Interlingual Machine Translation

Mallamma V Reddy[1], Dr. M. Hanumanthappa[2]
[1,2]Department of Computer Science and Applications,
Bangalore University, Bangalore, INDIA
[1]mallamma_vreddy@yahoo.co.in
[2]hanu6572@hotmail.com

***Abstract*—Interlingual is an artificial language used to represent the meaning of natural languages, as for purposes of machine translation. It is an intermediate form between two or more languages. Machine translation is the process of translating from source language text into the target language. This paper proposes a new model of machine translation system in which rule-based and example-based approaches are applied for English-to-Kannada/Telugu sentence translation. The proposed method has 4 steps: 1) analyze an English sentence into a string of grammatical nodes, based on Phrase Structure Grammar, 2) map the input pattern with a table of English-Kannada/Telugu sentence patterns, 3) look up the bilingual dictionary for the equivalent Kannada/Telugu words, reorder and then generate output sentences and 4) rank the possible combinations and eliminate the ambiguous output sentences by using a statistical method. The translated sentences will then be stored in a bilingual corpus to serve as a guide or template for imitating the translation, i.e., the example-based approach. The future work will focus on sentence translation by using semantic features to make a more precise translation.**

Keywords—Morphological analyser, Machine Translation [MT] , part-of-speech tagger

I. Introduction

Today, India has fifteen official languages. These languages originated from the Indo-Iranian branch of the Indo-European language family, the non-Indo-European Dravidian family, Austro-Asiatic, Tai-Kadai and the Sino- Tibetan language families. The languages that stem from the Dravidian family, are - Tamil, Kannada, Malayalam and Telugu, spoken in the South Indian states- Tamilnadu, Karnataka, Kerala and Andhra Pradesh. Most modern languages in North India, such as Hindi, Urdu, Punjabi, Gujarati, Bengali, Marathi, Kashmir, Sindhi, Konkani, Rajasthani, Assamese and Oriya, stem from Sanskrit and Pali.

Kannada or Canarese is a language spoken in India predominantly in the state of Karnataka, Making it the 25th most spoken language in the world. It has given birth to so many Indian languages like, Tulu, Kodava etc and one of the scheduled languages of India and the official and administrative language of the state of Karnataka [1]. Telugu is also one of the widely spoken languages in India especially in the state of Andhra Pradesh and the district of Yanam. Both Kannada and Telugu use the "UTF-8" / western windows encode and draw their vocabulary mainly from Sanskrit.

Various efforts have been made in developing machine translation (MT) systems for practical use. Historically, there are many approaches on MT research: transfer-based, interlingua-based, and etc. Among these approaches, the most distinctive are rule-based and corpus-based methods. Research on the corpus-based approach has emphasized on the importance of text corpora used as a source for linguistic and knowledge databases. There have been two major approaches among the corpus-based MT known as statistics-based and example-based. It might be said that all approaches have their own pros and cons. Therefore some MT [2] researchers have selected and combined them together for creating a new effective model. We also combine two potential approaches to produce our own strategy; namely, rule-based and example-based.

A. *Rule-Based and Example-Based Approaches*

The rule-based translation mostly consists of (1) a process of analyzing input sentences of a source language morphologically, syntactically and/or semantically and (2) a process of generating output sentences of a target language based on an internal structure or Interlingua. Each process is controlled by the dictionary and the rules. Meanwhile, the basic idea of example-based method [2] is to translate a sentence by using translation examples of similar sentences. The primary steps of example-based method are 1) collect examples in a database, 2) given an input, retrieve similar examples from the database, and 3) adapt the results of the similar examples to the current input and obtain the output.

B. *The hybrid translation method*

Many researchers apply both the rule-based and example-based methods as their own hybrid methods [3] propose a new hybrid translation method that combines a rule-based with an example-based method. An outline of the hybrid algorithm *is:* 1) find candidate sentences which are similar to the input sentence, 2) select the template: (a) rank the candidates by similarity to the input sentence (b) cluster the Translations of the candidate sentences (c) select the highest ranked pair of the best cluster, 3) translate input sentence by analogy to a selected template 4) output the adjusted sentence. For each difference, find it and translate using the rule-based modules.

