Dictionary Based Machine Translation from Kannada to Telugu

Sindhu D V Sagar B M

Research Scholar
R.V College of Engineering, India.
Sindhu.489@gmail.com

Associate Professor
R.V College of Engineering, India.
Sagar.bm@gmail.com

Abstract: Machine Translation is a task of translating from one language to another language. For the languages with less linguistic resources like Kannada and Telugu Dictionary based approach is the best approach. This paper mainly focuses on Dictionary based machine translation for Kannada to Telugu. The proposed methodology uses dictionary for translating word by word without much correlation of semantics between them. The dictionary based machine translation process has the following sub process: Morph analyzer, dictionary, transliteration, transfer grammar and the morph generator.

As a part of this work bilingual dictionary with 8000 entries is developed and the suffix mapping table at the tag level is built. This system is tested for the children stories. In near future this system can be further improved by defining transfer grammar rules.

Keywords: Bilingual dictionary, Transliteration, Morph analyzer, Morph generator, suffix mapping.

1. Introduction

Machine Translation (MT) is the oldest problem of research in artificial intelligence and computer science. During 1980’s on invention of personal computers changed MT researchers view in handling the languages. On the arrival web during 1990s resulted in increased data on the internet. This rapid growth of data on web was inspiration for the MT researchers to develop more commercial MT systems to provide a global communication. The practical objective of the MT is to develop the systems which are available to the common people at the nominal cost to reduce the language barrier. Machine translation can be considered as an area of applied research that draws ideas and techniques from linguistics, computer science, artificial intelligence, translation theory, and statistics [2].
MT systems can be designed either specifically for two languages called a bilingual system or from one language to many languages called multilingual systems. Usually multilingual systems are bidirectional, but most bilingual systems are unidirectional.

Even though machine translation was envisioned as a computer application in the 1950’s and research has been made for 60 years, machine translation is still considered to be an open problem. However, up to now despite considerable efforts fully automatic high-quality translation is not a realistic goal for the foreseeable future. There are many approaches for developing the MT systems, each system has their own advantages and disadvantages. Out of these approaches Statistical Machine Translation is widely used [2]. Statistical Machine Translation (SMT) system derives the rules from a bilingual parallel corpus which contains large number of high quality aligned sentences. The biggest challenge is to get the high quality corpus because of insufficient sources of the data. Example Based Machine Translation (EBMT) approach is used when less number of linguistic resources is available for the languages. In the example based translation, a system is defined which contains set of source language sentences and corresponding target language sentences. During the run time, example based translation use bilingual corpus as its database. This database is stored in the translation memory. When the system encounters the same sentence the system need not to translate again system directly retrieves from the translation memory.

The advantage of the example based translation is the translation memory saves the user effort of re translating the sentence and this saves the processor time and also the user time. The disadvantage of example based translation is the wastage of memory to store the database. Rule Based Machine Translation (RBMT) approach is used for the languages which are grammatical rich. Rule Based Machine Translation system is built using large number of the grammatical rules and the high quality of the bilingual dictionary with large number of entries. The RBMT system parses text and creates a morphological representation of the text and from the morphological information of the source language the target language is generated. This process requires extensive bilingual dictionaries with syntactic, and semantic information, and large sets of grammatical rules. The system uses the transfer grammatical rules for translating the source language into the target language. Rule-based Machine Translation systems give high quality translation when it uses large dictionaries and extensive grammar rules.

For Indian languages like Kannada and Telugu there is no predefined computational grammar, dictionaries and large amount of corpus. Hence in this paper Dictionary Based approach is used for translating from Kannada to Telugu.

2. Dictionary Based Machine Translation

From analysis of the current state of the research in machine translation it is proved that the bilingual dictionary with large number of entries is the basic requirement for the MT [5]. As of now the translation systems uses many guesses instead of the exact knowledge during translating from source language to the target language because of no adequate information for handling all the situations. The degree in which the above statement is true varies of course between different MT systems. For example it works well for the statistical machine translation because the translation model makes the guesses on the available corpus but the quality of the
translation is limited. The quality can be improved by increasing the bilingual corpus which covers all the aspects of the language model [4].

