Nr.: FIN-004-2018

A Parallel Corpus for Amharic–English Machine Translation

Andargachew Mekonnen Gezmu, Andreas Nürnberger,
Tesfaye Bayu Bati

Data and Knowledge Engineering Group

Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg

Technical report

Nr.: FIN-004-2018

A Parallel Corpus for Amharic–English Machine Translation

Andargachew Mekonnen Gezmu, Andreas Nürnberger,
Tesfaye Bayu Bati

Data and Knowledge Engineering Group

Technical report (Internet)
Elektronische Zeitschriftenreihe
der Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg
ISSN 1869-5078

Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg

Impressum (§ 5 TMG)

Herausgeber:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Der Dekan

Verantwortlich für diese Ausgabe:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Andargachew Mekonnen Gezmu
Postfach 4120
39016 Magdeburg
E-Mail: andargachew.gezmu@ovgu.de

http://www.cs.uni-magdeburg.de/Technical_reports.html

Technical report (Internet)
ISSN 1869-5078

Redaktionsschluss: 04.10.2018

Bezug: Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Dekanat

A Parallel Corpus for Amharic-English Machine Translation

Andargachew Mekonnen Gezmu, Andreas Nürnberger
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik, ITI/DKE
{andargachew.gezmu, andreas.nuernberger}@ovgu.de

Tesfaye Bayu Bati
Hawassa University
Faculty of Informatics
tesfayebayu@hu.edu.et

Abstract

This paper describes the acquisition, preprocessing, segmentation and alignment of an Amharic-English parallel corpus. In doing so we addressed language-specific issues such as normalization and end-of-sentence disambiguation. The corpus consists of 145,820 Amharic-English parallel sentences (segments) from various sources. This corpus is larger in size than previously compiled corpora. It is released[1] for research purposes and can be used to train or support Amharic-English machine translation systems.

1 Introduction

Amharic is a Semitic language that serves as the working language of the Federal Government of Ethiopia. Though it is lingua-franca in Ethiopia and plays several roles in the government, it is considered as a low-resource language with regard to its lack of basic tools and resources for natural language processing (Tachbelie et al, 2014; Strassel and Tracey, 2016; Gezmu et al, 2018).

One of the main challenges in machine translation of Amharic-English is the absence of a clean and sizable parallel corpus. There are a limited number of potential sources, which having both typed and scanned documents, of bilingual documents. Yet the absence of efficient optical character recognition (OCR) tools, and missing lemmatizer and bilingual lexicon with reasonable coverage, are some of the hurdles to extract and align sentences from these sources. Lexicons cannot contain all word forms of Amharic as a word can have many inflections. Amharic is morphologically rich. It is possible for an orthographic word to embed both lexical and grammatical words. Grammatical words like prepositions and conjunctions can be attached to lexical words. For example, ለኛና /ləɲana/ “for us and” embeds the preposition ለ /lə/ “for”, pronoun እኛ /ʔɨɲa/ “us”, and conjunction እና /ʔɨna/ “and”. The boundaries of morphemes in an orthographic words might be unclear. Besides, like other Semitic languages, Amharic verbs have a root-pattern morphology. These make the development of an efficient lemmatizer a difficult task.

By addressing these hurdles, in this project we made an effort to collect bilingual documents and align sentences from various sources.

2 Related Works

There have been parallel corpora of small sizes for Amharic-English machine translation. The most notable ones are the Amharic-English Bilingual Corpus (AEBC) and the Low Resource Languages for Emergent Incidents (LORELEI) Amharic representative language pack.

AEBC [2] is hosted by European Language Resource Association and contains parallel text from legal and news domains. The size of the corpus is 13,379 aligned segments (sentences).

LORELEI-Amharic [3] was developed by the Linguistic Data Consortium and is comprised of a monolingual and parallel Amharic text (Strassel and Tracey, 2016). It has 60,884 segments. Though LORELEI-Amharic is larger than AEBC, still it is not sufficient to train the current state-of-the-art machine translation algorithms. In addition, the parallel text was collected from discussion forums,

[1] Available at: http://dx.doi.org/10.24352/ub.ovgu-2018-145

[2] Available at: http://catalog.elra.info/product_info.php?products_id=1215

[3] Available at: https://catalog.ldc.upenn.edu/LDC2018T04

newswires, and weblogs. Discussion forums and weblogs are susceptible to spelling mistakes, especially for Amharic, since there is no spelling corrector to assist Amharic writers (Gezmu et al, 2018).