C. *Interlingual Machine Translation*

Interlingual is an artificial language used to represent the meaning of natural languages, as for purposes of machine translation. It is an intermediate form between two or more languages. Interlingual Machine Translation is a methodology that employs interlingual for translation. Ideally the interlingual representation of the text should be sufficient to generate sentences in any language. Languages can have different parts of speech. In some cases two or more words in one language have a equivalent single word in another language. Interlingua approach in *“Fig. 1”*addresses these structural differences between languages. The disadvantage is that the design of interlingual is too complex. This is due to the fact that there is no clear methodology developed so far to build a perfect interlingual representation.

An interlingual lexicon is necessary to store information about the nature and behavior of each word in the language. The information includes events and actions.

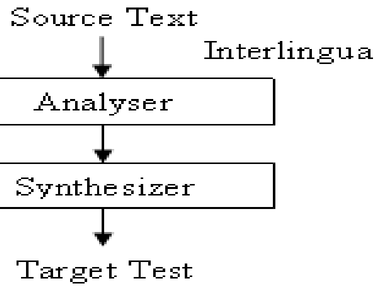


Figure 1. Interlingual Machine Translation.

A typical interlingual MT system has analyzer and synthesizer for each language. The analyzer produces interlingual representation of the meaning of the given text. The synthesizer produces one or more sentences with the meaning given by the analyzer.

II. Challenges in Machine translation

Machine translation[4] is the process of translating from source language text into the target language. Following is a list of challenges one has to face when attempt to do machine translation.

- Not all the words in one language have equivalent words in another language. In some cases a word in one language is to be expressed by group of words in another.
- Two given languages may have completely different structures. For example English has SVO structure while Kannada/Telugu has SOV structure.
- Sometimes there is a lack of one-to-one correspondence of parts of speech between two languages. For example, color terms of Kannada/Telugu are nouns whereas in English they are adjectives.
- The ways sentences are put together also differ among languages.
- Words can have more than one meaning and sometimes group of words or whole sentence may have more than one meaning in a language. This problem is called ambiguity.
- Not all the translation problems can be solved by applying values of grammar.
- It is too difficult for the software programs to predict meaning.
- Translation requires not only vocabulary and grammar but also knowledge gathered from past experience.
- The programmer should understand the rules under which complex human language operates and how the mechanism of this operation can be simulated by automatic means.
- The simulation of human language behavior by automatic means is almost impossible to achieve as the language is open and dynamic system in constant change. More importantly the system is not yet completely understood.

III. Machine Translation

The above mentioned challenges can be solved by using all the phases involved in machine translation depicted in the following *“Fig. 2”*.

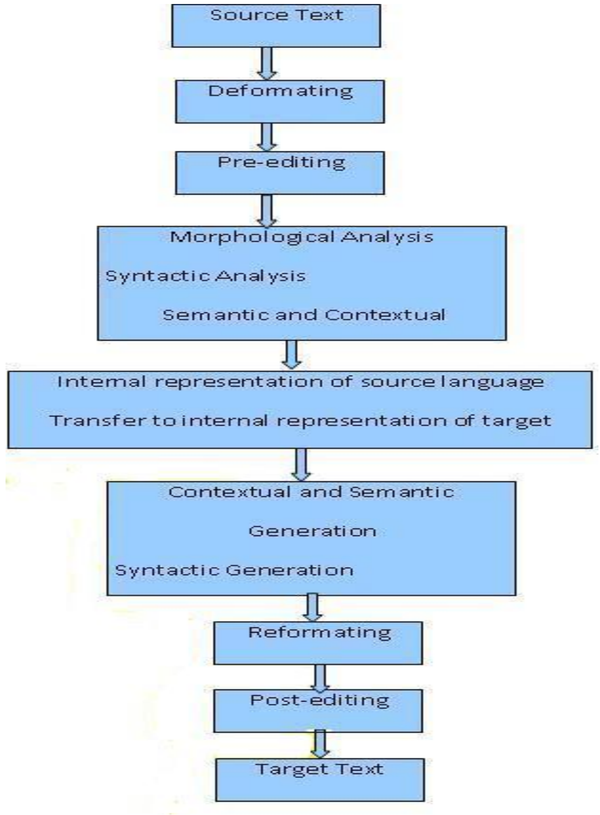


Figure 2. A Typical Machine Translation Process.

