English – Oromo Machine Translation: An Experiment Using a Statistical Approach

SisayAdugna, Andreas Eisele
Haramaya University, DFKI GmbH
Ethiopia, Germany
Email: sisayie@gmail.com, eisele@dfki.de

Abstract

This paper deals with translation of English documents to Oromo using statistical methods. Whereas English is the lingua franca of online information, Oromo, despite its relative wide distribution within Ethiopia and neighbouring countries like Kenya and Somalia, is one of the most resource scarce languages. The paper has two main goals: one is to test how far we can go with the available limited parallel corpus for the English – Oromo language pair and the applicability of existing Statistical Machine Translation (SMT) systems on this language pair. The second goal is to analyze the output of the system with the objective of identifying the challenges that need to be tackled. Since the language is resource scarce as mentioned above, we cannot get as many parallel documents as we want for the experiment. However, using a limited corpus of 20,000 bilingual sentences and 62, 300 monolingual sentences, translation accuracy in terms of BLEU Score of 17.74% was achieved.

1. Introduction

According to the Central Statistical Authority (CSA), in 2002, Oromo is the official language of the Oromiya region of Ethiopia that has a population of over 24 million. Since Oromo writing in Latin script began only in 1991 (Tilahun, 1993), there is an insufficient number of documents prepared in this language.

The Internet contains abundant amounts of useful documents in English, which are however inaccessible for most of the Oromo speakers due to lack of knowledge of the English language. Therefore, translation of documents from English to Oromo is necessary for making these useful online documents accessible for local use. Thus, the focus of this paper is on testing the possibility of automatic or machine translation from English to Oromo.

Studying how to make these documents available in local languages (such as Oromo) is vital in addressing the language barrier thereby reducing the effect of the digital divide.

2. Experiment

The architecture of the English – Oromo SMT system that is shown diagrammatically in Figure 2.1 includes the four basic components of SMT: Language Modeling, Translation Modeling, Decoding and Evaluation.

The Language Modeling component takes the monolingual corpus and produces the language model for the target language. The Translation Modeling component takes the part of the bilingual corpus as input and produces the translation model for the given language pairs. The Decoding component takes the language model, translation model and the source text to search and produce the best translation of the given text. The Evaluation component of the system takes the system output and the reference translation and compares them according to some metric of textual similarity like the BLEU score.

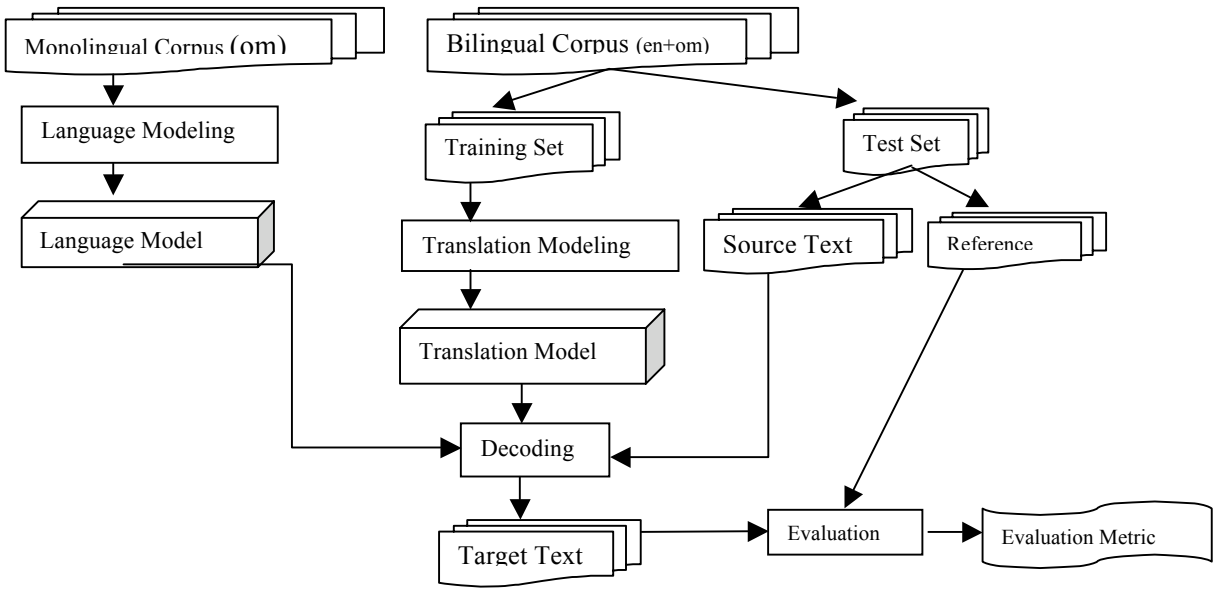

Figure 2.1 Architecture of the English – Oromo SMT System

2.1 Experimental Setup

2.1.1 Collection and preparation of the data

The parallel documents that were used are: Oromo versions of some chapters of the Bible, i.e., all the chapters of the Bible that are available in both English and Oromo as well as some spiritual manuscripts for which the English counterparts were available on the web, the United Nation's Declaration of Human Rights, the Kenyan Refugee Act, the Ethiopian Constitution, some medical documents, the proclamations and the regional constitution of Oromia Regional State.