3 Data Sources

We have carefully identified potential data sources that could serve as a basis for building a parallel corpus. Our goal was to build a corpus that covers modern-day Amharic on a broad range of topics. For practical reason of getting larger data with the open access, we have assessed newswires, newspapers, magazines, e-books and the Bible.

3.1 Newswires

Major newswires such as Deutsche Welle, BBC, Voice of America, Ethiopian News Agency, Ethiopian Broadcasting Corporation, Fana Broadcasting Corporate and Walta Information Center provide news stories in Amharic and English. In these newswires, the Amharic news stories are intended for the local public of Ethiopia. Because of this, only a small portion of the English news stories are translated into Amharic, or vice versa. For instance, in the Ethiopian News Agency, approximately one news story out of ten has a roughly translated version (Argaw and Asker, 2005). In most cases, the translation style is in a way to summarize the news stories in the target language rather than being faithful sentence-by-sentence translations.

3.2 Newspapers

Two newspapers, the Ethiopian Herald (አዲስ ዘመን in Amharic) and the Ethiopian Reporter, publish bilingual news articles in Amharic and English. Like in the newswires, the articles are not parallel translations.

Another important newspaper is the Federal Negarit Gazeta. Since 1995 Ethiopian proclamations and regulations (both in Amharic and English) have printed in the Federal Negarit Gazeta. In all of its volumes, the contents are divided into two columns, the left-hand side is for Amharic and the right-hand side is for English. The faithful sentence-by-sentence translations make the newspaper a good candidate for a parallel corpus. The problem is that most of the electronically available volumes are written in different fonts (even within a document). Some of the other volumes are scanned copies. The fonts used in the newspaper are not compatible with Unicode. For Amharic script, there are many different font encoding systems, all incompatible with each other. There is a need for the development of a reliable font conversion tool or an OCR system for Amharic.

3.3 Magazines

The Watchtower (መጠበቂያ ግንብ in Amharic) and Awake magazines (ንቁ in Amharic) are electronically available for the public since 2006. The faithful sentence-by-sentence translations make the magazines good candidates for a parallel corpus.

3.4 E-Books

The Ethiopian legal documents such as civil, penal, commercial and maritime codes along with its constitution have parallel translations into English. Though the documents are available as e-books, they are rendered in fonts that are not compatible with Unicode.

3.5 The Bible

The Bible is the world's most translated and easily available book. It was translated with great care and have a high coverage of modern-day vocabulary, as much as 85% (Chew et al, 2006). In addition, its content reflects the everyday living of human beings like love, war, politics, etc. However, older translations of the Bible used archaic language. We found out the recent translations of the bible use the modern-day (contemporary) language. For example, the Standard Version and the New World Translation use the modern-day language both in Amharic and English. This makes them good candidates for the parallel corpus.

4 Data Selection and Preprocessing

We have carefully selected documents from the data sources for inclusion in the parallel corpus. The selection criteria are the faithfulness of the translations and Unicode conformity of the contents of the documents. Awake magazine, the Bible, news articles, and Watchtower magazine were selected for inclusion in the corpus.

After the documents from the archives of the selected sources are downloaded, preprocessing was carried out as a preparation step for the next tasks of segmentation and alignment. The preprocessing mainly involves extraction of text from the documents, normalization of punctuation marks, and case normalization. From the EPUB and HTML documents, texts with paragraph tags were extracted using Python's epub and HTMLParser modules. Boilerplates such as headers, footers (including footnotes) and verse numbers (in the Bible) are removed from the documents. The news articles had already partially preprocessed and made available for research purposes by Geez Frontier Foundation as text files.