Text Input: This is the first phase in the **machine translation process** [4] and is the first module in any MT system. The sentence categories can be classified based on the degree of difficulty of translation. Sentences that have relations, expectations, assumptions, and conditions make the MT system understand very difficult. Speaker's intentions and mental status expressed in the sentences require discourse analysis for interpretation. This is due to the inter-relationship among adjacent sentences. World knowledge and commonsense knowledge could be required for interpreting some sentences.

Reformating and reformating: This is to make the machine translation process easier and qualitative. The source language text may contain figures, flowcharts, etc that do not require any translation. So only translation portions should be identified. Once the text is translated the target text is to be reformatted after post-editing. Reformatting is to see that the target text also contains the non-translation portion.

Pre-editing and Post editing: The level of pre-editing and post-editing depend on the efficiency of the particular MT system. For some systems segmenting the long sentences into short sentences may be required. Fixing up punctuation marks and blocking material that does not require translation are also done during pre-editing. Post editing is done to make sure that the quality of the translation is up to the mark. Post-editing is unavoidable especially for translation of crucial information such as one for health. Post-editing should continue till the MT systems reach the human-like.

Analysis, Transfers and Generation: Morphological analysis [5] determines the word form such as inflections, tense, number, part of speech, etc shown in following *"Table. I"* and*Table. II"*. Syntactic analysis determines whether the word is subject or object. Semantic and contextual analysis determines a proper interpretation of a sentence from the results produced by the syntactic analysis. Syntactic and semantic analyses are often executed simultaneously and produce syntactic tree structure and semantic network respectively. This results in internal structure of a sentence. The sentence generation phase is just reverse of the process of analysis.

TABLE I. FEW INFLECTIONS OF A VERB STEM AND ITS CORRESPONDING MEANINGS

Inflected Verb	**Meaning in English**	**Tense**	**Aspect**	**PNG**
ಮಾಡುವನು	He will do.	Future	Simple	3SM
ಮಾಡುತ್ತಿದ್ದಾನೆ	He is doing.	Present	Continuous	3SM
ಮಾಡಿರುವಳು	She has done.	Future	Perfect	3SF
ಮಾಡುತ್ತಿದ್ದಳು	She was doing.	Past	Continuous	3SF
ಮಾಡಿದಿರಿ	You did.	Past	Simple	2P-
ಮಾಡುತ್ತೇನೆ	I will do.	Future	Simple	1S-
ಮಾಡಿದ್ದರು	They did.	Past	Perfect	3P-
ಮಾಡಿರುತ್ತದೆ	It did.	Present	Perfect	3SN

TABLE II. DIFFERENT CASES AND THEIR CORRESPONDING CHARACTERISTIC SUFFIXES FOR NOUNS

Kannada Name	**English Name**	**Characteristic Suffix**
Prathama	Nominative	*0 (mu/ ru/ vu/ yu)*
Dwitiya	Accusative	*annu/ vannu/ rannu*
Tritiya	Instrumental	*iMda/ niMda/ riMda*
Chaturthi	Dative	*ge/ ige/ kke*
Pachami	Ablative	*deseyiMda*
Shashti	Genitive	*a/ ra/ da/ na*
Saptami	Locative	*alli/ nalli/ dalli/ valli*
Sambhodana	Vocative	*ee*

Morphological analysis and generation: Computational morphology deals with recognition, analysis and generation of words. Some of the morphological processes are inflection, derivation, affixes and combining forms as shown in *"Table. III"*Inflection is the most regular and productive morphological process across languages. Inflection alters the form of the word in number, gender, mood, tense, aspect, person, and case. Morphological analyser [5] gives information concerning morphological properties of the words it analyses.

