International Journal of Scientific & Engineering Research, Volume 5, Issue 4, April-2014
ISSN 2229-5518

A Typical Machine Translation System for English to Kannada

Mr. Chethan Chandra S Basavaraddi [1], Dr. H. L. Shashirekha [2]

Abstract— Machine Translation (MT) Research in India started in early 90's. Most of the operational and experimental systems are rule-based. Some important projects include: Mantra by CDAC, Pune, Anusaaraka by IIT-Kanpur and University of Hyderabad, AnglaBharti by IIT Kanpur, Shakti by LTRC, IIIT Hyderabad, MaTra by CDAC, Mumbai, ANUVAKSH by CDAC, Pune. IL-IL MT Consortia project: Language pairs include Malayalam-Tamil, Marathi-Hindi, Bengali- Hindi, Telugu-Hindi, Tamil-Telugu. E-IL MT Consortia project: Language pairs include English Hindi, Marathi, Oriya, Urdu and Bengali. Barriers in good quality MT output can be attributed to *ambiguity* in natural languages. Ambiguity can be classified into two: Structurally ambiguous and Lexically ambiguous. Structural ambiguity:"I saw the man on the hill with the telescope". Lexical ambiguity: Book that flight, I walked to the bank *(homonymy*), Bangalore is the capital city of Karnataka *(polysemy*), Cleaning fluids can be dangerous. The previous example illustrates the following facts Kannada follows SOV order as against SVO order of English. Kannada is a free word order language, Kannada is agglutinative, Difference occurs in Syntactic level i.e. word ordering Morphological level. The above facts justify the need for an efficient Syntax reordering module which takes care of syntactic differences. Morphological generator which takes care of complex morphology of the target language. Thus there is a need for the typical machine translation system for English to kannada and vice versa**.**

Keywords: Ambiguity, homonymy, Machine Translation (MT), Morphological generator, polysemy, subject object verb (SOV), subject verb object (SVO).

1. Introduction

Language processing:

Language processing refers to the way human beings use words to communicate ideas and feelings, and how such communications are processed and understood.

Natural language processing (**NLP**)**:** Natural language processing is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding -- that is, enabling computers to derive meaning from human or natural language input.

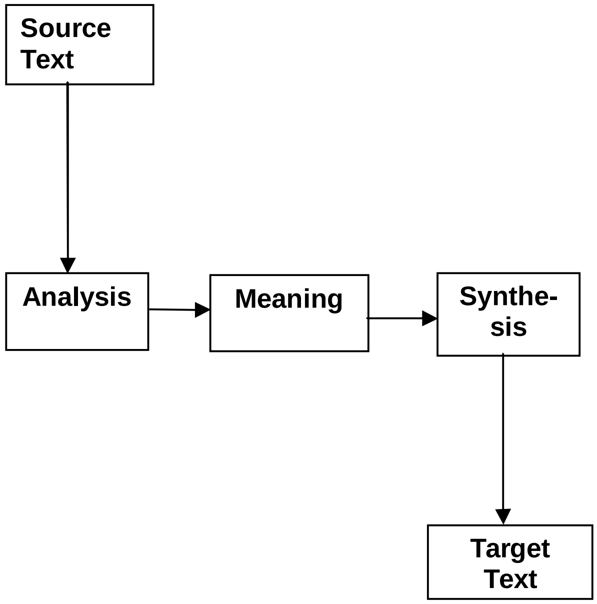
Machine translation is the study of designing systems that translate from one human language into another. This is a hard problem, since processing natural language requires work at several levels, and complexities and ambiguities arise at each of those levels. Some pragmatic approaches can be used to tackle these issues, leading to extremely useful systems. Machine translation (and natural language processing in general), is a difficult problem. There are two main reasons: Natural language is highly ambiguous. The ambiguity occurs at all levels – lexical, syntactic, semantic and pragmatic. A given word or sentence can have more than one meaning. '*party*' – 'a political entity' , or 'a social event, 'Deciding the suitable one in a particular case is crucial to getting the right analysis, and therefore the right translation. The second reason is that when humans use natural language, they use an enormous amount of common sense, and knowledge about the world, which helps to resolve the ambiguity. *Ex. "He went to the bank, but it was closed for lunch"* To get MT systems to exhibit the same kind of world knowledge in an unrestricted context requires a lot of effort.

The translation process may be stated as: Decoding the meaning of the source text; and Re-encoding this meaning in the target language.