Different styles of punctuation marks have been used in Amharic documents. For instance, for double quotation mark two successive single quotation marks (e.g., ' ') or similar symbols (e.g., ‹‹, ››, ``, « or ») are used; for end-of-sentence punctuation (። "Amharic full stop") two successive Amharic word separator (፡) that give the same appearance are used. Thus, normalization of punctuation marks is a non-trivial matter. We normalized all types of double quotations by ", all single quotations by ', question marks (e.g., ？ and ፧) by ?, old-forms of word separators (e.g., : and ፡) by plain space, full stops (e.g., :: and ፡፡) by ።, exclamation marks (e.g., ! and ！) by !, hyphens (e.g., :-, and ፡—) by ፡-, and commas (e.g., ፥ and ÷) by ፣. A similar approach is followed for normalization of English punctuations.

In Amharic script, there is no case difference. However, four of Amharic phonemes have one or more homophonic script representations and there are other peculiar labiovelars (e.g., ቍ /k'^{w}/, ጒ /g^{w}i/, and ጔ /g^{w}e/). In the modern-day Amharic writings, the homophonic characters are commonly observed to be used interchangeably and there is no uniform use of the peculiar labiovelars. For consistent spelling, the Ethiopian Languages Academy (ELA) proposed a spelling reform (ELA, 1970; Aklilu, 2004). Following their reform, homophonic characters are merged into their common forms: ሐ /ha/ and ኀ /ha/ are replaced with ሀ /ha/, ሠ /sə/ with ሰ /sə/, ዐ /ʔa/ with አ /ʔa/, and ፀ /s'ə/ with ጸ /s'ə/. The replacement also includes their variant forms. This process can be considered as case folding in English (Yacob, 2003). The peculiar labiovelars were normalized by substituting them with their closer counterparts (e.g., ቍ /k'^{w}/ with ቁ /k'u/).

5 Segmentation

Segmentation of sentences mainly involves disambiguation of end-of-sentence punctuation. To do so, punctuation marks are separated from word contents. The exceptions are abbreviations, initials of names, clitics, URLs, e-mail addresses, and hashtags since the punctuations are part of them. We created a list of known abbreviations and clitics for English which serves as negative wordlists for this process. End-of-sentence disambiguation facilitates automatic segmentation of sentences.

We considered isolated end-of-sentence punctuations (። for Amharic and period for English) and question marks as sentence boundaries. As end-of-sentence punctuations and question marks are stand out as the result of end-of-sentence disambiguation, sentence segmentation can be done automatically. However, there are ambiguities with direct speeches. Let's consider a pair of parallel-segments from the corpus as an example (see Figure 1). In the figure, from the linguistic point of view, the first segment should be segmented into two English sentences whereas the second one is a single Amharic sentence (the glosses of the Amharic sentence are given beneath the sentence). To avoid these kinds of ambiguities, end-of-sentence punctuations and question marks did not serve as sentence boundaries when they are used in direct speeches.

" Where did I go wrong ? " This question tormented Michael , from South Africa .

በደቡብ አፍሪካ የሚኖረው ማይክል " የተሳሳትኩት ነገር ምንድን ነው ? " የሚለው ጥያቄ እረፍት ይነሳዋል ።

(Gloss) *in-south Africa living Michael " wronged thing what is ? " saying question torment him .*

Figure 1: A pair of parallel segments.

6 Alignment

Since the selected documents were archived as EPUB and text files with corresponding file names for both languages, the document alignment was a trivial task. The challenge was in aligning sentences.

There are two main approaches to sentence alignment: length-based and lexical-based (Moore, 2002; Zaidan and Chowdhary, 2013). In length-based alignment approaches, the aligner relies on a probabilistic model that represents the source to target sentence length ratio for a pair of corresponding segments (sentences). Segment pairs that are likely under the length ratio model are taken as parallel segments.

In lexical-based alignment approaches, the aligner relies on a probabilistic model that describes the lexical similarity between a pair of segments. These approaches require external bilingual lexicons. Usually, their performance depends on the quality of the bilingual lexicons (Ma, 2006).