The idea behind the approach discussed here is that the bilingual dictionaries used for mapping the individual words of the source language to target language. Dictionary mapping alone cannot give the appropriate translation because the dictionaries alone can’t give the exact output in all the cases. To improve the quality of machine translation output transfer grammar rules are used along with the dictionary to produce the meaningful output. Transfer grammar rules are developed at the POS (parts of speech) tag level and a suffix mapping table is used for generating a target word.

Currently dictionaries, corpus and computational grammar are the source of problems with the translation engines, but (even with data-oriented approaches) when system has to give the accurate translation the above properties play a great role. These properties can be handled by defining the rule set based on the properties of the individual word [6]. This system can be improved based on the improvement of the quality of the bilingual dictionary and transfer grammar rule-set. Dictionary based approach involves Morph analyzer, bilingual dictionary, suffix mapping table and morph generator as shown in fig: 1. These modules are explained in the following section.

![Fig : 1 : Dictionary based machine translation](image)

**2.1 Morph Analyzer**

Morphological analyzer takes input as a word and produces output as the analysis of the word. Output holds the morphological information of the word. Kannada Saara system developed by Kavi Narayana Murthy at central university Hyderabad is used as morphological analyzer [1].

The output of morphological analyzer gives the morphological information of the given word. The morphological structure of Kannada verbs inflects for tense, person, gender, and number. The nouns inflect for plural, oblique, case and postpositions. The structure of verb is complex and capturing this complexity in a machine analyzable and generatable format is a challenging task [2]. Inflections of the Kannada verbs include finite, infinite, adjectival, adverbial and conditional markers.
Morph analyzer output:
raamanu || raama || N-PRP-PER-M.SL-NOM oLLeya || oLLeya ADJ-ABS huDuga || huDuga ||
N-COM-COU-M.SL-NOM

2.2 Bilingual Dictionary

Bilingual dictionary play a crucial part in machine translation. In machine translation system there are ideally three different types of dictionaries: source language dictionary, bilingual dictionary and a target language dictionary. In this paper bilingual dictionary is used. In a machine translation system building a dictionary is very costly in terms of time and data compared to building a grammar. Machine translation system gives good accuracy if the bilingual dictionary has more number of entries.

For most of the Indian languages translation systems don’t have the bilingual dictionaries in the electronic form. Hence the user should add the entries either manually or through automated way. For the automated way of building the dictionary large amount of the parallel corpus is required if the corpus is not available the only way is to add the entries in to the dictionary. If the machine translation system is developed for any domain then the user can only concentrate on adding the domain entries only. For domain independent system the user should add all the entries.

2.3 Morph Generator

The aim in morphological generation is to generate the inflected form of a word by using the morphological features obtained from the morphological analyzer. Defining or collecting any kind of linguistic resources is the time consuming process. Hence in order to save the developer time the linguistic resources that are already defined for analyzer can be reused. From experimental point of view, morphological generation is the reverse process of morphological analysis [8]. It is the process of converting the internal representation of a word to its surface form. Morph generator uses the suffix adding method. The morphological generation is the process of adding the corresponding suffixes to the root word to generate a target word.

Target word = root word + suffix
Example: raamuDu = raamu +Du

Input for the morphological generator would be the root word and this root word is inflected by the corresponding suffix to derive the target form of a word. The Morphological structure of Kannada verb is complex it inflects for the person, gender, and number markings and also it considers the auxiliaries which indicate aspect, mood etc. While generating the verb, the features should include all the inflections of the verb to select the appropriate suffix to generate the exact word [3].

