PRELIMINARY EXPERIMENTS ON ENGLISH-AMHARIC STATISTICAL MACHINE TRANSLATION

Mulu Gebreegziabher Teshome[1], *Laurent Besacier (Prof.)*[2]
(1) Addis Ababa, Ethiopia: IT Doctoral Program, Addis Ababa University
(2) Grenoble, France: University Joseph Fourier

ABSTRACT

This paper discusses the preliminary experiment conducted to translate from English to Amharic using the Statistical Machine Translation (EASMT) approach. The experiment on the EASMT system is being conducted on training corpus of both languages based on expressions that are found in parallel documents. The experiment involves collecting of a total of 632 Parliamentary corpora of which 115 have been used in the experiment. The corpus coverage is 15 years from Aug 21, 1995 to July 16, 2010. Each document contains data, which are translations of each other. The experiment has been conducted using 18,432 English-Amharic sentence pairs extracted from these corpora in order to measure the accuracy of the translation system. Accordingly, the baseline phrase-based BLEU score result is 35.32%. A 0.34% increase in BLEU has been achieved by applying morpheme segmentation to the tokens of the Amharic output result and the reference of the baseline system. The increase is 0.92% when compared with the same segmented reference between the baseline and the segmented system.

Index Terms— Statistical Machine Translation, Parallel Corpus, Word Segmentation

1. INTRODUCTION

Amharic is a Semitic language spoken in Ethiopia. It is the official language of the Federal Democratic Republic of Ethiopia which is also the second most-spoken Semitic language in the world, after Arabic. Ethiopia is the second most populous country of Africa having more than 73.75 million inhabitants according to 2007 Census. Based on a projection of annual growth rate of 2.6 percent, Ethiopia's population up to July 2011 is more than 82 million and Amharic is spoken as a first language by more than 27 million people of the country [1].

Corpus based approaches to Machine Translation (MT) have been on the rise especially for languages such as Amharic which is considered as one of the Natural Language Processing (NLP) scarce resource language. Previous researches towards developing English Amharic NLP systems in general and MT researches in particular indicate that there is a need to start empirical researches for Amharic NLP.

The challenge to develop MT using rule-based approach to Amharic, which is considered as one of the NLP scarce resource language, is enormous. The same might not be true for well developed NLP resourced languages. What makes it challenging for under resourced languages is that the rule-based MT heavily employs integrated linguistic knowledge, rules and resources of both the source and target languages. The linguistic knowledge includes tagging, parsing, morphology, syntactic, semantic, and lexical knowledge of the source and target languages. The linguistic rules comprise rules for analyzing, transferring (including syntactic, semantic & lexical), and generating the source and/or target languages. The transfer rules may include translation of idioms and word sense disambiguation. The linguistic knowledge and rules employ one or more linguistic resources that include rich bilingual dictionaries, taggers, parsers, and Treebank [2] [3]. However, it is almost impossible to develop a MT system using the rule-base method for Amharic at least in the near future as it is under resourced language with respect to the different linguistic knowledge, rules and resources.

The statistical approach on the other hand relies heavily on bilingual parallel aligned corpora of the source and target languages. The challenge is minimized since the statistics based approach requires very limited computational linguistic resources compared to the rule-based approach that might take so many years to develop some or all of the mentioned linguistic resources.

Thus, the preliminary experiment on the EASMT system is being conducted based on sentences that are found in parallel Amharic-English Parliamentary corpus.

2. EXPERIMENTAL SETTINGS

SMT systems require training on bilingual corpora. It is critical to develop a bilingual English-Amharic corpus by using automatic acquisition of available corpora from the web or by customizing available resources. The corpus associates probabilities with translations empirically by counting co-occurrences in the data. Estimates of

probabilities get more accurate as the size of data increases [2] [3]. Most translation systems use parallel corpus for training data from constitutions called Hansard Corpus [2] [3]. Similarly, the English-Amharic parallel corpus from parliamentary documents that exist online including those collected manually are used for the preliminary experiment on EASMT.

2.1. Preprocessing the parallel corpus

The first step involves collecting of raw English-Amharic corpus from the Parliament of the Federal Democratic Republic of Ethiopia. The parliamentary corpora coverage is from Aug 21, 1995 to July 16, 2010 and is published under the Federal Negarit Gazeta. A total of 115 parliamentary corpora have been processed out of the 632 collected raw corpora. Table 1 summarizes counts in terms of the aligned documents, sentences, tokens and vocabularies of the parallel corpus.