When we try to inspect the output of the sentence-aligner for the Oromo documents, we noticed that some words are split into three tokens (one to the left of the apostrophe, the apostrophe itself, and one to the right of the apostrophe). This is due to the fact that, the sentence-aligner does not consider a word having " ' " (apostrophe) as one word. So, we have to modify the preprocessor in such a way that it should not separate a word into three pieces, rather as a word having the apostrophe as a character.

In Oromo, when this symbol comes at the end or beginning of a word, it is used as a single quote. Otherwise, it is used to represent a sound called 'Hudhaa' that should be dealt with at preprocessing. Here, if "'" (apostrophe) appears to be within a word, the tokenizer should not consider it to stand by itself rather it should keep the characters to left of it, itself and the character to the right of it as one token.

Though not exhaustive enough, list of the abbreviations for Oromo that is used for tokenization and sentence alignment was prepared manually.

2.1.2 Size of the Data

We used 62,300 sentences (1,024,156 words) of monolingual corpus (including the half part of the bilingual corpus - the Oromo part) to train the Language Modeling subsystem. Bilingual corpus of 21,085 sentences (384,881 words) of English and 20,848 sentences (308,051 words) of Oromo was used to build the initial translation model of the system. From this, 90% of the bilingual corpus has been used for training and the remaining 10% has been used for testing the system.

2.1.3 Software Tools Used

The widely used language modeling tool SRILM toolkit (Stolcke, 2002) was used for language modeling because the Moses MT system has a support for SRILM as a language modelling tool. For word-alignment, the state-of-the-art tool is GIZA++ (Och and Ney, 2003), which implements the word alignment methods IBM1 to IBM5 and HMM. Sentence alignment was done using a script that is available from http://www.statmt.org. Decoding was done using Moses (Koehn et. al, 2007), which is an SMT system that is also used to train translation models to produce phrase tables. Evaluation was done using the BLEU (Papineni et al., 2002) scoring tool.

3. Result and Discussion

From the test of up to 9-gram, the overall BLEU score of 5.06% was found initially. A closer look at the source, reference and target documents sentence-by-sentence showed on the raw test set that the sentences are not really aligned due to the impurity of the data. That is, as it was not prepared for this particular purpose, it was not as pure as it is needed for the system and one sentence in one language may be equivalent to more than one sentence in another. This will lead the evaluation system to the comparison of sentences which are not meanings of one another while their equivalents are also compared with some another sentences. For example, in the following screenshot, the output sentence is the translation of the source sentence except the repetition of the word "kaffaltii" in the output. However, it was compared with the 'reference sentence' wrongly assumed to be the human translation of the source sentence.

In order to eliminate problems like this, we tried to align the source and the reference sentences before running the system on the test set again. The BLEU score found after manually correcting the sentence alignment of the test data was improved from 5.06% to 17.74%. Individually seen, the highest BLEU score of 43.96% is observed for 1-gram scoring. However, the n-gram score sharply drops as n increases, i.e., the n-gram score for values of n equals 1, 2, 3, 4, 5, 6, 7, 8 and 9 is observed to be 43.96%, 21.57%, 14.42%, 10.72%, 8.04%, 5.52%, 3.76%, 2.23% and 1.30% respectively as shown in the graph in Figure 4.2..