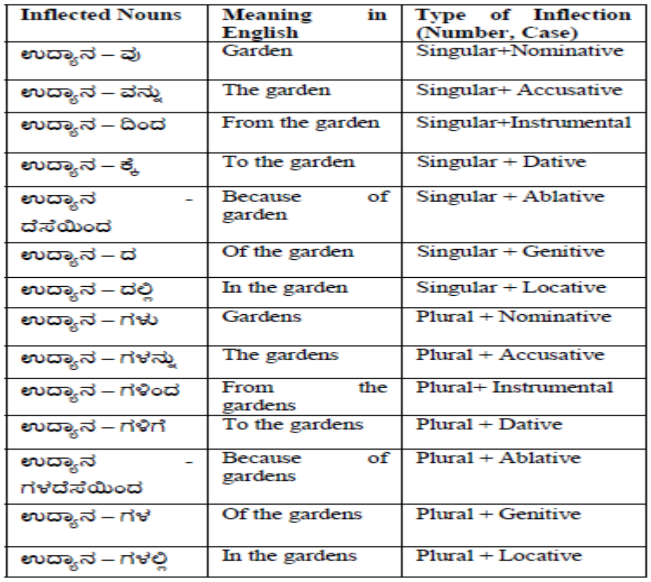
In Kannada, adjacent words are often joined and pronounced as one word. Such word combinations occur in two ways- *Sandhi* and *Samasa*. *Sandhi* (Morphophonemics) deals with changes that occur when two words or separate morphemes come together to form a new word. Few *sandhi* types are native to Kannada and few are borrowed from Sanskrit. We in our tool have handled only Kannada *sandhi*. However we do not handle *Samasa*.

TABLE III. SANDHI TYPES AND EXAMPLES FOR WORD COMBINATION

Complex word	**Simple/inflected words**	**Sandhi type**
ಚೆಂಡಾಟ	ಚೆಂಡು + ಆಟ	ಲೋಪ ಸಂಧಿ
ಸುಂದರವಾದ	ಸುಂದರ + ಆದ	ಆಗಮ ಸಂಧಿ
ಕೈದೋಟ	ಕೈ + ತೋಟ	ಆದೇಶ ಸಂಧಿ

Syntactic analysis and generation: As words are the foundation of speech and language processing, syntax can considered as the skeleton. Syntactic analysis concerns with how words are grouped into classes called parts-of-speech shown in *"Table. IV"*, how they group their neighbors into phrases, and the way in which words depends on other words in a sentence. Example

TABLE IV. INFLECTIONS OF A NOUN STEM AND ITS CORRESPONDING MEANINGS



Inflected Nouns	Meaning in English	Type of Inflection (Number, Case)
ಉದ್ಯಾನ – ವು	Garden	Singular+Nominative
ಉದ್ಯಾನ – ವನ್ನು	The garden	Singular+ Accusative
ಉದ್ಯಾನ – ದಿಂದ	From the garden	Singular+Instrumental
ಉದ್ಯಾನ – ಕ್ಕೆ	To the garden	Singular + Dative
ಉದ್ಯಾನ - ದೆಸೆಯಿಂದ	Because of garden	Singular + Ablative
ಉದ್ಯಾನ – ದ	Of the garden	Singular + Genitive
ಉದ್ಯಾನ – ದಲ್ಲಿ	In the garden	Singular + Locative
ಉದ್ಯಾನ – ಗಳು	Gardens	Plural + Nominative
ಉದ್ಯಾನ – ಗಳನ್ನು	The gardens	Plural + Accusative
ಉದ್ಯಾನ – ಗಳಿಂದ	From the gardens	Plural+ Instrumental
ಉದ್ಯಾನ – ಗಳಿಗೆ	To the gardens	Plural + Dative
ಉದ್ಯಾನ - ಗಳದೆಸೆಯಿಂದ	Because of gardens	Plural + Ablative
ಉದ್ಯಾನ – ಗಳ	Of the gardens	Plural + Genitive
ಉದ್ಯಾನ – ಗಳಲ್ಲಿ	In the gardens	Plural + Locative

Grammar formalism: Grammar formalism is a framework to explain the basic structure of a language. Reserachers propose the following grammar formalisms: Phrase Structure Grammar (PSG), Dependency Grammar, Case Grammar, Systematic Grammar, and Montague Grammar.