Collected		**Amharic**	**English**	**Total**
Aligned	Documents	115	115	230
	Sentences	19,115	25,730	44,845
	Tokens	219,430	283,578	503,008
	Vocabularies	32,299	17,908	50,207

Table 1: Counts of Parliamentary English-Amharic parallel proclamation corpus

Some common features of each corpus are that they are readily available in PDF format. Each corpus contains information related to the government, title of the Gazette, notes, content description, publisher, date and place of publication, either in Amharic or English or both. Furthermore, there is also pagination, header and footer content attached to all Gazeta. The full content of the proclamation, which is located in the center of the Gazette, is divided into two columns. The left side column is for Amharic and the right side column is for the translation of the text in English.

A pre-processing task is performed on each corpus in order to retain and convert the full content into a valid format suitable for the EASMT system. Some of these pre-processes include text conversion, trimming, sentence splitting, sentence aligning and tokenization. The process of trimming is performed before and after aligning at document level. The sentence splitting has been done before starting aligning at sentence level while tokenization is performed after aligning at the sentence level.

The first pre-process is to convert the corpus from PDF to RTF then to Unicode text. The number of successfully converted corpora from the total 632 is 115. The low turnout is due to some of the oldest Gazeta, which are saved as jpg image formats. Aligning into Amharic and English is already done as both are incorporated on the same page. The most challenging task was converting from the RTF to Unicode text file. This is because each corpus can have at least 8 different Amharic fonts. The good quality converter tool found at the time was the Power Geez 2010 version. The Amharic fonts recognized by the converter include: VG Geez Numbers, VG2000 main, Ge'ez-1, Ge'ez-1 Numbers, Ge'ez-2, Ge'ez-3, VG2 main, VG2 Title, Visual Geez Unicode, Visual Geez Unicode Title, and Power Geez Unicode1. If a word contains more than two fonts during conversion, then the converter automatically converts the word using the first encountered font. The words with other fonts will contain weird characters after the automatic conversion of the full document is complete. As a result, it takes time to manually select and reconvert those words that have been wrongly converted.

Trimming has been performed by removing any part of the corpus except the text that contains the full content of the proclamation. After automatically trimming the corpora, the process of splitting each paragraph into sentences using sentence endings is performed. The Amharic sentence endings and punctuations have been converted to English to make it easy to apply similar pre-processing tools used for English. The converted Amharic punctuations include the Ethiopic comma (፣), colon (፡), semi-colon (፤) and full stop (።) to their English counterparts (',', ':', ';', '.') respectively.

The alignment at the sentence level has been done using a sentence aligner called Hunalign similar to [4]. Hunalign aligns bilingual text at sentence level using sentence-length information. In the simplest case, its output is a sequence of bilingual sentence pairs. In the presence of a dictionary, it combines this information with sentence-length information [4]. A small English-Amharic bilingual dictionary, which is adopted from [5], of word lists sized 8,212 have been used. The aligner was able to align 19,115 Amharic sentences and 25,730 English sentences.

2.2. Size of the parallel corpus

The aligner was able to align 18,434 English Sentence to Amharic sentences in 0-1, 1-1, or 1-2. Those sentences that do not have matching translations (0-1 or 1-0) have been dropped. Those sentence pairs with more than 200 tokens in length have been dropped as well in order to get a better performance of the decoder. As a result, 18,432 English sentences aligned with Amharic sentences have been retained and used for this experiment. That is why the count of sentences before doing the alignment at sentence level is much higher than that of after the alignment.

Out of the total collected data, 90% or 16,432 randomly selected sentence pairs have been used for training while the remaining 10% or 2,000 sentence pairs are used for tuning and testing. Thus, the preliminary experiment is developed using a total of 18,432 English-Amharic bilingual parallel and 254,649 monolingual corpora. The monolingual corpus is used for the Language Modeling (LM). The LM

corpus contains those data related to parliamentary documents that are not included in the bilingual corpora and news items collected from the Ethiopian News Agency. Table 2 provides statistics of the detail counts at sentence, token and vocabulary level used in the preliminary experiment.

