Indic Language Machine Translation Tool: English to Kannada/Telugu

Chapter · January 2013

DOI: 10.1007/978-81-322-1143-3_4

Mallamma Reddy
Rani Channamma University Belgavi

hine Translation Tool for NLP

Mallamma V Reddy[1], Dr. M. Hanumanthappa[2]
[1,2]Department of Computer Science and Applications,
Bangalore University, Bangalore, INDIA
{[1]mallamma_vreddy,[2]hanu6572}@bub.ernet.in

Abstract: Natural Language Processing is a field of computer science, AI and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output. In NLP, the major task is machine translation, the process of automatically translating text from one human language to another. This paper proposes a new model MT system in which Rule-Based, Dictionary-Based approaches are applied for English-to-Kannada/Telugu Language Identification and MT. The future work will focus on sentence translation by using Semantic-Structures/features to make a more precise translation.

.

Keywords- Artificial Intelligence [AI], Dictionary-Based, Natural language processing [NLP], Morphological analyzer, Machine Translation [MT], Part-of-speech tagger

1. INTRODUCTION

India has 18 officially recognized languages: 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 the age of Internet, the multiplicity of languages makes it even more necessary to have sophisticated machine translation systems. In this paper we are presenting the Machine translation system particularly from English to Kannada/Telugu and vice-versa, Kannada or Canarese is one of the 1652 mother tongues spoken in India. Forty three million people use it as their mother tongue. Telugu is a Central Dravidian language primarily spoken in the state of Andhra Pradesh, India, where it is an official language. According to the 2001 Census of India, Telugu is the language with the third largest number of native speakers in India (74 million), 13th in the Ethnologies list of most-spoken languages world-wide, and most spoken Dravidian language. As the English Language has ASCII encoding system for identifying the specification of a character, similarly Indian Languages have encoding systems named *Unicode* such as ñUTF-8ò , ñUTF-16ò,òUTF-32ò, ISCII. We are here using the character encoding system for Indian Languages is Unicode [1] Text Format ñUTF-8ò. This Machine Translation Model broadly classified into three modules

- **Language Identification Module:** Identifying the Language [2] of the Document(s) by uploading file(s).
- **Transliteration Module:** Transliteration is mapping of pronunciation and articulation of words written in one script into another script preserving the phonetics.
- **Translation Module:** Change in language while preserving meaning.

2. LANGUAGE IDENTIFICATION

The language identification problem refers to the task of deciding in which natural language a given text is written this is the one of the major challenge in the Natural Language Processing. Several corpora were collected to estimate the parameters of the proposed models and to evaluate the performance of the proposed approach. Using the *unigram statistical approach* [3] for each Language, the proposed model [4] is learnt with a training data set of 100 text lines from each of the three Languages- English, Kannada and Telugu the output is shown in Figure. 1.

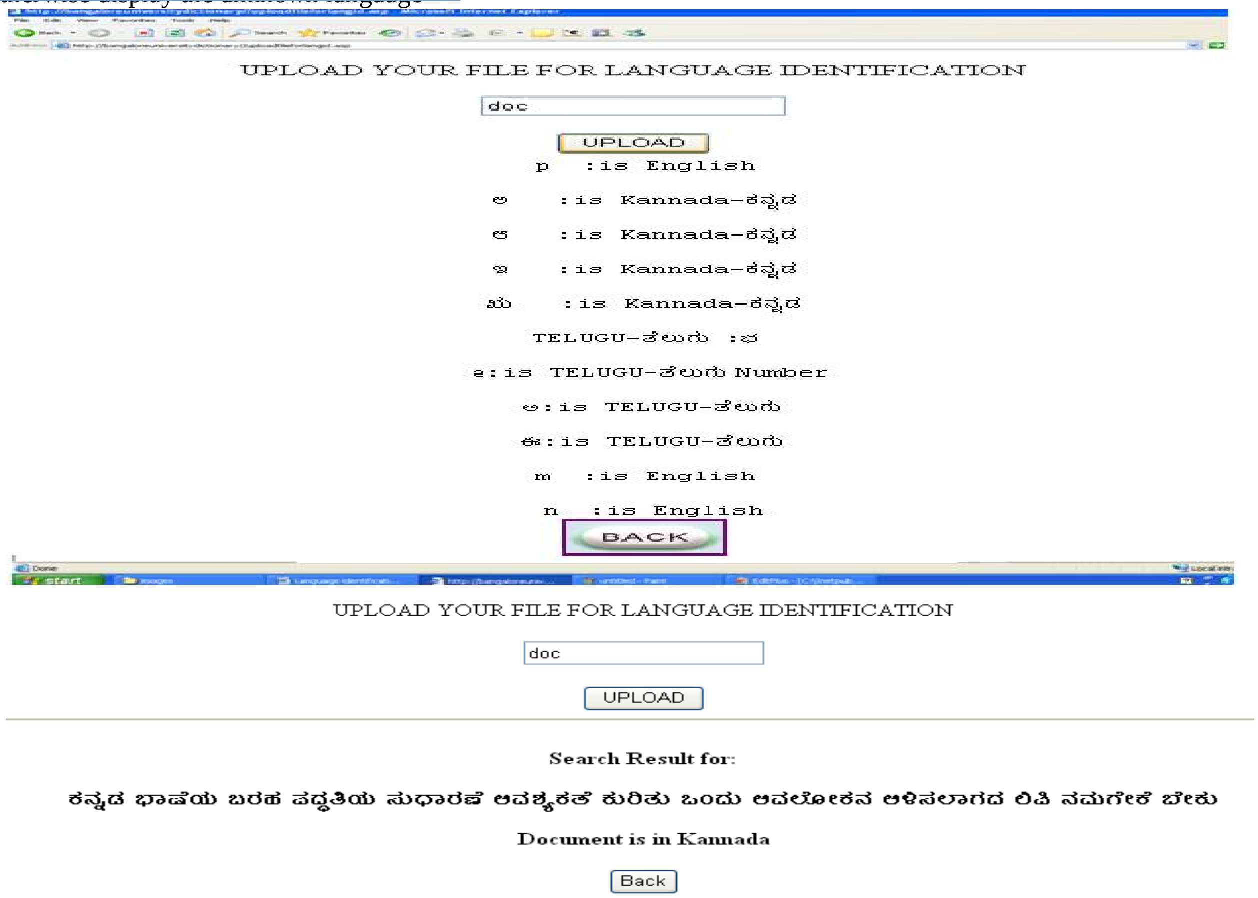
Language Identification algorithm used in the proposed model is.