Morph Generator output: Raamu + Du manchi bala +Du

2.4 Algorithm for Morphological Generator

This section describes the algorithm used to develop the morphological generator.
Algorithm

Step 1: Get the morphological information from the morph analyzer
Step 2: Check the bilingual dictionary for the entered word.
Step 3: If the entered word in the bilingual dictionary get the equivalent else transliterate the word.
Step 4: Extract the suffix and find the equivalent suffix in the suffix mapping table.
Step 6: Add suffix to the replaced root word to generate the target word
Step 7: Exit

3. Results

Machine Translation system uses 0 to 5 matrix system to evaluate the performance. Here 0&1 indicates not acceptable, 2 indicates acceptable with lot of corrections, 3 is acceptable with few corrections and 4 & 5 are acceptable. The system is tested on 100 sentences taken from Vijaya Karnataka children stories section. Out of which 10 sentences have got the score of 0, 12 sentences have got the score of 1, 24 sentences have got the score of 2, 44 sentences have got the score of 3 and 10 sentences have got the score of 4. The output pattern is shown below.

Input sentence: raamanu oLLeya huDuga

Morphological analysed output: raamanu || raama || N-PRP-PER-M.SL-NOM oLLeya || oLLeya ADJ-ABS huDuga || huDuga || N-COM-COU-M.SL-NOM

Morphological generated Output: Raamu(-Du) manchi bala (-Du)

Fig 2: Output of MT system
4. Conclusion

Machine Translation is one of the most focused areas of research in the field of Natural Language Processing. Machine Translation plays a very important role in breaking the language barrier and promoting the Interlingua communication in a multilingual country like India. In this paper, dictionary based approach is used for developing the MT system for Kannada and Telugu. The dictionary based approach is well suited for the languages which have the minimal linguistic resources and for the languages with the similar structure. For dictionary based approach bilingual dictionary is the crucial resource. Here bilingual dictionary with 8000 entries is developed. Morph generator uses the suffix adding method to generate the target word. Dictionary based approach can be further improved by developing the transfer grammar for both the languages.

Acknowledgement

The author wish to thank Dr. kavi Narayana Murthy, Professor at school of computer and information sciences, central university, Hyderabad

References

[1] K. Narayana Murthy “A Machine Assisted Translation System” ,Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, INDIA.
[2] W. John Hutchins and Halord L. Somers, “An Introduction To Machine Translation”, Academic Press Ltd., 1992, pp 1-9.
[3] Krishnamurti Bh J P L Gwunn “A Grammar of Modern Telugu “ chapter 5 : the structure of Telugu Orthography 1985.
[4] S. Kereto, C. Wongchaisuwat, Y. Poovarawan, “Machine translation research and development’. In proceedings of the Symposium on Natural Language processing in Thailand, pages 167-195, March 1993.
[5] K. Narayana Murthy “A Machine Assisted Translation System” ,Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, INDIA.
[6] Murthy. K,“MAT: A Machine Assisted Translation System”, In Proceedings of Symposium on Translation Support System( STRANS-2002), IIT Kanpur, 2002, pp. 134-139.
[7] Balajapally, P., Bandaru, P., Ganapathiraju, M., Balakrishnan, N., & Reddy, R. (2006). Multilingual Book Reader: Transliteration, Word-to-Word Translation and Full-text Translation. In VAVA 2006.
[8] Ambati, V., & Rohini, U. A Hybrid Approach to EBMT for Indian Languages. ICON 2007.

[9] Goyal, V., & Lehal, G. S. Evaluation of Hindi to Punjabi Machine Translation System. IJCSI International Journal of Computer Science, 2009, 4(1), pp.36-39.

[10] Dwivedi, S. K., & Sukhadeve, P. P. Machine Translation System in Indian Perspectives. Journal of Computer Science, 2010, 6(10), 1111-1116.

Sindhu D.V completed B.E form R L Jalapa institute of Technology under Visvesvaraya Technological University, Belgaum. She is currently pursuing dual degree (MTech+PhD) under VTU in the field of Natural Language Processing.

Dr. Sagar B.M completed B.E from Kuvempu University, Karnataka. He has done his Master degree from Visveswaraya Technological University, Karnataka and completed his from PhD in Natural Language Processing Avinashilingam University, Coimbatore. Currently he is working as Associate Professor in R.V College of Engineering. Published 45 research papers in National and international conferences. He holds life time membership in ISTE &CSI.