Set	Language	Sentences	Tokens	Vocabulary
Train	English Amharic	16,432 16,432	377,542 292,742	10,941 24,295
Dev	English Amharic	1,000 1,000	22,470 17,303	3,072 5,006
Test	English Amharic	1,000 1,000	23,054 17,635	3,172 5,172
LM	Amharic	254,649	4,874,713	220,723

Table 2: Training, development and test data set counts

2.3. Software resources

The preliminary experiment is built upon previously developed software resources that are readily available to the research community to aid research in statistical machine translation. MOSES is one of such tools. MOSES is an open source statistical machine translation system that allows researchers to automatically train translation models for any language pair [6]. Moses is a decoder more suitable for phrase based SMT. In addition to the Moses tools, Perl scripts have been developed for pre-processing purposes.

The other resources used are statistical analysis toolkits such as word alignment, language modeling (LM), translation modeling, and evaluation. The toolkit used to build the language model is SRILM [9]. Whereas to build the statistical translation model, an open source Giza++ toolkit as well as some scripts that are with the MOSES suite are used [10]. The BLEU metric is used to evaluate the performance of the test result. All mentioned toolkits are integrated with MOSES. These toolkits are used by many researches, as research shows they have proved successful [2] [6]. The basic architecture of the EASMT in Figure 1 shows the major components of the system and indicates where each toolkit fits during translation.

3. RESULTS AND DISCUSSION

3.1. Phrase-based translation

The EASMT system has been trained using the English-Amharic parallel Training Set as translation examples and tested using the English Source Text as new sentences that gives an output Target Text of translated Amharic sentences (Figure 1). The output then has been scored using the BLEU evaluation metrics. Accordingly, the baseline phrase-based BLEU score result indicates that 35.32% translation BLEU score has been achieved.

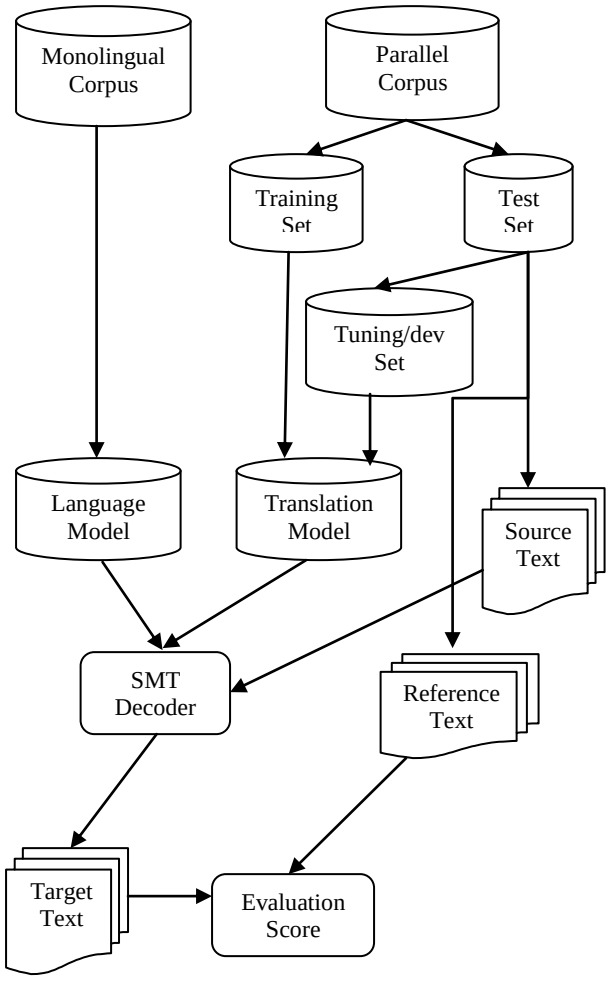


Figure 1: The basic architecture of the EASMT

	Source English Text	Target Amharic Translation	Reference
Date	6th day of June, 2000	ግንቦት 29 ቀን 1992	ግንቦት 29 ቀን 1992
Geez number	7	፯	7
synonym	rectification Any person fraud	የሚታረሙበትን ማንኛውም ሰው በማጭበርበር	እንዲስተካከሉ ማናቸውም ሰው በማታለል
redundancy	association apply government	ማህበር ተፈፃሚ መንግሥት	ማኅበራት ተፈጻሚ መንግስት
Direct	fund	ፈንድ	የገንዘብ ምንጮች
Insertion	general manager	የአስተዳደሩ ሥራ አስኪያጅ	የዋናው ሥራ አስኪያጅ

Table 3: Sample translated English text to Amharic

The preliminary experiment result shows that the EASMT has the ability to translate the basic meaning of the English sentence when translating into Amharic sentence. However, there are some strong as well as weak points in performance of the EASMT.