Algorithm1: LandId ()

```
Input: Pre-processed text lines of English, Kannada and Telugu text Documents
Output: Identify the Language of the document.
1. Do for i = 1 to 3 Language document types
```

language

4. Otherwise display the unknown language


UPLOAD YOUR FILE FOR LANGUAGE IDENTIFICATION

doc

UPLOAD

p :is English

ಅ :is Kannada-ಕನ್ನಡ

ಆ :is Kannada-ಕನ್ನಡ

ಇ :is Kannada-ಕನ್ನಡ

ಋ :is Kannada-ಕನ್ನಡ

TELUGU-తెలుగు :ఐ

౨:is TELUGU-తెలుగు Number

అ:is TELUGU-తెలుగు

ఈ:is TELUGU-తెలుగు

m :is English

n :is English

BACK

UPLOAD YOUR FILE FOR LANGUAGE IDENTIFICATION

doc

UPLOAD

Search Result for:

ಕನ್ನಡ ಭಾಷೆಯ ಬರಹ ಪದ್ಧತಿಯ ಸುಧಾರಣೆ ಆವಶ್ಯಕತೆ ಕುರಿತು ಒಂದು ಅವಲೋಕನ ಅಳಿಸಲಾಗದ ಲಿಪಿ ನಮಗೇಕೆ ಬೇಕು

Document is in Kannada

Back

Fig 1: Language identification for English, Kannada and Telugu by uploading document(s).

3. TRANSLITERATION

The Language transliteration is one of the important area in natural language processing. Machine Transliteration is the conversion of a character or word from one language to another without losing its phonological characteristics. It is an orthographical and phonetic converting process. Therefore, both grapheme and phoneme information should be considered. Accurate transliteration of named entities plays an important role in the performance of machine translation and cross-language information retrieval process. Transliteration should not be confused with translation, which involves a change in language while preserving meaning. CLIR [5] is the acronym of a great variety of techniques, systems and technologies that associate information retrieval (normally from texts) in multilingual environments. Dictionaries have often been used for query translation in cross language information retrieval. However, we are faced with the problem of translating Names and Technical Terms from English to Kannada/Telugu. The most important query words in information retrieval are often proper names. The transliteration [6] [7] example as shown in Figure. 2.

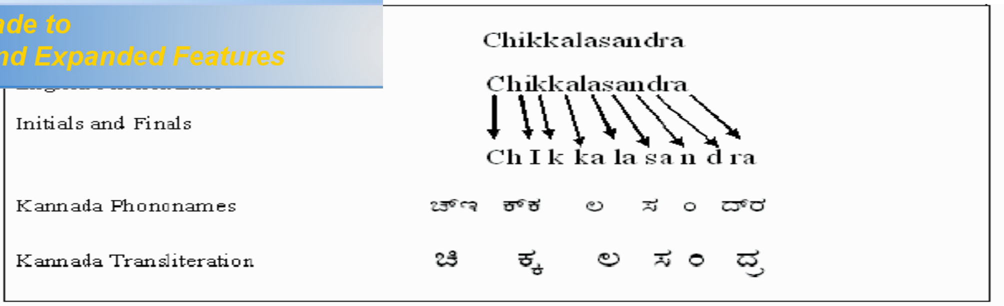


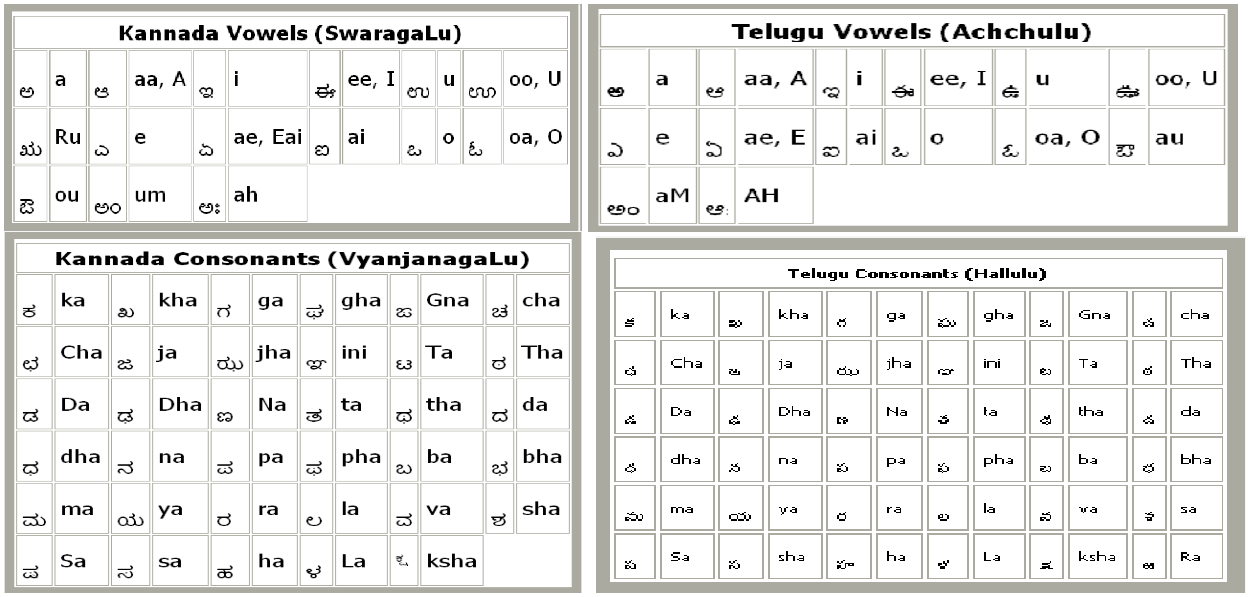
Fig 2: Example: English-Kannada Name Transliteration

3.1 Transliteration standards

- **Complete**: Every well-formed sequence of characters in the source script should transliterate to a sequence of characters from the target script, and vice versa.
- **Predictable**: The letters themselves (without any knowledge of the languages written in that script) should be sufficient for the transliteration, based on a relatively small number of rules. This allows the transliteration to be performed mechanically.
- **Pronounceable**: The resulting characters have reasonable pronunciations in the target script. Transliteration is not as useful if the process simply maps the characters without any regard to their pronunciation.
- **Reversible:** It is possible to recover the text in the source script from the transliteration in the target script. That is, someone that knows the transliteration rules would be able to recover the precise spelling of the original source text.