State-of-the-art aligners follow hybrid approaches. They rely both on sentence length and lexical similarity or word correspondences. Notable examples are Moor[4] (2002) and Hunalign[5] (Varga et al, 2005). While Hunalign optionally requires a bilingual lexicon, Moore's algorithm does not require any external resources, bilingual lexicon or parallel training data. Amharic words are highly inflectional and the available bilingual lexicons contain only lemmas of common words. With the absence of an efficient lemmatizer, there is no benefit in using these lexicons for alignment.

Selected Documents	**Number of Documents**	**Number of Segments**
Awake Magazine	53	16,491
The Bible	66	48,651
News Articles	421	7,710
Watchtower Magazine	124	72,968
Total	664	145,820

Table 1: The number of segments aligned in each bilingual document.

We tested Hunalign and Moor's aligner on the bilingual text from Awake magazine, without using any external resources. Hunalign was able to correctly align 15,025 segments whereas Moor's aligner aligned 16,491 segments. Since we have got better performance results with Moor's aligner, we used it to align segments in the remaining bilingual documents. The number of segments aligned in each bilingual document is given in Table 1.

7 Conclusion and Future Works

We collected, preprocessed, segmented and aligned 145,820 Amharic-English parallel sentences (segments) from various sources. In doing so we addressed language-specific issues such as normalization and end-of-sentence disambiguation. In addition, the corpus is larger in size than previously compiled corpora. It can be used to train Amharic-English machine translation algorithms. In the future, we plan to increase the size of the corpus by collecting text from scanned documents and documents with incompatible fonts by employing OCR systems and font conversion tools.

Acknowledgement

We would like to thank Geez Frontier Foundation for the news articles.

References

Aklilu, Amsalu. 2004. Sabean and Ge'ez symbols as a guideline for Amharic spelling reform. In *Proceedings of the first international symposium on Ethiopian philology*, pages 18-26.

Argaw, Atelach Alemu and Lars Asker. 2005. Web Mining for an Amharic-English Bilingual Corpus. In *Proceedings of the First International Conference on Web Information Systems and Technologies,* pages 239-246.

Chew, Peter A., Steve J. Verzi, Travis L. Bauer and Jonathan T. McClain. 2006. Evaluation of the Bible as a resource for cross-language information retrieval. In *Proceedings of the workshop on multilingual language resources and interoperability,* pages 68-74.

ELA. 1970. የአማርኛ ፡ ፊደል ፡ ሕግን ፡ አንዲጠብቅ ፡ ለማድረግ ፡ የተዘጋጀ ፡ ራፖር ፡ ማስታወሻ, (Engl. A memorandum for standardization of Amharic spelling) *Journal of Ethiopian Studies*, 8(1):119-134.

Gezmu, Andargachew Mekonnen, Andreas Nürnberger and Binyam Ephrem Seyoum. 2018. Portable Spelling Corrector for a Less-Resourced Language: Amharic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 4127-4132.

Ma, Xiaoyi. 2006. Champollion: A robust parallel text sentence aligner. In *Fifth International Conference on Language Resources and Evaluation,* pages 489-492.

Moore, Robert C. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas,* pages 135-144.

Strassel, Stephanie and Jennifer Tracey. 2016. LORELEI Language Packs: Data, Tools, and

[4] The implementation is available at: https://www.microsoft.com/en-us/download/details.aspx?id=52608

[5] The implementation is available at: http://mokk.bme.hu/resources/hunalign/

Resources for Technology Development in Low Resource Languages. In *Tenth International Conference on Language Resources and Evaluation,* pages 3273-3280.

Tachbelie, Martha Yifiru, Solomon Teferra Abate, Laurent Besacier. 2014. Using different acoustic, lexical and language modeling units for ASR of an under-resourced language–Amharic. *Speech Communication*, 56:181-194.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of Recent Advances in Natural Language Processing*, pages 590-596.

Yacob, Daniel. 2003. Application of the Double Metaphone Algorithm to Amharic Orthography. In *The XVth International Conference of Ethiopian Studies*, pages 921-934.

Zaidan, Omar and Vishal Chowdhary. 2013. Evaluating (and Improving) Sentence Alignment under Noisy Conditions. In *Proceedings of the Eighth Workshop on Statistical Machine Translation,* pages 484-493.