The EASMT performed well in respect to handling date information, Geez numbers, synonyms, syllable redundancy, direct translation and insertion (Table 3). It correctly translates the Gregorian Calendar (G.C.) date formats to Ethiopian Calendar (E.C.). There is a date, month and year difference between the E.C. and G.C and one needs a converter to get the exact date information. As indicated in Table 3, the date information is translated correctly as if it was done using a converter. A converter may subtract seven from the date information and adjusts the month accordingly. Additionally, it was able to translate the year information correctly by subtracting the year difference, which is eight. So, the date in G.C "6th day of June, 2000" has been correctly translated to "May 29, 1992" to E.C. Another good result is the Arabic numerals or Geez numbers that are used interchangeably in Amharic writing. The translated numbers are in one of the two number systems. If the translated number is in Geez, it is translated correctly as shown in Table 3. Another good performance of the EASMT is with regard to synonym. The translated synonym text is different but has identical meaning as the Reference text, which is the human translation of the Source text. With regard to redundancy, there are some Amharic syllables (e.g. 'ጸ' and 'ፀ') that represent the same sound that can be used interchangeably [7]. The complete list of the Amharic redundant syllables is taken from the Amharic IPA notation table (Table 4). Finally, the direct translation is also another strong point where some words are translated to Amharic exactly as the English but written using the Amharic syllables. This is useful to get translations of the English text that do not have equivalent translation or that are not commonly used Amharic words like this example. The last was the insertion where a word has been added that is not in the Reference. In this example, "the Administrator" is inserted into the translated Amharic text to get "the Administrator general manager". The text "the Administrator" can be used optionally some time.

Some weaknesses of the EASMT include non translated words, wrongly translated, insertion, deletion, alignment problem, preposition usage, and morphological errors. To address some of these problems, we have used word segmentation [8] on the Target side, which is the Amharic.

3.2. Using word segmented corpus

Two types of EASMT called segmented and un-segmented systems have been developed. The un-segmented system is the baseline system normally trained, developed and evaluated without using the segmenter [8]. The segmented system has been developed by segmenting all Target texts. These include the Amharic sentences of the Training Set, the Test Set (Tuning Set and Reference), the output Target text test result and the language model (Figure 1). The segmentation was performed on the Geez syllables without transliterating to Latin characters. This was done with little modification by making the segmenter compatible to Unicode characters.

The segmenter was able to segment a word into smaller morphemes. For example, the Amharic word "በተፈጥሮአዊ" is segmented in to three morphemes "በ", "ተፈጥሮ" and "አዊ" each correspond to a preposition, noun, and adjective marker respectively. The segmenter was able to segment prefixes and suffixes. The following words "ሀላፊነትና, ሀላፊነትን, በሀላፊነት, በሀላፊነትና, ከሀላፊነት, ከሀላፊነትና" were all segmented to "ሀላፊነት" by removing the prefix, suffix or both. The segmentation has contributed a lot to the vocabulary size and the percentage decrease is 22%, 5% and 19% for the training, development and test set respectively (Table 6).

	*	ə	u	i	a	e	o	ɨ	ʷi	ʷa	ʷe	ʷ
*			ኡ ዑ	ኢ ዒ	አ ኣ ዐ ዓ	ኤ ዔ	እ ዕ					
g							ጎ ጐ					
h	ሀ ሐ ኀ ኸ		ሁ ሑ ኁ ኹ	ሂ ሒ ኂ ኺ	ሀ ሃ ሐ ሓ ኀ ኃ ኻ	ሄ ሔ ኄ ኼ	ሆ ሖ ኆ ኾ		ኈ ኹ	ሏ ኋ ኻ	ኌ ኼ	ኍ ኾ
k							ከ ኰ					
k'							ቆ ቈ					
q							ቐ ቒ					
s	ሥ ስ	ሠ ሰ	ሡ ሱ	ሢ ሲ	ሣ ሳ	ሤ ሴ	ሦ ሶ			ሧ ሷ		
s'	ጽ ፅ	ጸ ፀ	ጹ ፁ	ጺ ፂ	ጻ ፃ	ጼ ፄ	ጸ ፆ					