3.2 Transliteration for Characters

The number of consonants and vowels in Baraha Kannada/Telugu Both Kannada and Telugu use the ñUTF-8ò / western windows encode and draw their vocabulary mainly from Sanskrit, and their English equivalent transliteration is shown in Figure.3.



Kannada Vowels (SwaragaLu)

ಅ	a	ಆ	aa, A	ಇ	i	ಈ	ee, I	ಉ	u	ಊ	oo, U
ಋ	Ru	ಎ	e	ಏ	ae, Eai	ಐ	ai	ಒ	o	ಓ	oa, O
ಔ	ou	ಅಂ	um	ಅಃ	ah						

Telugu Vowels (Achchulu)

అ	a	ఆ	aa, A	ఇ	i	ఈ	ee, I	ఉ	u	ఊ	oo, U
ఎ	e	ఏ	ae, E	ఐ	ai	ఒ	o	ఓ	oa, O	ఔ	au
అం	aM	అః	AH								

Kannada Consonants (VyanjanagaLu)

ಕ	ka	ಖ	kha	ಗ	ga	ಫ	gha	ಙ	Gna	ಚ	cha
ಛ	Cha	ಜ	ja	ಝ	jha	ಞ	ini	ಟ	Ta	ಠ	Tha
ಡ	Da	ಢ	Dha	ಣ	Na	ತ	ta	ಥ	tha	ದ	da
ಧ	dha	ನ	na	ಪ	pa	ಫ	pha	ಬ	ba	ಭ	bha
ಮ	ma	ಯ	ya	ರ	ra	ಲ	la	ವ	va	ಶ	sha
ಷ	Sa	ಸ	sa	ಹ	ha	ಳ	La	ಕ್ಷ	ksha		

Telugu Consonants (Hallulu)

క	ka	ఖ	kha	గ	ga	ఘ	gha	ఙ	Gna	చ	cha
ఛ	Cha	జ	ja	ఝ	jha	ఞ	ini	ట	Ta	ఠ	Tha
డ	Da	ఢ	Dha	ణ	Na	త	ta	థ	tha	ద	da
ధ	dha	న	na	ప	pa	ఫ	pha	బ	ba	భ	bha
మ	ma	య	ya	ర	ra	ల	la	వ	va	శ	sa
ష	Sa	స	sha	హ	ha	ళ	La	క్ష	ksha	ఱ	Ra

Fig 3: English-Kannada/Telugu Character Mapping

3.2.1 Algorithm

We constructed a Dictionary with the help of training data that stores the possible mappings between English characters and Kannada/Telugu characters. Mapping was created between single English to single Kannada/Telugu annada/Telugu characters. Algorithm followed for making dictionary

Algorithm 2: Dictionary for Transliteration

```
for each (name_english, name_Kannada) in the training data:
index = 0
while index ! = len(name_english) and index != len(name_Kannada):
map name_english[index] to name_Kannada[index]
if index<len(name_english)-1
map (name_english [index] + name_english[index+1]) to name_Kannada [index]
index ++
index_english = len(name_english) - 1
index_Kannada = len(name_Kannada) - 1
while index_Kannada>-1 and index_english>-1:
map name_english[index_english] to name_Kannada[index_Kannada]
if index_english >0:
map (name_english[index_english-1]+name_english[index_english]) to name_Kannada[index_Kannada]
index_english
index_Kannada
```

For example:
If in training data English string is E1 E2 E3 E4 and corresponding Kannada string is K1 K2 K3.
Then we made the following mappings:

E1 --> K1
E2 --> K2
E3 --> K3
E4 --> K3
E3 --> K2
E2 --> K1

E1 E2 --> K1
E2 E3 --> K2
E3 E4 --> K3

If in training data English string is E1 E2 E3 and corresponding Kannada string is K1 K2 K3 K4.
Then we made the following mappings:

E1 --> K1
E2 --> K2
E3 --> K3
E3 --> K4
E2 --> K3
E1 --> K2

E1 E2 --> K1
E2 E3 --> K2
E2 E3 --> K4
E1 E2 --> K3

4. TRANSLATION

We use a Query Translation [8] based approach in our system since it is efficient to translate the query vis-a-vis documents. It also offers the flexibility of adding cross-lingual capability to an existing monolingual IR engine by just adding the query translation module. We use machine-readable bi-lingual KannadaŸ English and TeluguŸ English dictionaries created by BUBShabdasagar for query translation [9] [10] this is depicted in Figure.4. The KannadaŸ English bi-lingual dictionary [11] has around 14,000 English entries and 40,000 Kannada entries. The TeluguŸ English bi-lingual has relatively less coverage and has around 6110 entries.

