International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 30195-30201
© Research India Publications
http://www.ripublication.com

A Comparative Study of English To Kannada Baseline Machine Translation System With General and Bible Text Corpus

Shiva Kumar K M[1], Namitha B N[2], Nithya R[3]
Amrita VishwaVidyaPeetham, Mysore
Email:shivadvg19@yahoo.com ,nachappanamitha95@gmail.com ,nithya.rnair7@gmail.com

Abstract

In this paper we present the insights gained from a detailed study of Kannada-English Statistical machine translation system with reference to corpus creation. We propose approaches to create a quality corpus which can enhance class categories in translation modelling so that we can get improved machine translation.

Statistical machine translation (SMT) is an approach to MT that is characterized by the use of machine learning methods. The accuracy of these systems depends crucially on the quantity, and domain pf the data. In SMT system data is pre-processed consistently. The agglutinative and morphologically rich Indian language require a huge amount of corpus creation because SMT treats morphological variants of a word as a separate to kenrather than a related token. So we need to create related words and sentences as unique entries in a corpus. Working with English-Kannada language pair with a small data set of 2500 sentences and a big openly available Bible corpus we show that the impact of token types and their frequency plays a major role in improving BLEU score of our Baseline MT System. We report comparative result of experiments conducted on these two corpus for English to Kannada Baseline MT System.

Key Word: LM, BLEU SCORE, MORPHOTACTICS, BASE LINE, GIZA++, SMT.

Introduction

India is a multilingual country with Kannada as an official language spoken and used in Karnataka and other parts of the country. Kannada is being spoken and used by around 6croses of people in India. We found a rich literature written in Kannada since 4000 years.

Machine translation is the process of translating words from source language to target language by the computer system. In India this translation is of great importance as India has 18 officially recognized languages those are Assamese, Bengali, English, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. Clearly, India owns the language diversity problem. In internet age, multiplicity of language makes using resources on internet even more difficult and there arises necessity of translation. There are four main approaches to machine translation.

Literature Survey

Papers those are taken for reference describes the methods to analyze different approaches and work with languages which are having highly asymmetrical morphological structures with the use of different translation probabilities [1]. Technique which is mainly used is Statistical Machine Translation (SMT). In SMT a sentence can be translated from one language to another in many possible ways and this approach takes the views as every sentence in target language is possible translation of the input sentence but sentence with high probability is considered that is using probability distribution function p(e/f) where f is source language sentence and e is target language sentence[3]. Here language model (LM), translation model(TM) and decoders to systematically address the problem [2].Other than SMT we have different approaches. Approach one here studies on the use of morphological analysis and finding frequently used words in the language to identify the source language. Then using POS (Parts of Speech) to translate source language sentences to target language sentence but this approach is time consuming since manually words are collected and count is maintained to recognize the language [4]. In real time, SMT is used in (Google and Bing) free online translator tools with word to word translation and it keeps increasing language option but efficiency is not up to the mark when sentence to sentence translation is considered [5]. Literature shows that the rule based machine translation process is extremely time consuming, difficult and failed to analyze accurately a large corpus of unrestricted text. Approach two to increase the efficiency, combination of SMT and rule based approach is studied but both approaches are standalone and work in sequence to give results [6] but this approach is time consuming as they are standalone. They are more complex and benefit of using SMT may be ineffective as it is merged with rule-based approach. Approach three EBMT Example Based machine translation is based on the idea of reusing the already translated examples. Example based translation involves three major steps - Example acquisition, Matching and Recombination [7]. This approach is effective when we work with large set of data that is examples and will not give desired result when we opt for different set source sentences other than example sentences. Approach four using comparable corpora and PBSMT (Phrase Based SMT) but it is shown that restricting phrases to linguistic phrases or statistically motivated phrases decreases the quality of translation. To avoid this, processing of sentences have to be done with target morphological structure [8]. Bilingual dictionary with comparable corpora approach may cover large part of vocabulary but not morphology of language. To increase efficiency, large set of parallel corpus will have to be considered [9]. Factored Machine Translation Systems is an approach where a word has multiple representations in target language and with linguistic information that words are integrated with target language but Pre-processing of the corpora has to be done [10]. The motivation for using SMT is to take advantages of the robustness of the SMT system and the linguistic knowledge of morphological analysis and through system combination approach and usage of language model, translation model and decoder systematically address the problem with the usage of large volume of bilingual corpora will increase the efficiency of sentence to sentence translation compared to other approaches.