Table 4: Redundant Amharic Syllables in IPA notation

System / Reference	Un-segmented	Segmented
Un-segmented	35.32	-
Segmented	35.66	36.58

Table 5: BLEU test score for Segmented and Un-Segmented System

In Table 5, the outputs of both systems have been scored with two types of reference: segmented and unsegmented. In the case of the segmented reference and the un-segmented system, the output of the system was re-segmented to be made consistent with the reference. However, for the dual case (segmented system and unsegmented reference) we were not able to re-stick (unsegment) the system output to make it compatible with the unsegmented reference. We see that the vocabulary size reduction has contributed for the overall performance of the segmented system that has shown better performance compared to the baseline phrase-based system. When compared with the same segmented reference (fair comparison), the BLEU score for the segmented system is 36.58%, which is a 0.92% increase from the baseline system that has a BLEU score 35.66%. It also showed an improvement by 0.34% by segmenting the output result and reference of the unsegmented system.

Set	Tokens		Vocabulary	
	Un-Seg	Seg	Un-Seg	Seg
Train	292,742	307,422	24,295	18,882
Development	17,303	18,141	5,006	4,770
Test	17,635	18,618	5,172	4,205
LM	4,874,713	5,147,719	220,723	101,240

Table 6: Counts for segmented and un-segmented training, development, test and LM data sets

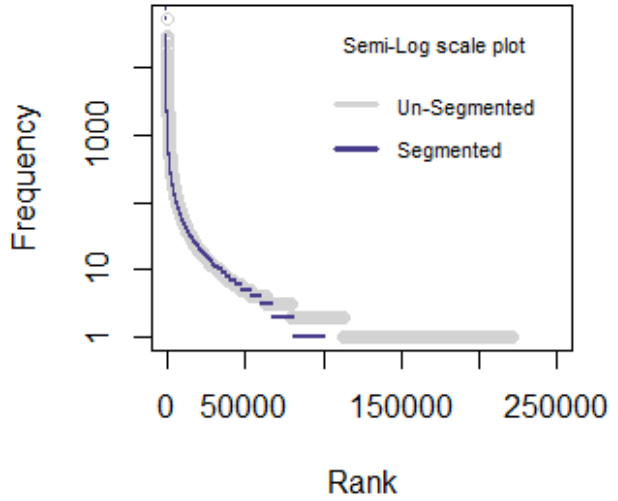
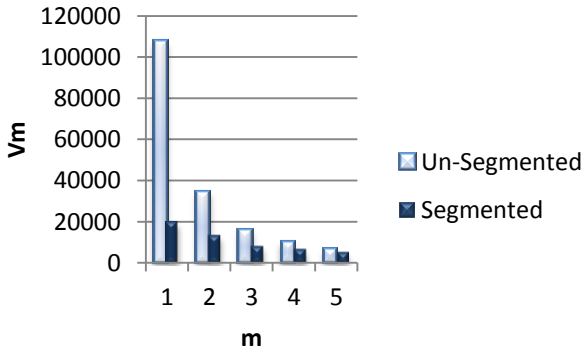


Figure 2: distribution of ranks among frequencies

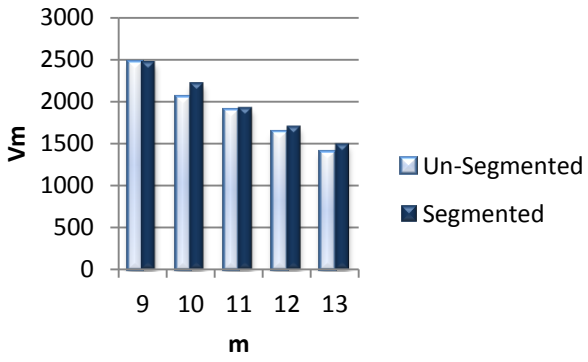
The long tail in Figure 2 shows that the lowest most semi-log plot is messy due to the data sparseness problem for the un-segmented than that of the segmented Amharic LM tokens. This visibly indicates that there are more tokens (Vm) that appear only once (m) in the unsegmented than in the segmented corpus (Figure 3a). The symbol V is for vocabulary size (number of types), and Vm is the vocabulary size that have frequency m. Specifically, at V1 the number of vocabulary types that occur only once in the un-segmented and segmented corpus are 107,304 and 19,593 respectively. At V2, the numbers of vocabulary types that occur twice are 34,122 and 13,132 for the unsegmented and segmented corpus respectively.

