiple-sentence verses. However, *MT-sent* models are tested on 651 single-sentence verses (*Test-2*) extracted from *Test-1*.

MT system	System	MT-models	BLUE	METEOR	TER
MT-verse	System-1	sys-base	15.6	19.7	74.2
	System-1b	sys-base-stem	19.8	21.1	**71.0**
	system-1c	sys-base-segm	19.3	20.9	**71.0**
	System-2	sys-stem	**20.9**	22.7	72.7
	System-3	sys-segm	20.7	**23.3**	71.7
MT-sent	System-4	sys-base-sent	13.0	17.8	76.7
	System-5	sys-stem-sent	18.8	21.1	74.4
	System-6	sys-segm-sent	**19.8**	**22.9**	**72.5**

Table 5: MT-verse and MT-sent: BLEU, METEOR, and TER scores

System	Test tokens	OOV count	OOV ratio (%)	Perplexity
System1-base	21042	1408	6.7	270
System2-stem	29808	757	2.5	69
System3-segm	32500	664	**2.0**	**52**
System4-base-sent	12943	1106	8.5	317
System5-stem-sent	18535	604	3.3	79
System6-segm-sent	20196	**532**	**2.6**	**59**

Table 6: Perplexity: Test tokens with OOVs included

Therefore, in order to compare *MT-verse* with *MT-sent* we evaluated both systems under *Test-2* dataset. The evaluation and perplexity scores are presented in Tables 7 and 8. We observe that this version of *MT-verse* outperforms *MT-sent* under all metrics. This may be attributed to the fact that OOVs are greatly reduced when using the smaller test set, *Test-2* against *MT-verse*, which uses larger training set.

Test data/model	BLUE	METEOR	TER
System1-Test2/unseg	14.5	18.9	75.8
System2-Test2/stem	**20.0**	22.2	73.5
System3-Test2/segm	19.8	**22.9**	**72.5**

Table 7: MT-verse: BLEU, METEOR, and TER scores tested on Test2

System	Test2 tokens	OOV count	OOV ratio (%)	Perplexity
System1-Test2/base	12943	913	7.1	291
System2-Test2/stem	18535	477	2.6	70
System3-Test2/segm	20196	**422**	**2.1**	**53**

Table 8: MT-verse: Perplexity and OOV evaluated on Test-2

4.2.2. Unsegmented, stemmed and morphologically segmented models

Overall, we observe that segmentation has improved the machine translation quality compared to the unsegmented baseline. In the *MT-verse* system, the baseline for *sys-stm* model is *sys-base-stm* while that of *sys-segm* is *sys-base-segm*. We see a BLEU score improvement of 1.1 and 1.4 over the baseline compared to the stemmed and segmented models. Although the BLEU score of the stemmed model is marginally better than the segmented model (20.9 vs 20.7), the METEOR and TER metrics show that the morphologically segmented model outperforms the others. Moreover, the metrics for the segmented models of the *MT-sent* system consistently show the best results. The analysis of OOVs and perplexity on Table 6 further clarifies the reason for the performance gain. The OOV count decreases from 1408 for the baseline to 664 for the segmented model; which is a reduction of 2.1% in the OOV count. Similar reduction pattern is also reflected in the *MT-sent* models. Since the test data size is different for the models, the OOV ratio may better explain the OOV size in relation to the test data. Accordingly, we see higher rates of OOV in the unsegmented models while the ratio of OOVs to the test data decreases with finer segmentation. We note that this ratio has negative effect on perplexity; the higher the OOV ratio, the worse the perplexity. Model *system3-segm* in Table 6 has the lowest OOV ratio and perplexity among all the models while s*ystem6-segm-sent* has the lowest score among the *MT-sent* models. Therefore, based on these findings, we understand that the fine grained segmentation was better in improving quality of English-Tigrinya machine translation. In general, although our training data is relatively small, the performance gain observed from the segmented systems demonstrates the usefulness of word segmentation strategies for English-Tigrinya machine translation.

4.2.3. Translation outputs

The examples in Table 9 demonstrate translation output of two reference sentences from all models of the MT system. We note very interesting insights both in terms of morphological and syntactic transfer from the source to the target sentence.