Mr. Chethan Chandra S Basavaraddi,
Research Scholar,
Dept. of studies & research in computer science,
Mangalore University, Mangalagangotri-574199,
Mangalore, Karnataka, India, ph-9844508359,
E mail: raddi04@yahoo.com.

Dr. H. L. Shashirekha,
Chairperson,
Dept. of studies & research in computer science,
Mangalore University, Mangalagangotri-574199,
Mangalore, Karnataka, India, ph- 9480573679,
E mail: hlsrekha@gmail.com.

IJSER



2. The Machine translation

Machine translation is the study of designing systems that translate from one human language into another. This is a hard problem, since processing natural language requires work at several levels, and complexities and ambiguities arise at each of those levels. Some pragmatic approaches can be used to tackle these issues, leading to extremely useful systems.

2.1. General Issues in Machine Translation

- Machine translation (and natural language processing in general), is a difficult problem.
- There are two main reasons:
- Natural language is highly ambiguous. The ambiguity occurs at all levels – lexical, syntactic, semantic and pragmatic.
- A given word or sentence can have more than one meaning.

  '*party*' – 'a political entity' , or 'a social event,'

- Deciding the suitable one in a particular case is crucial to getting the right analysis, and therefore the right translation.
- The second reason is that when humans use natural language, they use an enormous amount of common sense, and knowledge about the world, which helps to resolve the ambiguity.

  Ex. "He went to the bank, but it was closed for lunch"
- To get MT systems to exhibit the same kind of world knowledge in an unrestricted context requires a lot of effort.

2.2. Different Categories of Machine Translation Systems

The three categories of machine translation systems are:

1. Machine Aided Human Translation
2. Human Aided Machine Translation
3. Fully Automated Machine Translation

IJSER

2.3.Various Approaches to Machine Translation

Linguistic or Rule Based Approaches

a. Direct Approach

b. Interlingua Approach

c. Transfer Approach

Non-Linguistic Approaches

d. Dictionary Based Approach

e. Corpus Based Approach

   i. Example Based Approach

   ii. Statistical Approach

Hybrid Approach

2.4. Linguistic or Rule Based Approaches

- The rule based approach is the principal methodology that was developed in machine translation.
- It requires a lot of linguistic knowledge during the translation.
- It uses grammar rules
- The computer programs will be helpful in analysing the text for determining grammatical information and features for each and every word in the source language,
- The source language is translated by replacing each word by the target equivalent or word that have the same context in the target language.
- Linguistic knowledge is required in order to write the rules for this type of approaches. These rules will play a vital role during the different levels of translation.
- The benefit of rule based machine translation method is that it can intensely examine the sentence at its syntax and semantic levels.
- There are complications in this method such as the prerequisite of vast linguistic knowledge and the vast number of rules needed to maintain the balance between source and target languages.

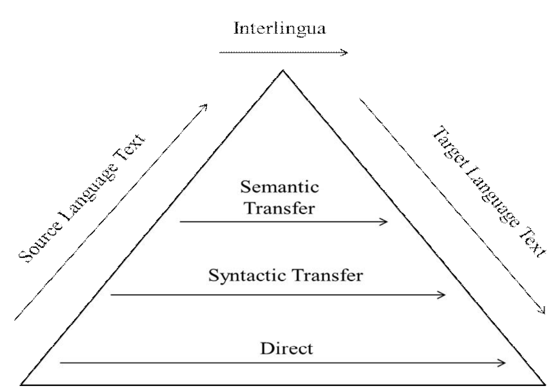
The three different approaches that require linguistic knowledge are as follows:

1. Direct Approach
2. Interlingua Approach
3. Transfer Approach

2.5. Components of a typical MT System

- We can divide the machine translation task into two or three main phases –
    - The system has to first analyse the source language input to create some internal representation.
    - It then typically manipulates this internal representation to transfer it to a form suitable for the target language.
    - Finally, it generates the output in the target language.
- Thus a typical MT system contains components for analysis, transfer and generation.

2.6. The Vaqueros Triangle of Machine Translation



2.7. Levels Of Analysis Of Source Language

Lexical level (or word level)

Syntactic level (or sentence level)

ISSN 2229-5518

Semantic level (meaning level)

Discourse level

Pragmatic level

2.8. Components of a typical MT System

(a).ANALYSIS COMPONENT – ANALYSER

1. Tokanizer : Tokenizes the sentence into words/ tokens
2. Morphological analyzer: It analyses words into morphemes and gives a meaning representation
3. POS tagger: Assign Parts-of-speech to the tokens / words
4. Syntactic parser: Analyses sentences into phrases or chunks.