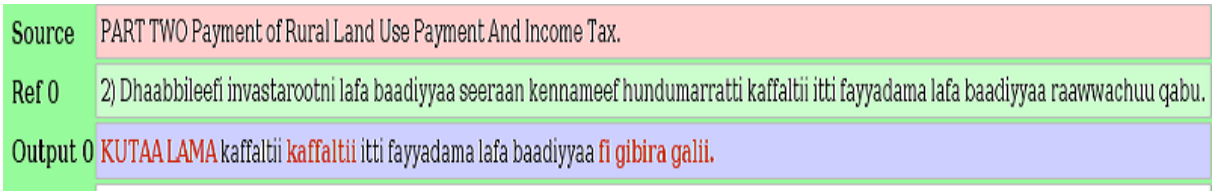

Figure 4.1 Comparing output with incomparable reference translation

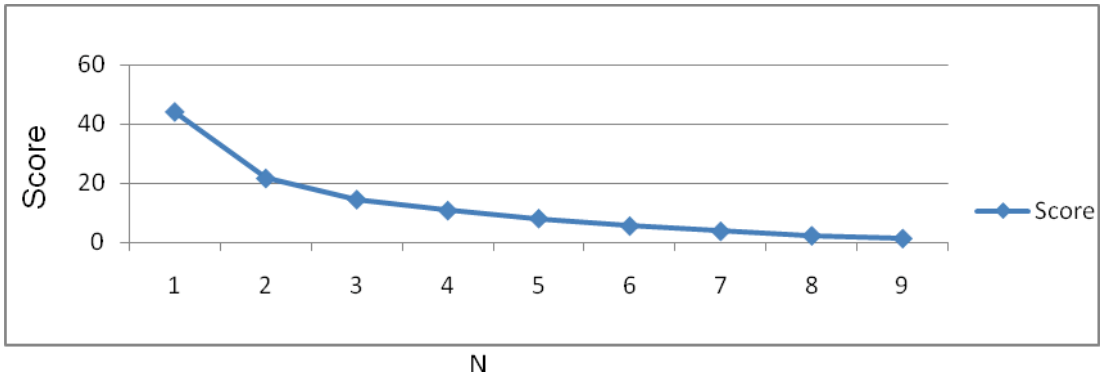

Figure 4.2 Individual N-gram Scoring

In addition to the limited size of the training corpus, the overall BLEU score of the system is attributed to the following major reasons:

3.1. Availability of a single reference translation

A source sentence can have many possible translations. A very good MT output might look like one human translation, but very different from the other one. Thus, automatic evaluation metrics judge MT output by comparing with multiple human translations and taking the average. However, in this experiment, there is only one human translation for the given source sentences. As a result, a little deviation of the MT output from this single human translation will count against the score of this system, which could not have been the case if there were many human translations.

3.2. Domain of the test data

The test data is composed mainly of three major domains – legal, medical, and religious. The domain diversity of the test data has affected the BLEU score significantly. As the majority of our data is from the religious domain, upon evaluating the output, we have investigated the bias introduced by the training data from religious domain. That is, the system performs better if it is tested on religious documents than documents from other domain. When seen separately, the BLEU score for the test data from the legal, medical, and religious domains are 13.69%, 1.97%, and 21.72% respectively.

4. Conclusions

Although Oromo is among resource-scarce languages (Kula et. al., 2008), the result of this experiment, which is an average BLEU score of 17.74%, shows that the amount of data available can be used as a good starting point to build SMT system from English to Oromo.

Despite the fact that there is not any existing MT system for the English to Oromo language pair with which one can compare the result of this system, we believe that comparing the score with other existing systems' scores for other language pairs will enable one to judge at least somewhat the level of achievement. For instance, in the shared task of the last WMT workshop, the best reported BLEU scores for several language pairs were lower than 17%, and still the performance of these systems is regarded as promising. Thus, one can conclude that our system performed not too low as compared to the systems built on a relatively sufficient amount of resource.

5. Next Steps

We consider the current system as a first step towards a more systematic construction of linguistic resources. In particular, we imagine that the output of our preliminary MT system should be made available in a form that users can post-edit the results and feed the result back into the MT system, thereby contributing to the growth of parallel corpora for this language pair. As a similar setup is under development within the EuroMatrix Plus project, it should be straightforward to apply it also to translation between English and Oromo. We will furthermore consider the collection and use of comparable corpora, from which we could draw additional lexical coverage. Such methods are currently investigated in several European research projects (see e.g. Eisele & Xu, 2010), and it will be interesting to see how the results can be ported to a language pair involving Oromo. Finally, some additional coverage might be gained by building linguistic models of Oromo morphology in a suitable finite-state formalism.

In the long run, this may lead to a situation in which the lack of suitable training data for this language pair is not as dramatic as it is today.

6. Acknowledgment

This work was supported materially by the EU projects EuroMatrix and EuroMatrix Plus via Saarland University and DFKI GmbH, and financially by Addis Ababa University and German Academic Exchange Service (DAAD).

7. References

_____. (2002). Statistical Abstract. Central Statistical Authority, Federal Democratic Republic of Ethiopia

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In Proceedings of the Fourth Workshop on Statistical Machine Translation. 1–28. Athens, Greece.

Eisele, A. & Xu, J. (2010). Improving Machine

Translation Performance Using Comparable Corpora. In Proceedings of 3rd Workshop on Building and Using Comparable Corpora. Valletta, Malta.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Proc. of ACL Demo and Poster Sessions. 177–180.

Kula, K.,Varma, V. and Pingali, P. (2008). Evaluation of Oromo-English Cross-Language Information Retrieval, IIIT, Hyderabad, India.

Och, F. J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19–51.

Papineni, K., S. Roukos, T. Ward, & W. Zhu. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. ACL

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing, vol 2, pp 901-904. Denver, Colorado, USA.

Tilahun, G. (1993). QubeAffan Oromo: Reasons for Choosing the Latin Script for Developing an Oromo Alphabet. The Journal of Oromo Studies 1(1)