The relatively shorter reference sentence (b) has been correctly translated by all models. This may suggest that the models perform well with short sentences. Therefore, we discuss the syntactic structure and meaning preservation aspects of the translation taking the longer sentence (a) as an example. For easier high level discussion, we simplify the source sentence (a) by abstracting it into representative sense sub-phrases (enclosed in square brackets […]). We also convert the reference and Tigrinya outputs of *MT-verse* system into similar sub-phrases as follows:

Source sentence (a):

" [but of the fruit of the tree of the knowledge of good and evil][you may not take ;] [for on the day when you take of it ,][death will certainly come to you .]"

Reference sentence (a):

"[kabIta xIbuQInI kIfuInI ItefIllITI omI gIna :] [kabIa mIsI ItIbelIOI meOalItIsI] [motI kItImewItI iKa Imo :][kabIa ayItIbIlaOI : ilu azezo .]"

Conversion to English sub-phrase units:

Source:	[of-tree][**don't-take**][if-you-take][die]
Reference:	[from-tree][if-you-eat][die][**don't-eat**]
Translated-unseg:	[from-tree][**don't-eat**] [if-you-eat][die]
Translated-stm:	[from-tree][**don't-eat**][if-you-eat][die]
Translated-seg:	[tree-you][**don't-eat**][if-you-eat][death]

Generally, Tigrinya has subject-object-verb structure while English follows subject-verb-object ordering. In all the translations, the phrase order of the Tigrinya translation output aligns better with the English source rather than the Tigrinya reference. The boldface sub-phrases demonstrate this observation. In this specific case, the order alteration does not make sentence comprehension very difficult. However, invalid ordering may create ungrammatical translations which can also make the meaning difficult to understand. The *translated-seg* is more difficult to understand because the original meaning is not entirely preserved. Some studies have shown that aggressive segmentation into very fine units might actually hurt the translation quality by unnecessarily enlarging the phrase table and worsening the uncertainty of choosing the correct phrase candidate (Haj and Lavie, 2010). There are two problems with *translated-seg (system 3 in* Table 9*)* . First, the beginning phrase is translated to 'you are the tree' which is different from the original phrase that has the sense of 'from-tree'; and second the last phrase 'you will die' is wrongly translated as 'it is death'. In comparison, the *translated-stm* output preserves the meaning in the references better than the segmented model. However, *translated-seg* seems to have better token coverage compared to the other models. For example, the word '*ilu azezo'* in the reference was only found in the *translated-seg* models. This could be the reason why *translated-seg* scores are better because BLEU is a token level metric. Therefore, in post-processing, a de-tokenization step is required to attach morphemes with their root words and then make the evaluation from the words. We plan this type of analysis for future research.

System	Sentence
Source	(a) but of the fruit of the tree of the knowledge of good and evil you may not take ; for on the day when you take of it , death will certainly come to you . (b) and the name of the second river is gihon : this river goes round all the land of cush .
Reference	(a) kabIta xIbuQInI kIfuInI ItefllITI omI gIna : kabIa mIsI ItIbelIOI meOalItIsI motI kItImewItI iKa Imo : kabIa ayItIbIlaOI : ilu azezo . (b) sImI Iti KalIayI rIba dIma gihonI Iyu : nIsu nIKWla mIdIri kushI yIzora .
Baseline (unsegm.)	(a) gInaKe : kabI fIre omI ayItIbIlaOI : kabIa mIsI ItIbelIOI meOalItIsI : motI kItImewItI iKa . (b) sImI Iti KalIayI rIba dIma gihonI Iyu : nIsu nIKWla mIdIri kushI yIzora .
System 2 (stemmed)	(a) kabI fIre Ita gIna kI* f +uInI Ite* fIlITI omI dIma Imo : kabIa ayItIbI* laOI : kabIa mIsI ItI* belIOI meOalItIsI motI kItI* mew +ItI iKa : nI* sI +Ka . (b) sImI Iti KalI +ayI rIba dIma giho +nI Iyu : nIsu nI* KWla mIdIri kushI yI* zora .
System 3 (morph-seg)	(a) gIna +Ke It +i fIre It +a xIbuQI +nI kIfuI +nI i +Ka Imo : kabI +a ayI* tI* bIlaOI : beta meOalIti Iti +a : kabI +a mIsI ItI* belIOI meOalItI sI motI Iy +u : il +u azez +o . (b) sImI It +i KalIayI rIba dIma giho +nI Iy +u : nIsu nI* KWla mIdIri kushI yI* zor +a .
System 4 (MT-sent-unseg)	(a) gInaKe Iti fIre Ita xIbuQInI kIfuInI ItefllITI omI dIma iKa Imo , beta meOalIti mIsI : motI kItImewItI iKa . (b) sImI Iti KalIayI rIba dIma gihonI Iyu : nIsu nIKWla mIdIri kushI yIzora .
System 5 (MT-sent-stm)	(a) gInaKe : kabI fIre omI dIma iKa Imo : kabIa ayItIbI* laOI : beta meOalIti Itia : kabIa mIsI ItI* belIOI meOalItIsI motI kItI* mew +ItI iKa . (b) sImI Iti KalI +ayI rIba dIma giho +nI Iyu : nIsu nI* KWla mIdIri kushI yI* zora +nI .
System 6	(a) gIna +Ke It +i fIre It +a xIbuQI +nI kIfuI +nI i +Ka Imo : kabI +a