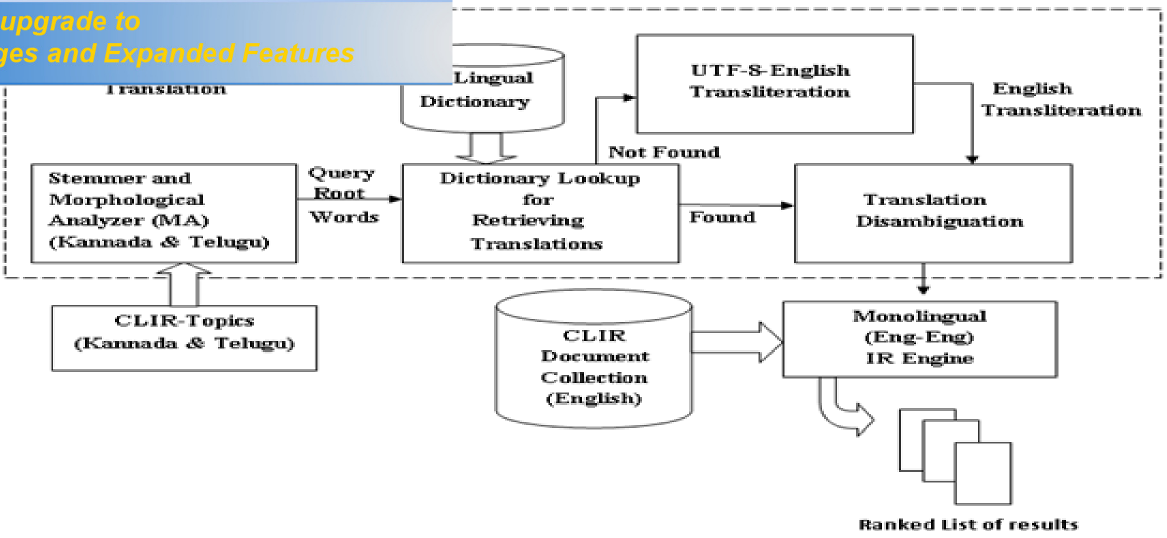

Translation
Lingual Dictionary
UTF-8-English Transliteration
English Transliteration
Not Found
Stemmer and Morphological Analyzer (MA) (Kannada & Telugu)
Query Root Words
Dictionary Lookup for Retrieving Translations
Found
Translation Disambiguation
CLIR-Topics (Kannada & Telugu)
CLIR Document Collection (English)
Monolingual (Eng-Eng) IR Engine
Ranked List of results

Fig 4: Query Based Translation Module

Kannada and Telugu, like other Indian languages, are morphologically rich. Therefore, we stem the query words before looking up their entries in the bi-lingual dictionary. In case of a match, all possible translations from the dictionary are returned. In case a match is not found, the word is assumed to be a proper noun and therefore transliterated by the UTF-8 English transliteration module. The above module, based on a simple lookup table and corpus, returns the best three English transliterations for a given query word. Finally, the translation disambiguation module disambiguates the multiple translations/transliterations returned for each word and returns the most probable English translation of the entire query to the monolingual IR engine.

4.1.1 Kannada Morphology

Kannada is a morphologically rich language in which morphemes combine with the root words in the form of suffixes. Kannada grammarians divide the words of the language into three categories namely:

i) **Declinable words** (namapada): Morphology of declinable words, as in many Dravidian languages is fairly simple compared to verbs. Kannada words are of three genders- masculine, feminine and neutral. Declinable and Conjugable words have two numbers- singular and plural.

ii) **Verbs** (kriyapada) or Conjugable words: The verb is much more complex than the nouns in Figure.5. There are three persons namely first, second and third person. Tense of verbs is past, present or future. Aspect may be simple, continuous or perfect. Verbs in Figure.6. Occur as the last constituent of the sentence. They can be broadly divided into finite or non-finite forms. Finite verbs have nothing added to them and are found in the last position of a sentence. They are marked for tense with Person-Number-Gender (PNG) markers. Non-finite verbs, on the other hand cannot stand alone. They are always marked for tense without PNG marker.

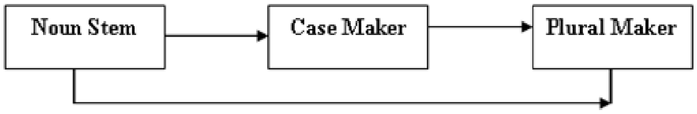


Fig 5: A formal Grammar for Kannada Nouns

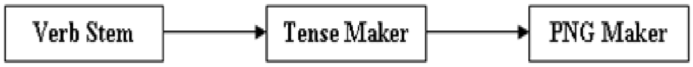


Fig 6: A formal Grammar for Kannada Verbs

ected words may be classified as adverbs, postpositions, conjunctions e words of this class are *haage, mele, tanaka, alli, bagge, anthu* etc.

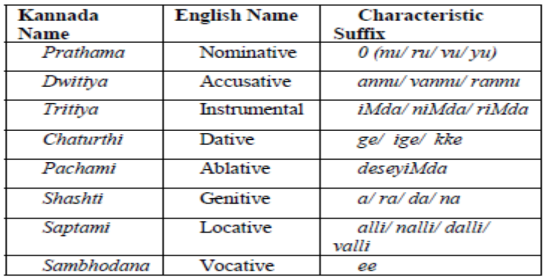
4.2 Morphophonemics

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.*

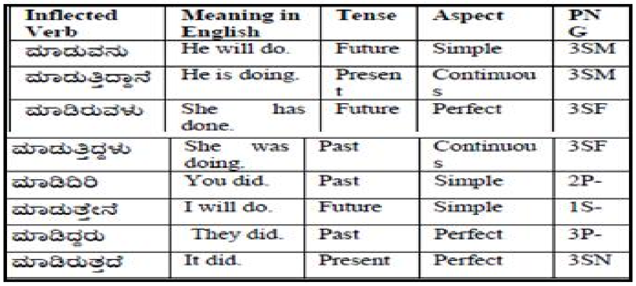
Kannada sandhi is of three types - *lopa, agama* and *adesha* sandhi. While *lopa* and *agama* take place both in compound words and in the junction of the crude forms of words and suffixes, *adesha* sandhi occurs only in compound words. Detailed description of sandhi types can be found in [12].

4.2.1 Algorithm For Morphological Analyzer and Generator

Morphological analysis [12] 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.