Experimental Setup

English is structurally classified as Subject-Verb-Object (SVO) language with a Kannada language is highly agglutinative and morphologically rich, Kannada follows subject- object -verb structure whereas English follows sub-verb-obj. Kannada sentences maintains the coherence factor with in the sentence from first word to last word using gender and case markers between subject and verb, this makes difficult to apply SMT for English Kannada language pair which we summarize in this section.

English- Kannada: the basic structural differences between English and Kannada is a large distance between subject and verb.

Compared to other Indian languages Kannada is morphologically richer with respect to inflection case and gender markers. Example: (ram ne market gayatha, sita ne market gayithi,) (rama nu marukattegehodanu, siteyumarukattegehodalu)- in this way Kannada has more inflections than other languages.

Textual coherence in Kannada

Consider the Kannada sentence:

□□□□□□□0□□□□□□□□□□□□□0□□□.

Transliteration:

[Annanu thangige baleyannu thandannu]

Translation:

[*The elder brother bought bangles to his sister]*

In the above Kannada sentence, the subject annanu is associated with to morphological entries (annanu,**nu**). The morphological inflection “nu” symbolises gender (masculine). In Kannada, the subject, object, the nouns are associated with grammatical information like person name gender (PNG).

In the above sentence, the inflection **nu** will repeat with the verb thandanu which is the last word of the sentence. Hence the inflectional suffix **nu** has its influence throughout the end of the sentence. Hence the coherence factor in Kannada sentence is more compared to English or Hindi sentences.

Consider the Hindi sentence:

□□□□□□□□□□□□□□□□□□□□□□ |
Bhayane behen keliye chudiya laaya
(Translation of Hindi sentence)
The elder brother bought bangles to his sister

In the above sentence, the inflection ne will not repeat with the verb **laya** which is the last word of the sentence. Hence the inflectional suffix ne does not have its influence throughout the end of the sentence. Hence the coherence factor in Hindi sentence is less compared to Kannada.

It is difficult to apply machine translation on Kannada sentences because of its morphological richness and grammatical structure. In this work we are proposing the comparative study of SMT with two different corpus in which one with tokens having higher frequency and other with very low frequency. The first corpus is a general corpus which comprises normal conversations. The second corpus is the publicly available bible corpus.

We process the Kannada corpus file through Indic tokenize which is available from IIT Bombay website, since Kannada tokens are highly agglutinative the tokenizer available in Moses tool kit will not normalize correctly.

English corpus normalization: English data file is tokenized and normalized using tokenizer. perl and truecase. perl provided in Moses toolkit.

We use the Moses toolkit for carrying out our SMT experiments. In this study we are proposing the base line system results on 20,000 parallel Kannada-English bible corpus and 3000 general text Kannada-English Parallel corpus.

Baseline source corpus : In English-Kannada baseline SMT the target language is Kannada , it requires more linguistic effort to change the English sentence word order and we don't have earlier research results with respect to BLEU score of English-Kannada SMT , we are retaining the source language and target language sentences in their normal form,.

Results

Corpus statistics: General corpus

Maximum sentence length: 10 words
Minimum sentence length: 3 words

Corpus Statistics

Bible Text corpus features

General Text Corpus Details

steps	sentence		words	
Training	Eng	Kan	Eng	Kan
	3000	3000	7195	9988
tuning	200	200	833	664
testing	50	50	336	249

Bible Text Corpus Details

steps	sentence		words	
Training	Eng	Kan	Eng	Kan
	18000	18000	4,56,040	2,79,168
tuning	1000	1000	29175	18200
testing	1000	1000	31707	19197