The variants of PSG are: Context Free PSG, Context Sensitive PSG, Augmented Transition Network Grammar (ATN), Definite Clause (DC) Grammar, Categorical Grammar, Lexical Functional Grammar (LFG), Generalised PSG, Head Driven PSG, and Tree Adjoining (TAG).

Not all the grammars suit a particular language. PSG, for example, does suit Japanese while dependency grammar does suite. Case grammar is popular as sentence in different languages that express the same contents may have the same case frames.

Parsing and Tagging: Tagging means the identification of linguistic properties of the individual words and parsing is the assessment of the functions of the words in relation to each other.

Semantic and Contextual analysis and Generation: A semantic analysis composes the meaning representations and assigns them the linguistic inputs. The semantic analyser uses lexicon and grammar to create context independent meanings. The source of knowledge consists of meaning of words, meanings associated with grammatical structures, knowledge about the discourse context and commonsense knowledge.

IV. Approach

The following approach is designed to produce an experimental system in translating English into Kannada/Telugu by using the 4 basic sentence patterns as a template. After that the output sentences will be stored as raw data for further applying an example-based method. The outline of the system is as follows:

1. Morphological analysis
2. Pattern mapping
3. Looking up bilingual dictionary
4. Disambiguating possible combinations

A. *Morphological Analysis*

An input sentence is first segmented into a word, written English sentences are automatically segmented, that is, each word is separated by a pause or space, then analyzed morphologically into a morpheme (in the form of a stem or root) by applying morphological analysis rules as shown in *"Fig. 3"*:

```
if check_RightPos (1) ="s" then
    if check_RightPos (2) ="es" then
        if check_RightPos (3) {"ies","ves" } then
            cut_RightPos (3) ;
            if check_RightPos (3) = "ies" then
                Add_char ("y") ;
            else
                Add_char ("f") ;
            end if
            if Search Dic( ) = TRUE then
                break ;
            end if
        else
            cut RightPos (2) ;
            if Search Dic( ) = TRUE then
            break ;
            end if
        end if
    else
        cut_RightPos (1) ;i
        f Search Dic( ) = TRUE then
        break ;
    end if
end if
end if
```

Figure 3. Sample of morphological rules for cutting off the suffixes of English plurality.

B. *Pattern mapping*

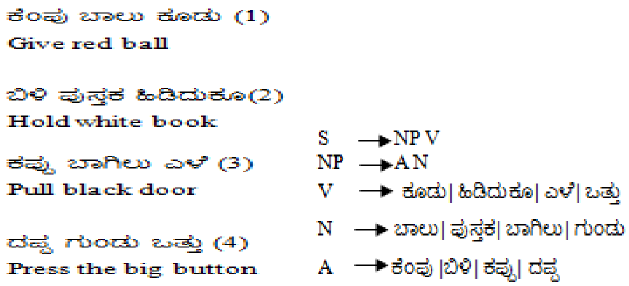
We make an attempt to map each pair of patterns from the simplest one to the least by using their similarity as the basis. In brief, a pair that can be mapped should be identical both in surface and deep structure. The two syntactic and semantic criterions, based on Phrase Structure Grammar [7] and Case Grammar, respectively, of pattern mapping that we have presumed is:

a) Each entry or word in a pair should have or represent the same syntactic relationship such as "subject", "verb" and "object", lying in linear order from left to right,

b) Each entry should underlie the same semantic relationship such as an "agent" of the action, an "object" or an "experiencer" etc.

Pattern mapping or transfer between the two languages involves a few steps. First, an English input sentence is syntactically analyzed into a series of non-terminal symbols (NP, VI, VT, ADJ, etc.). This string will be checked with the table of E-K sentence pattern mapping (*"Fig. 4).*" If the pattern of input sentence is identical to any pattern of English, it will be mapped to the Kannada/Telugu sentence pattern that is correspondent. Next, each English lexical entry will be reordered according to word ordering of Kannada/Telugu [6] sentence pattern. If the different sections are found, the rules can be of help before entering the next stage.