Kannada Name	English Name	Characteristic Suffix
Prathama	Nominative	*0 (nu/ 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*



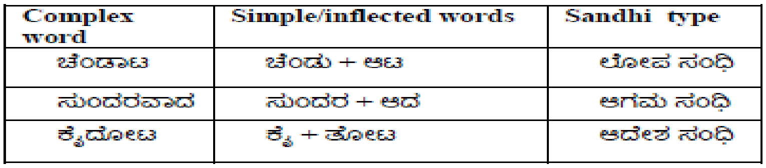
Inflected Verb	Meaning in English	Tense	Aspect	PNG
ಮಾಡುವನು	He will do.	Future	Simple	3SM
ಮಾಡುತ್ತಿದ್ದಾನೆ	He is doing.	Present	Continuous	3SM
ಮಾಡಿರುವಳು	She has done.	Future	Perfect	3SF
ಮಾಡುತ್ತಿದ್ದಳು	She was doing.	Past	Continuous	3SF
ಮಾಡಿದಿರಿ	You did.	Past	Simple	2P-
ಮಾಡುತ್ತೇನೆ	I will do.	Future	Simple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

TABLE 1:DIFFERENT CASES AND THEIR CORRESPONDING CHARACTERISTIC SUFFIXES FOR NOUNS

TABLE 2. FEW INFLECTIONS OF A VERB STEM AND ITS CORRESPONDING MEANINGS

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 analyzer [13] 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*.



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

TABLE III. SANDHI TYPES AND EXAMPLES FOR WORD COMBINATION

e new algorithm which is developed for morphological analyzer [13] rithm is simple and accurate.

Algorithm

Step 1: Get the word to be analyzed.
Step 2: Check whether the entered word is found in the Root Dictionary.
Step 3: If the word is found in the dictionary, stop;
Else
Step 4: Separate any suffix from the right hand side
Step5: If any suffix is present in the word, then check the availability of the suffix in the dictionary.
Then
Step 6: Remove the suffix present,
Then re-initialize the word without identified suffix, Go to Step 2.
Step 7: Repeat this process until the Dictionary finds the root/stem word.
Step 8: Store the English root/stem word in a variable and then get the corresponding Kannada word from the bilingual dictionary
Step 9: Check what all grammatical features does the English word have given and then generate the corresponding features for the Kannada word
Step 10: Exit.

4.2.2 Part Of Speech Tagger

Traditional grammar classifies the words into different categories. These are called parts of speech; the eight parts of speech are: the verb, the noun, the pronoun, the adjective, the adverb, the preposition, the conjunction, and the interjection. Each part of speech explains how the word is used. In fact, the same word can be a noun in one sentence and a verb or adjective in the next. Tag is a keyword or term assigned to a piece of information. The process of marking up a single ówordô, corresponding to a particular part of speech(like verb or noun) based on both its definition, as well as its context-i.e., relationship with adjacent and related words in a phrase, sentence, or paragraph, is known as Part of Speech Tagging (POST).

The Stanford part of speech tagger [14] is used for obtaining the part of speech of query term in context of the sentence.

PART-OF-SPEECH	**TAG**	**EXAMPLES**
ÅAdjective	JJ	happy, bad
ÅAdjective, comparative	JJR	happier, worse
ÅAdjective, cardinal number	CD	3, fifteen
ÅAdverb	RB	often, particularly
ÅConjunction, coordination	CC	and, or
ÅConjunction, subordinating	IN	although, when
ÅDeterminer	DT	this, each, other, the, a, some
ÅDeterminer, post determiner	JJ	many, same
ÅNoun	NN	aircraft, data
ÅNoun, plural	NNS	women, books
ÅNoun, proper, singular	NNP	London, Michael
ÅNoun, proper, plural	NNPS	Australians, Methodists
ÅPronoun, personal	PRP	you, we, she, it
ÅPronoun, question	WP	who, whoever
ÅVerb, base present form	VBP	take, live

Algorithm 3: POS tagging

Input: Untagged English Sentence
Output: Tagged Translated Kannada Sentence

```
        If the Word is tagged
                Stop
        Else
                Tag<=Word
        End if
End For
Return Tagged Translated Kannada Sentence
```

4.2.3 Dictionary based approach

One of the major factors that can potentially degrade the effectiveness of dictionary-based cross-language information retrieval is the ambiguity in translating query words [15]. In the efforts to resolve this translation ambiguity [16], several recent studies [17] have suggested the strategy of translation selection by exploiting word co- occurrence patterns. Usually a similarity measurement between two translation words is defined in the form of word co-occurrence statistics. With the word similarities, we can then measure the coherence of a translation word with regard to a query. Only translation words with high coherence scores will be selected for the translation of the query.

The Dictionary based method consists of dictionaries, multilingual thesauri. The process of Dictionary based method shown in Figure.7. Query translation [18] is relatively efficient and can be performed as needed. The principal limitation of query translation is that queries are often short and short queries provide little context for disambiguation.

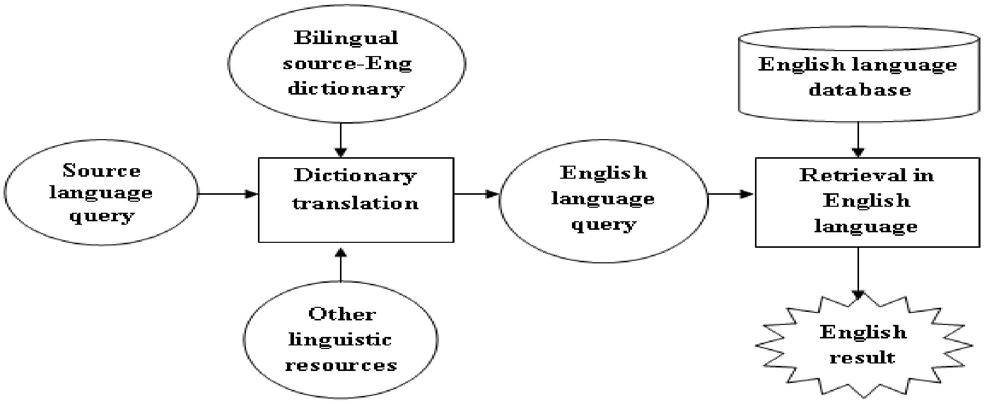

Bilingual source-Eng dictionary
English language database
Source language query
Dictionary translation
English language query
Retrieval in English language
Other linguistic resources
English result

Fig 7: Dictionary Based method for Query Translation

4.2.4 Rule-Based Approach

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. Each process is controlled by the dictionary and the rules.

5. THE SELECTION OF WORD TRANSLATION

Normally in CLIR words that are not included in phrases are translated word-by-word shown in Figure 8. However, this does not mean that they should be translated in isolation from each other. Instead, while translating a word, the other words (or their translations) form a "context" that helps determine the correct translation for the given word.