(b). TRANFER COMPONENT – TRANSFER SYSTEM

- It transfers parsed source language syntactic structure into target language syntactic structure

(c). GENERTION – GENERATOR

1. Generates target language words form source language morphological information
2. Generates target language sentence from target language's morphological and syntactic information.

3. Challenges For English - Kannada MT

The major challenges that English-Kannada MT face are : i) the difference in the word order of English and Kannada, ii) Morphological and agglutinative nature of Kannada and iii) PNG (Person Noun Gender) and tense markers of Kannada.

The most significant challenge in MT is the difference in the word order or chunk order of English and Kannada languages. Word order plays an important role in positional languages like English and normally follow right-branching with Subject-Verb-Object orders. Unlike English language Kannada language has syntax of relatively free word order . These languages are verb final languages and all the noun phrases in the sentence normally appear to the left of the verb.

The subject noun phrase may also appear in many different positions relative to other noun phrases in the sentence. The word order does not determine the functional structure in Kannada language and permits scrambling. But normally Kannada language follows Subject Object Verb (SOV) order in contradiction with Subject Verb Object (SVO) order of English. The underlying structural differences between the source and target languages which forms a major weighing factor for the low translation quality and manifest themselves as a relatively poor translation. In this work the problem of structural differences between source and target languages are successfully overcome with reordering task using reordering rules.

The second challenge that really matters in the MT system is the morphological difference between English and Kannada. South Dravidian languages like Kannada is morphologically very rich than English. The morphological difference is because of the agglutinative nature of Kannada language, in which different word forms are formed by inserting different morphemes to the root word serially. For the highly agglutinative nature of Kannada, it is possible to form more than two thousand different word forms from a single root word. The following example illustrates the agglutinative nature of the Kannada. The different meaningful parts of the word „ಓದಕೆ ೊಂಡಿದ್ವನ" . (OdikoMDiddavana) -> „the one (masculine) who was reading" is:

ಓದನ + ಇ + ಕೆ ಳನು +ಂ ೊಂಡ್ + ಉ + ಇಯನ + ದ್್ + ಅ + ಅವನನ + ಅ

Odu + i + koLLu +MD+ u + iru + dd + a + avanu + a

Root + VBP+ AUXV +PST+ VBP + AUXV + PST+ RP + PRON-3SM + ACC

From this research work, it was proven that with use of morphological information, especially for morphologically rich languages like Kannada, the performance of the translation system can be substantially improved.

The next important challenge is based on the complexities of PNG and Tense Markers of Kannada language. The PNG and the tense marker concatenated to the verb stems are the two important aspects of verb morphology in Kannada.

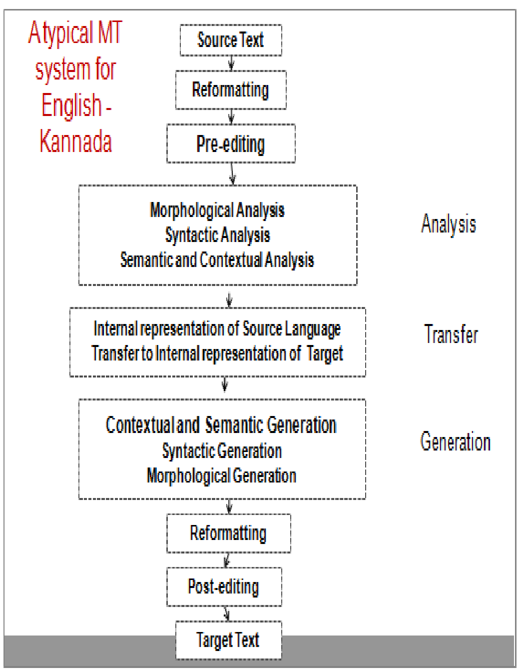
IJSER

The verbal inflectional morphemes attached to the verbs provide information about the syntactic aspects like number, person, case-ending relation and tense. The PNG features of the head noun of the subject NP determine the agreement marker of the verb. Usually the Kannada language verbs follow the regular pattern of suffixation.

Kannada morphology is more complex when compare with other South Dravidian language like Tamil and Malayalam. Generally PNG markers are not used in Malayalam and the same PNG markers are used in Tamil regardless of the type of tense in the sentence.

4. A typical MT System for English - Kannada

In order to achieve a reasonable translation quality in open source tasks, corpus based MT approaches require large amounts of parallel corpus which are not always available, especially for less resourced language pairs like English to Kannada. So the rule based method is the more realistic approach to tackle the MT problem between less resourced language pairs like English to Dravidian language like Kannada.