(MT-sent-morphseg)	ayI* tI* bIlaOI : beta meOalIti Iti +a : kabI +a mIsI ItI* belIOI meOalItI +sI motI Iy +u : il +u azez +o . (b) sImI It +i KalIayI rIba dIma giho +nI Iy +u : nIsu nI* KWla mIdIri kushI yI* zor +a .

Table 9: Sample Translations from the verse based and sentence based models

5. Conclusion and Future work

In this research we investigated the effect of morphological segmentation on the performance of English-to-Tigrinya statistical machine translation. Machine translation between English and Tigrinya is challenging since the target language, Tigrinya, is highly inflected and both languages have morphological and syntactic divergence. Segmentation was performed to help both languages converge to better word alignment, reduce OOVs and improve the language model. We explored two segmentation schemes; one based on longest-affix segmentation and another based on fine grained morphological segmentation. We used a relatively small parallel corpus derived from the Bible translation of both languages. The Bible text was extracted automatically and aligned properly on verse-level and sentence-level. We employed phrase-based translation using tools from Moses toolkit. The experimental results show a promising improvement in translation quality using both schemes. Segmentation reduced the OOVs ratio and perplexity of models and as a result the BLEU, METEOR and TER scores improved. In general, the morphologically segmented models scored better results than the unsegmented baseline and affix segmented models.

Language models are created from monolingual corpus which is easier to build than parallel corpus. In the future, we want to study the effect of large Tigrinya language models on translation quality. Statistical machine translation approaches require large bilingual text to achieve reasonable translation quality. However language resources are a big challenge for under-resourced language such as Tigrinya. In the future, we would like to create a large English-Tigrinya parallel corpus for effective machine translation.

6. References

Badr I., Zbib R., and Glass J., 2008, Segmentation for English-to-Arabic Statistical Machine Translation, In *Proceedings of ACL-08: HLT,* short papers (Companion Volume), pp. 153–156.

Habash N. and Sadat F., 2006, Arabic Preprocessing Schemes for Statistical Machine Translation, In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. Association for Computational Linguistics,* pp. 49–52.

Haj H. A. and Lavie A., 2010, The Impact of Arabic Morphological Segmentation on Broad-coverage

English-to-Arabic Statistical Machine Translation, In *Conference of the Association for Machine Translation in the America (AMTA)*, Denver, Colorado.

Koehn P. et al., 2007, Moses: Open Source Toolkit for Statistical Machine Translation, In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Stroudsburg, PA, USA, pp. 177–180.

Mulu G. T. and Besacier L., 2012, Preliminary Experiments on English-Amharic Statistical Machine Translation, In *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU)*.

Phillips J. D., 2001, The Bible as a basis for Machine Translation, In *proceedings of the Pacific Association for Computational Linguistics*, PACLING-2001.

Popovi´c M. and Ney H., 2004, Towards the Use of Word Stems and Suffixes for Statistical Machine Translation, In *Proceedings of The International Conference on Language Resources and Evaluation*.

Resnik P., Olsen M. B., and Diab M., 1999, The Bible as a Parallel Corpus: Annotating the Book of 2000 Tongues, In *Computers and the Humanities: Selected Papers from TEI 10: Celebrating the Tenth Anniversary of the Text Encoding Initiative*, vol. 33, no. 1/2, Denver, Colorado, pp. 129–153.

Sarikaya R. and Deng Y., 2007, Joint Morphological-Lexical Language Modeling for Machine Translation, In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. Association for Computational Linguistics (NAACL)*, pp. 145–148.

Singh N. and Habash N., 2012, Hebrew Morphological Preprocessing for Statistical Machine Translation, In *Proceedings of the 16th EAMT Conference European Association for Machine Translation*.