Working in this principle of translation our assumption is that the correct translations of query words tend to co-occur in target language documents and incorrect translations do not. Therefore, given a set of original source hem the best translation word such that it co-occurs most often with documents. For example

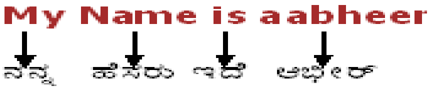


Fig 8: word-by-word translation

Finding such an optimal set is computationally very costly. Therefore, an approximate greedy algorithm is used. It works as follows: Given a set of *m* original query terms {$a_1,\ldots, an$}, we first determine a set T_i *of* translation words for each a_i through the dictionary. Then we try to select the word in each T_i that has the highest degree of *cohesion* with the other sets of translation words. The set of best words from each translation set forms our query translation.

Cohesion is one of the aspects that are taken into consideration in the textual analysis of translations. Cohesion is the study of textual equivalence defining it as the network of lexical, grammatical, and other relations which provide links between various parts of a text.

The cohesion is based on term similarity. The EMMI weighting measure has been successfully used to estimate the term similarity in [15]. We take a similar approach. However, we also observe that EMMI does not take into account the distance between words. In reality, we observe that local context is more important for translation selection. If two words appear in the same document but at two distant places, it is unlikely that they are strongly dependent. Therefore, we add a distance factor in our calculation of word similarity. Formally, the similarity between terms *x* and *y* is

$$SIM(x,y) = p(x,y)\times\log_2\left(\frac{p(x,y)}{p(x)\times p(y)}\right) - K\times\log_2 Dis(x,y) \qquad \textbf{(1)}$$

Where

$$p(x,y) = \frac{c(x,y)}{c(x)} + \frac{c(x,y)}{c(y)} \qquad \textbf{(2)}$$

$$p(x) = \frac{c(x)}{\sum_x c(x)} \qquad \textbf{(3)}$$

c(x,y) is the frequency that term *x* and term *y* co-occur in the same sentences in the collection, *c(x)* is the number of occurrence of term *x* in the collection, *Dis*(*x*,*y*) is the average distance (word count) between terms *x* and *y* in a sentence, and *K* is a constant coefficient, which is chosen empirically. (*K*=0.8 in our experiments). The cohesion of a term *x* with a set *X* of other terms is the maximal similarity of this term with every term in the set, i.e.

$$Cohesion(x,X) = Max_{y\in X}\, SIM(x,y) \qquad \textbf{(4)}$$

Algorithm 4: Greedy Algorithm for Word Translation using Cohesion Technique

```
Step 1: For each source query word $ai$ (i = $1$ to $n$), retrieve a set of translations $Ti$ from the lexicon;
Step 2: For each set $Ti$ ( $i = 1$ to $n$), do
Step 3: For each term $tij$ in $T$i, do
Step 4:  For each set $Tk$ ($k=1$ to $n$ & $k\neq i$),
```

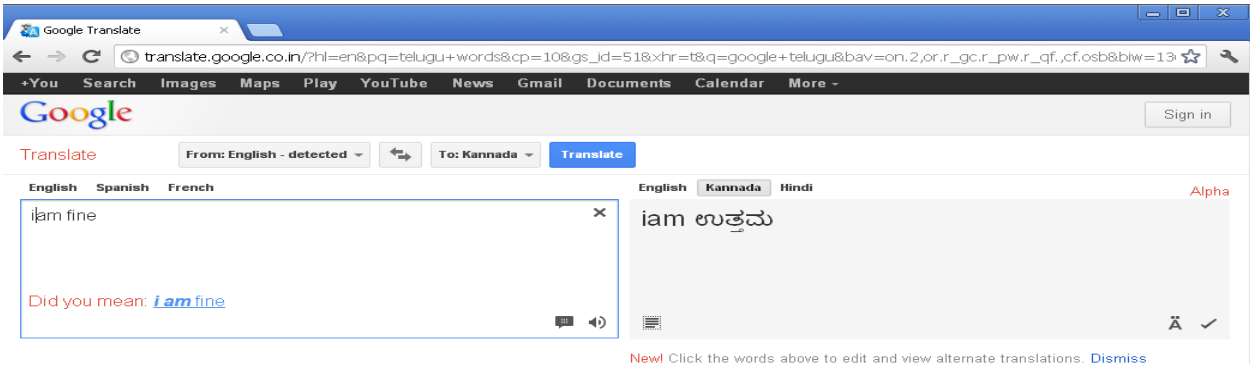
);

Cohesion (tij, Tk) (k=1 to n & k≠ i);

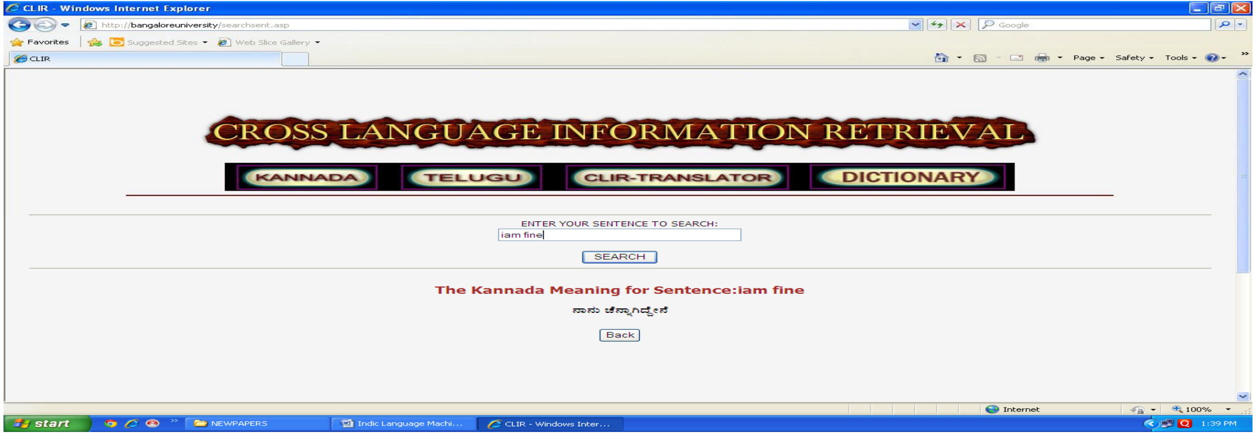
score, and add the selected sense into the set *T*.