The proposed English to Kannada MT was developed based on the following three motivations:

1) Kannada language is a morphologically rich language and translation between a language with simple morphology like English and a language with complex morphology like Kannada is generally a complex task. Syntactic and semantic variance makes the problem much harder.

2) The second motivation is based on the fact that English-Kannada MT is not explored much.

3) Finally with the absence of any English-Kannada MT system so far, even a reasonable domain specific English-Kannada MT can find its immediate applications in government and education sectors.

5. Other Important issues

There other important issues to be addressed

1. Word Sense Disambiguation (WSD)
2. Name Entity Recognition
3. Multiword Expressions
4. Anaphora Resolution/coreferential resolution

5.1. Word sense disambiguation (WSD)

Word sense disambiguation (WSD) is the ability to identify the meaning of words in context in a computational manner. WSD is considered an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence. We introduce the reader to the motivations for solving the ambiguity of words and provide a description of the task. There are at least three kinds of WSD system: supervised, unsupervised, and knowledge-based approaches.

* "An empty can." – "Can he do that?" – "Can it!"
* "Run fast" – "Stand fast" (antonymy).
* "Time flies like an arrow."
* "Book that flight".
* I walked to the bank *(homonymy)*.
* Bangalore is the capital city of Karnataka *(polysemy)*.
* Cleaning fluids can be dangerous.

5.2. Multiword Expressions

Compounds of the type N+N, V+V, Adj+Adj, Adv+Adv, verb+auxiliary verb, N+Verbalizer, onomatopoeic compounds, etc. need to be taken together while parsing.

e.g. Old ladies hostel.

5.3. Anaphora resolution

Anaphora resolution is the replacing of words such as pronouns, which are semantically vacant, with the appropriate entity to which they refer. The city councilors refused the demonstrators a permit because they feared violence. The city councilors refused the demonstrators a permit because they advocated revolution.

5.4. Named entity recognition

Named entities include names of person, place, commodity, etc. They have to identified and separated out of other words which are found in common dictionaries. They need to be recognized as they undergo inflections like ordinary words.

ex. Madurai Kamaraj University
World Health Organization

6. Machine Translation in India

MT Research in India started in early 90's. Most of the operational and experimental systems are rule-based. Some important projects include:

-Mantra by CDAC, Pune
-Anusaaraka by IIT-Kanpur and University of Hyderabad
-AnglaBharti by IIT Kanpur
- Shakti by LTRC, IIIT Hyderabad
- MaTra by CDAC, Mumbai
-ANUVAKSH by CDAC, Pune

IL-IL MT Consortia project: Language pairs include Malayalam-Tamil, Marathi-Hindi, Bengali- Hindi, Telugu-Hindi, Tamil-Telugu.

E-IL MT Consortia project: Language pairs include English Hindi, Marathi, Oriya, Urdu and Bengali.

7. Acknowledgment:

I wish to thanks,

Dr. H. L SHASHIREKHA, Chairperson, Department of Studies & Research in Computer Science, Mangalore University., for her valuable motivation, guidance and suggestion, which helped me for completion of this Research paper.

8. Conclusion

The previous example illustrates the following facts, Kannada follows SOV order as against SVO order of English. Kannada is a free word order language. Kannada is agglutinative. Difference occurs in, Syntactic level i.e., word ordering, Morphological level. The above facts justifies the need for an efficient Syntax reordering module which takes care of syntactic differences. Morphological generator which takes care of complex morphology of the target language.

References:

[1] Antony P J, Ajith V P and Soman K P, "Statistical Method for English to Kannada Transliteration", International Conference on Recent Trends in Business Administration and Information Processing", Trivandrum, India, Published in Springer LNCS-CCIS, Volume 70, 356-362, 26-27 March 2010.

[3] Siva Reddy- Cross Language POS Taggers (and other Tools) for Indian Languages: An. Experiment with Kannada using Telugu Resource Nov8, 2011. www. aclweb.org / anthology/W11-3603.

[2] Parameswarappa. S., Bharathi, G. N., A novel approach to build Kannada web Corpus. Institute of Electrical and Electronics Engineers -Feb 28, 2012.

[4] S. Parameswarappa and V.N. Narayana, "Kannada Verbs and their Automatic Sense Disambiguation," To appear in the Proceedings of Int. Conf. on Global wordnet, GWC-2012, Kunibiki messe, Japan, 2012.

[5] Dr.Rajendran S, Amrita Universtiy,Coimbatore (Machine translation, workshop on natural language processing ,feb 25th- march 1st. Manglore University, Manglore)2014.

IJSER