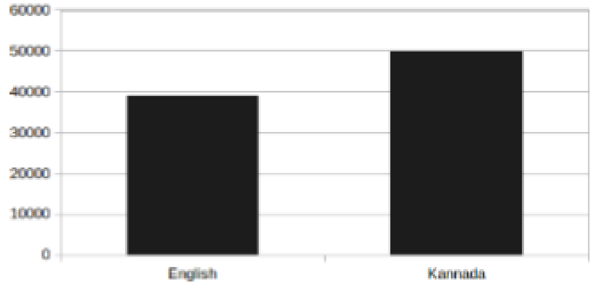
En-Kn MT Results

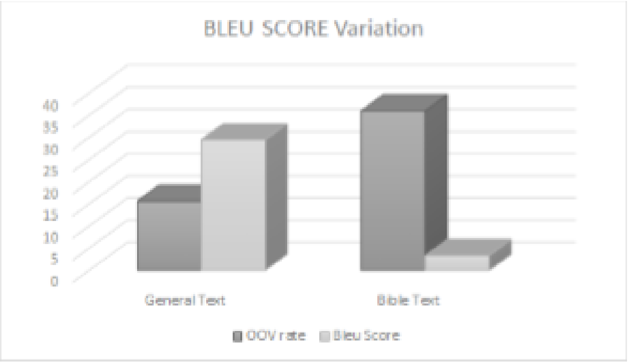
Corpus type	OOV Rate (%)	BLEU Score
General Text	15.6	29.47
Bible Text	36	3.36

Maximum sentence length: 25words
Minimum sentence length: 8 words

Language	No of unique words	Highest freq. words	Lowest freq. words	Total words
English	9,954	39,015	1	4,56,040
Kannada	3914	49,732	1	2,79,168

Frequency of Words





Conclusion

In English corpus we see more no. of repeated words in each class , thus the no. of tokens are less and because of high frequency of words there perplexity factor while assigning the probability. Kannada has more number of unique words and thus the number of tokens are more and frequency is less compared to English language, therefore perplexity factor will also be less.

Acknowledgment

We would like to give deep gratitude for the blessings of Amma and all those who have supported us to complete the implementations we have presented in this paper. Very special thanks to the Management, Principal and Head of department, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysore campus, for all their support & giving us the resources required completing the development work.

Reference

[1] Young-Su Lee “Morphological Analysis for Statistical Machine Translation” IBM T. J.Watson Research Center, Yorktown Heights, NY 10598.

[2] Nakul Sharma “English To Hindi Statistical Machine Translation System” Thapar university Patiala.

[3] Unnikrishnan P, Antony P J, Dr. Soman K P “A Novel Approach for English to South Dravidian Language Statistical Machine Translation System”.

[4] Bhojraj Singh Dhakar, Sitesh Kumar Sinha, Krishna Kumar Pandey “A Survey of Translation Quality of English to Hindi Online Translation Systems (Google and Bing)” International Journal of Scientific and Research Publications, Volume 3, Issue 1, January 2013.

[5] Arafat Ahsan, Prasanth Kolachina, Sudheer Kolachina, Dipti Misra Sharma, Rajeev Sangal “Coupling Statistical Machine Translation with Rule-based Transfer and Generation” LTRC, IIIT Hyderabad.

[6] Anju E S, Manoj Kumar K V “Malayalam To English Machine Translation: An EBMT System” Dept. of Computer Science & Engineering, Govt. Engineering College Thrissur.

[7] Mallamma V Reddy, Dr. M. Hanumanthappa “Natural Language Identification and Translation Tool for Natural Language Processing” Department of Computer Science and Applications, Bangalore University, Bangalore, INDIA.

[8] Raghavendra Udupa U, Hemanta K. Maji “Computational Complexity of Statistical Machine Translation” IBM India Research Lab New Delhi India, Dept. of Computer Science University of Illinois at Urbana-Champaigne.

[9] Rui Wang, Petya Osenova and Kiril Simov "Linguistically-Augmented Bulgarian-to-English Statistical Machine Translation Model" Language Technology Lab DFKI GmbH Saarbr ucken, Germany, Linguistic Modelling Department, IICT Bulgarian Academy of Sciences Sofia, Bulgaria

[10] Ann Irvine, Chris Callison-Burch "Combining Bilingual and Comparable Corpora for Low Resource Machine Translation" Center for Language and Speech Processing Johns Hopkins University ,Computer and Information Science Dept. University of Pennsylvania.

[11] St ephane Huet, Elena Manishina and Fabrice Lef`ever "Factored Machine Translation Systems for Russian-English" niversit e d'Avignon, LIA/CERI, France.

[12] Santanu Pal , Partha Pakray , Sudip Kumar Naskar "Automatic Building and Using Parallel Resources for SMT from Comparable Corpora" Universität Des Saarlandes, Saarbrücken,Germany, Computer & Information Science, Norwegian University of Science and Technology, Trondheim, Norway, Department of Computer Science & Engineering, Jadavpur University, Kolkata, India

[13] Philipp Koehn and Hieu Hoang. Factored Translation Models, Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic, June 2007.