6. EXPERIMENTAL SETUP AND COMPARATIVE STUDY

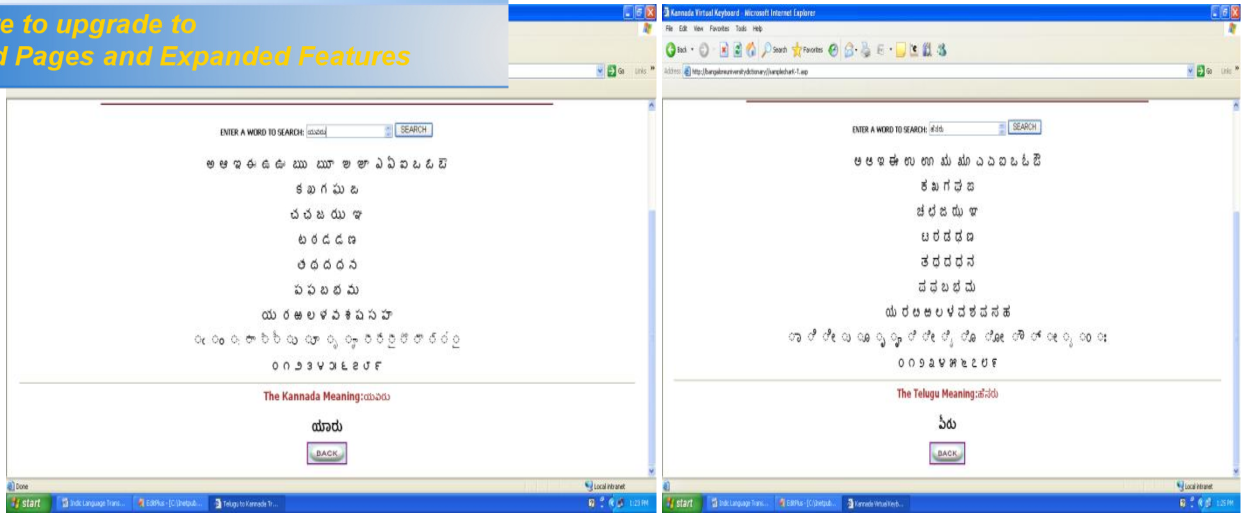
Several corpora were collected to estimate the parameters of the proposed models and to evaluate the performance of the proposed approach. Cross Language Information Retrieval Tool [19] is built by using the ASP.NET as front end and Database as back end, the Kannada/Telugu is encrypted by using the ñUTF-8ò /Encoding system. Telugu [20] and Kannada [21] and vice-versa are the source language and the target language, respectively, in our query translation. All the experiments carried out here involve the same set of Kannada/Telugu queries and the same query expansion, translation and retrieval method. The only difference between the experimental conditions is in what dictionaries are used in the query translation. We have trained the systems with corpus size of 200, 500 and 1000 lexicons and sentences respectively. Performances of the systems were evaluated with the same set of 500 distinguished sentences/Phases that were out of corpus. The experiment results as shown in Figure 9 and Figure 10. The comparative results are shown in result 1 and result 2


Google Translate
translate.google.co.in/?hl=en&pq=telugu+words&cp=10&gs_id=51&xhr=t&q=google+telugu&bav=on.2,or.r_gc.r_pw.r_qf.,cf.osb&biw=13
From: English - detected
To: Kannada
iam fine
Did you mean: *i am* fine
iam ಉತ್ತಮ
New! Click the words above to edit and view alternate translations. Dismiss

Result 1: Google Translation


CLIR - Windows Internet Explorer
http://bangaloreuniversity/searchsent.asp
CROSS LANGUAGE INFORMATION RETRIEVAL
ENTER YOUR SENTENCE TO SEARCH:
iam fine
The Kannada Meaning for Sentence:iam fine
ನಾನು ಚೆನ್ನಾಗಿದ್ದೇನೆ