The variation is very high at V1 and it becomes less and less as we move further until V9. From this point on wards, the vocabulary size for the segmented corpus is either greater than or equal to but not less than that of the unsegmented corpus (Figure 3b). The data sparseness problem has significantly improved as the vocabulary size decreases.

The vocabulary size has played a major role for the improved performance of the segmented system. Reduction of the vocabulary size has improved the frequency of tokens and the performance of the EASMT system. To some extent, the drawbacks stated under section 3.1 have been solved. These includes duplication, non-translated words, wrongly translated, preposition usage, and morphological errors. However, some of the segmented affixes wrongly showed up in the output.



a. The vocabulary types with frequency range 1-5



b. The vocabulary types with frequency range 9-13

Figure 3: summary of a frequency distribution in terms of Vocabulary types (Vm) that have frequency m

4. CONCLUSION AND FURTHER WORK

According to these results, more experimentation and research is required to further improve the translation accuracy of the EASMT. The experiment done so far is encouraging as the translation is done from less inflected English language to Amharic, which is a morphologically rich language. Collecting and processing the corpora has been the most challenging. The major challenge was to get a bilingual corpus as Amharic is one of the languages that suffer severely from lack of computational linguistic resources. Therefore, it can be concluded that the research can be further conducted using these corpus.

The other major hurdle to use freely available resources used for other languages is due to the characteristics of the Amharic writing system that uses the Geez fonts. It was almost difficult and time consuming to adopt the existing tools. Such tools have been developed with the intension to use them for the major European languages that have more or less similar writing system and punctuations. The tools also demand some other related resources such as dictionaries, morphology analyzers, grammar, parsers, taggers, xml-based corpus, spell checkers, etc. Such resources are almost absent in Amharic.

Thus, most of the preprocessing steps have been done by writing small scripts using Perl. The developed scripts have been used for counting, trimming, splitting paragraphs to sentences and aligning at document level. Adopted tools have been used to align at the sentence level.

Finally, it is important to put a roadmap for the next experiment. There is a need to further look at the analysis of the experiment more closely and to perform the next level experiment using Moses. The task to be performed includes improving the translation quality by applying morphological and syntactical linguistic knowledge.

5. ACKNOWLEDGEMENTS

The authors thank the Laboratoire d'Informatique de Grenoble - LIG of France that covered a round trip travel expense and Dr. Pedro Moreno from Google, who was responsible for arranging the Google grant funding to cover the registration and accommodation expenses.

6. REFERENCES

[1] CSA. The 2011 Ethiopia Statistical Abstract, Central Statistical Agency, Addis Ababa: Ethiopia, 2011.

[2] Daniel Jurafsky and James H. Martin, Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd ed. New Jersey: Prentice Hall, 2009.

[3] Christopher D. Manning and Hinrich Schutze, Foundations of statistical natural language processing. Cambridge: The MIT Press, 1999.

[4] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. Association for Computational Linguistics, 1993.

[5] Armbruster-Bender-Fulas-Merged-2007. http://nlp.amharic.org/resources/lexical/word-lists/verbs/ ArmbrusterVerbs-20070120.txt/

[6] Moses: a statistical machine translation system. http://www.statmt.org/moses, 2012.

[7] Daniel Yacob. "Developments Towards an Electronic Amharic Corpus", TALN 2005, Dourdan, 6-10 June 2005.

[8] Creutz, M. and Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

[9] A. Stolcke, "SRILM-an extensible language modeling toolkit", in J. H. L. Hansen and B. Pellom, editors, Proc. ICSLP, vol. 2, pp. 901-904, Denver, Sep. 2002.

[10] Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.