Following is the Kannada Grammatical Productions for a Robot to explain simple instructions like:



ಕೆಂಪು ಬಾಲು ಕೊಡು (1)
Give red ball

ಬಿಳಿ ಪುಸ್ತಕ ಹಿಡಿದುಕೊ(2)
Hold white book

ಕಪ್ಪು ಬಾಗಿಲು ಎಳೆ (3)
Pull black door

ದಪ್ಪ ಗುಂಡು ಒತ್ತು (4)
Press the big button

S → NP V
NP → A N
V → ಕೊಡು| ಹಿಡಿದುಕೊ| ಎಳೆ| ಒತ್ತು
N → ಬಾಲು| ಪುಸ್ತಕ| ಬಾಗಿಲು| ಗುಂಡು
A → ಕೆಂಪು |ಬಿಳಿ| ಕಪ್ಪು| ದಪ್ಪ

Figure 4. E-K sentence pattern mapping.

C. *Looking up bilingual dictionary and generating*

The bilingual dictionary of 10,000 entries is created in dbase format and looked up for mapping Kannada/Telugu equivalent entries onto the input string. Then a Kannada/Telugu output sentence is generated. Due to multiple meanings of one word, there is a large number of possible combinations produced inevitably by this process. Therefore we plan to use the statistical data to determine what the most likely one should be. At least it can help in reducing the number of candidates.

D. *possible combinations*

In this step the statistical method is used to calculate the probabilities of word that should be translated. In other words, we search through the statistical data stored and pick out the most likely word for our translation. With this method, we can eliminate a large number of possible combinations or candidate sentences. The output sentences that are ambiguous or have nonsensical meaning will be deleted as much as possible. As a result, we can obtain the most accurate and accepted outcome. For example, for a query "ಹಸುರು ತೋರಣ". the translation for ಹಸುರು is {green} and the translations for ತೋರಣ are {hang, designing}. Here, based on the context, we can see that the choice of translation [8] for the second word is water since it is more likely to co-occur with river.

Conclusion

In this paper we have explained the concepts and algorithms presented while implementing Bilingual Translation System for English to Kannada/Telugu which translates given input sentence in source language into target language using hybrid approach. New rules have been added to the proposed system in order to make the system more efficient. This work can be extended to other domains with the addition of new rules.

Acknowledgement

This is the major research project entitled ***Cross-Language Information Retrieval*** sanctioned to Dr. M. Hanumanthappa, PI-UGC-MH, Department of computer science and applications by the University grant commission. We thank to the UGC for financial assistance. This paper is in continuation of the project carried out at the Bangalore University, Bangalore, India.

References

[1] The Karnataka Official Language Act, "Official *website of Department of Parliamentary Affairs and Legislation*" Government of Karnataka. Retrieved 2007-06-29.

[2] Prof. Abdullah H. Homiedan," Machine Translation"

[3] Satoshi Shirai, Francis Bond and Yamato Takahashi. 1997. A hybrid rule and example-based method for machine translation. In *proceedings of the Natural anguageProcessing PacificRimSymposium1997,pages49-54,* December.

[4] S. Kereto, C. Wongchaisuwat, Y. Poovarawan. 1993. Machine translation research and development.In *proceedings of the Symposium on Natural Language processing in Thailand*, pages 167-195, March

[5] Dr. Ramakanth Kumar P, et.al." Kannada Morphological Analyser and Generator Using Trie" published in IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.1, January 2011

[6] Ganapathiraju Madhavi, Balakrishnan Mini,Balakrishnan N, Reddy Raj, "Om: One tool for many (Indian) languages",Journal of Zhejiang University SCIENCE,Vol 6A, No. 11, pp 1348-1353, Oct 2005.

[7] Wittaya Nathong. 1988. *Contrastive analysis of English and Thai.* Ramkhamhaeng University Press,Bangkok.

[8] Mallamma.V.Reddy,.Hanumanthappa.M, *"Kannada and Telugu Native Languages to English Cross Language Information Retrieval"* published in (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5) , Sep-Oct 2011, page-1876-1880. IISN: 0975-9646.