Result 2: CLIR Translation



ENTER A WORD TO SEARCH: యవరు SEARCH

అ ఆ ఇ ఈ ఉ ఊ ఋ ౠ ఌ ౡ ఎ ఏ ఐ ఒ ఓ ఔ

క ఖ గ ఘ ఙ

చ ఛ జ ఝ ఞ

ట ఠ డ ఢ ణ

త థ ద ధ న

ప ఫ బ భ మ

య ర ఱ ల ళ వ శ ష స హ

ః ం ః ా ి ీ ు ూ ృ ౄ ె ే ై ొ ో ౌ ్

౦ ౧ ౨ ౩ ౪ ౫ ౬ ౭ ౮ ౯

The Kannada Meaning:యవరు

ಯಾರು

BACK

Kannada Virtual Keyboard - Microsoft Internet Explorer

Address http://bangaloreuniversitydictionary//samplechart-1.asp

ENTER A WORD TO SEARCH: ಹೆಸರು SEARCH

ಅ ಆ ಇ ಈ ಉ ಊ ಋ ೠ ಎ ಏ ಐ ಒ ಓ ಔ

ಕ ಖ ಗ ಘ ಙ

ಚ ಛ ಜ ಝ ಞ

ಟ ಠ ಡ ಢ ಣ

ತ ಥ ದ ಧ ನ

ಪ ಫ ಬ ಭ ಮ

ಯ ರ ಱ ಲ ಳ ವ ಶ ಷ ಸ ಹ

ಾ ಿ ೀ ು ೂ ೃ ೄ ೆ ೇ ೈ ೊ ೋ ೌ ್ ಂ ಃ

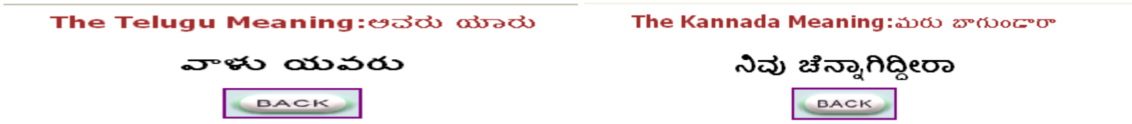
೦ ೧ ೨ ೩ ೪ ೫ ೬ ೭ ೮ ೯

The Telugu Meaning:ಹೆಸರು

పేరు

BACK

Fig. 9: Sample Results for Word



The Telugu Meaning:ಅವರು ಯಾರು

వాళు యవరు

BACK

The Kannada Meaning:మరు బాగుండారా

ನಿವು ಚೆನ್ನಾಗಿದ್ದೀರಾ

BACK

Fig. 10: Sample Results for Sentences

7. EVALUATION METRIC AND PERFORMANCE

In the experiment, the performance of word translation extraction was evaluated based on precision and recall rates at the word. Since, we considered exactly one word in the source language and one translation in the target language at a time. The word level recall and precision rates were defined as follows:

[illegible] (5)

[illegible] (6)

From the experiment we found that the performances of our systems are significantly well and achieves very competitive accuracy by increasing the corpus size as shown in Figure 11.

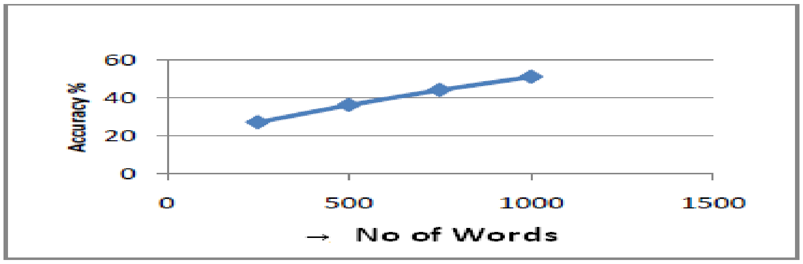


Fig 11: Performance Graph

WORK

guŸ English CLIR system developed for the Ad-Hoc bilingual Task. Our approach is based on query Translation using bi-lingual dictionaries. Transliteration of words which are not found in the dictionary is done using a simple rule based approach. Dictionary-based query translation has been widely used in CLIR because of its simplicity and the increasing availability of machine-readable bilingual lexicons. However, besides the problem of completeness of the lexicon, we are also faced with the problem of selecting the best translation word(s) from the dictionary.

In this paper, we also presented a method to identify and separate text lines of English, Kannada and Telugu documents from a trilingual document is presented. The approach is based on the analysis of the *Unigram statistical approach* of individual text lines and hence it requires character or word segmentation. In future we can also use this language identification module for translation with the help of bilingual dictionary. This will be very useful for machine translation from English to Kannada/Telugu language. One of the major challenge and future work is that English has Subject Verb Object (SVO) structure while Kannada has Subject Object Verb (SOV) structure in Machine translation will be unraveled by using morphology of Natural Languages.

Acknowledgement

I owe my sincere feelings of gratitude to Dr. M. Hanumanthappa, for his valuable guidance and suggestions which helped me a lot to write this paper. 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. http://www.ssec.wisc.edu/~tomw/java/unicode.html#x0C80
2. Penelope Sibun, And Jeffry C Reynar, ñ*Language Identification: Examining the issues"*
3. Tommi Vatanen, Jaakko J. V¨Ayrynen, Sami Virpioja, ñ*Language Identification of Short Text Segments with N-gram Models*ò
4. Bashir Ahmed, Sung-Hyuk Cha, And Charles *Tappert "Language Identification from Text Using N-gram Based Cumulative Frequency Addition"*published in Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 7th, 2004
5. Prasad Pingali, Vasudev Varma, *Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006*. In working notes for the CLEF 2006 workshop (Cross Language Adhoc Task), 20-22 September, Alicante, Spain.
6. Jong-Hoon Oh Key-Sun Choiò*Machine Learning Based English-to-Korean Transliteration using Grapheme and Phoneme Information*ò IEICE TRANS.INF. & SYST., VOL.E88-D, NO.7, july2005, pp 1737-1748.
7. Jasleen kaur, Gurpreet Singh josan, ñ*Statistical Approach to Transliteration from English to Punjabi"* published in International Journal on Computer Science and Engineering (IJCSE) ISSN: 0975-3397 Vol. 3 No. 4 Apr 2011 1518.
8. Ballesteros, L. and Croft, W.B., "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval", in Proceedings of ACM SIGIR Conference, 20: 84-91. 1997.
9. Prof. Abdullah H. Homiedan *"Machine Translation".*
10. 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

*the Dictionaries in a Machine Translation System*ò. In Lawson,
ine Translationò. North-Holland. 1982.

13. Jisha P.Jayan, Rajeev R R, S Rajendran, ñ*Morphological Analyser and Morphological Generator for Malayalam - Tamil Machine Translation"* published in International Journal of Computer Applications (0975 ï 8887) Volume 13ï No.8, January 2011.
14. Part of speech tagger is available at http://nlp.stanford.edu/software/tagger.shtml
15. M. Adriani, ñUsing statistical term similarity for sense disambiguation in cross-language information retrievalò. *Inf. Retr.*, 2(1):7182, 2000.
16. A. Maeda, F. Sadat, M. Yoshikawa, and S. Uemura,ò Query term disambiguation for web cross-language information retrieval using a search engine. In *IRAL '00: Proceedings of the fifth international workshop on Information retrieval with Asian languages*, pages 2532. ACM Press, 2000.
17. S. H. M. Myung-Gil Jang and S. Y. Park. Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th annual meeting of the association for computational linguistics*, 1999.
18. J. Gao, J.-Y. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang. Improving query translation for cross-language information retrieval using statistical models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96104. ACM Press, 2001.
19. Mallamma.V.Reddy, M.Hanumanthappa, *"CLIR Project (English to Kannada and Telugu)"* http://bangaloreuniversitydictionary//menu.asp
20. The Karnataka Official Language Act". *Official website of Department of Parliamentary Affairs and Legislation*. Government of Karnataka. Retrieved 2012-07-29.
21. http://en.wikipedia.org/wiki/Telugu_language