Indic Language Machine Translation Tool: English to Kannada/Telugu

Mallamma V. Reddy and M. Hanumanthappa

Abstract Natural Language Processing is a field of computer science, AI and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, computer extracts meaningful information from natural language input and/or producing natural language output. The major task in NLP is machine translation, which automatically translates text from one human language to another by preserving its meaning. This paper proposes new model for Machine-Translation system in which Rule-Based, Dictionary-Based approaches are applied for English-to-Kannada/Telugu Language-Identification and Machine Translation. The proposed method has four steps: first, Analyze and tokenize an English sentence into a string of grammatical nodes second, Map the input pattern with a table of English–Kannada/Telugu sentence patterns, third, Look-up the bilingual-dictionary for the equivalent Kannada/Telugu words, reorder and then generate output sentences and fourth step is to Display the output sentences. The future work will focus on sentence translation by using semantic features to make a more precise translation.

Keywords Natural language processing (NLP) · Language identification · Transliteration · Morphological analyzer · Machine translation (MT)

M. V. Reddy (✉) · M. Hanumanthappa
Department of Computer Science and Applications, Bangalore University, Bangalore, Karnataka, India
e-mail: mallamma_vreddy@yahoo.co.in

M. Hanumanthappa
e-mail: hanu6572@hotmail.com

P. P. Swamy and D. S. Guru (eds.), *Multimedia Processing, Communication and Computing Applications*, Lecture Notes in Electrical Engineering 213,
DOI: 10.1007/978-81-322-1143-3_4,

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 [1, 2] systems. In this paper we are presenting the Machine translation system particularly from English to Kannada/Telugu and vice versa, Kannada [3] or Canarese is one of the 1,652 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 [4] 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 [5] such as "UTF-8", "UTF-16", "UTF-32", ISCII. Machine Translation Model broadly classified into three modules.

- **Language Identification Module:** Identifying the Language [6] of the Document(s) by uploading file(s) or by entering the text
- **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 is the major challenge in Natural Language Processing. Several corpora were collected to estimate the parameters of the proposed models to evaluate the performance of the proposed approach. Using the unigram statistical approach for each Language, the proposed model [7, 8] is learnt with a training data set of 100 text lines from each of the three Languages- English, Kannada and Telugu. Language Identification [9] algorithm is described and result is shown in Fig. 1.

```
Algorithm LandId ()
Input: Pre-processed text lines of English, Kannada and
Telugu text Doc(s)
Output: Identify the Language of the document.
```

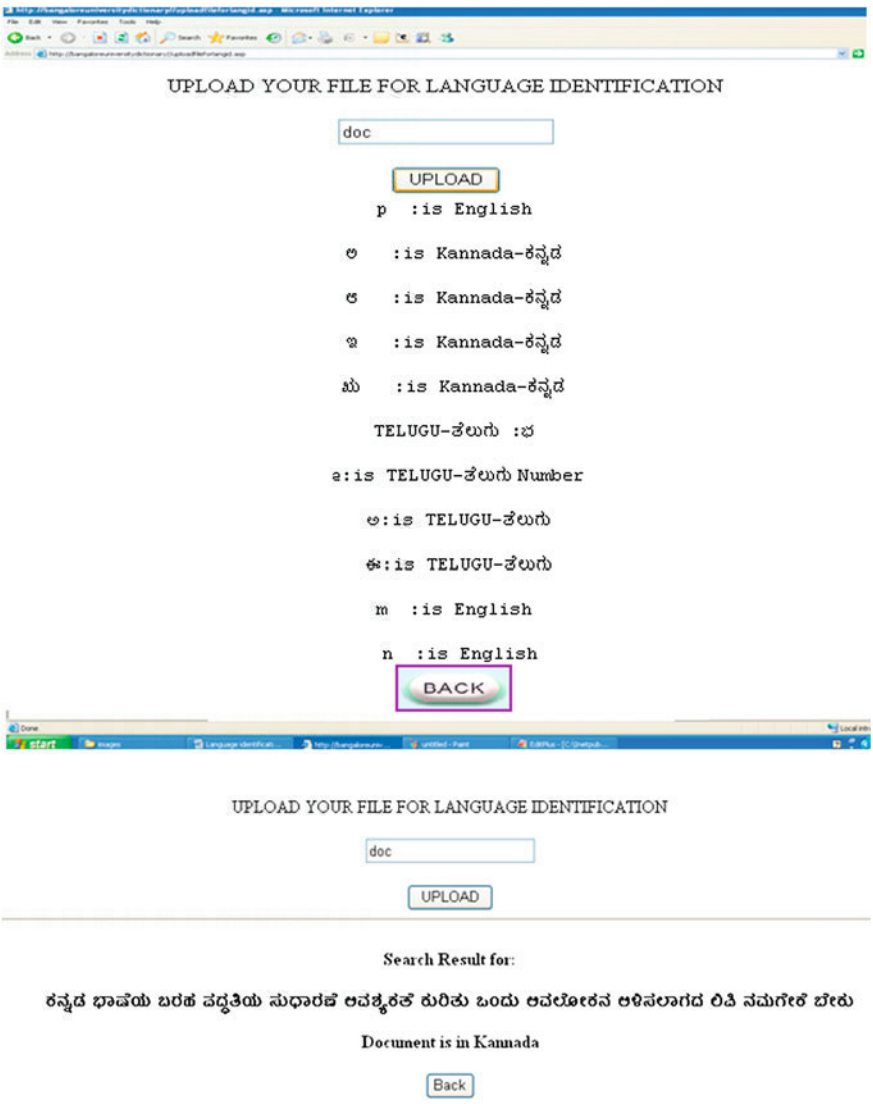

Fig. 1 Language identification for English, Kannada and Telugu by uploading docs

```
Do for i = 1 to 3 Language document types
Do for k = 1 to 100 text lines of ith document
Compare until i == k if yes display the type of Lang.
Otherwise display the unknown language
```

3 Transliteration

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 CLIR [10]. 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. Mapping of characters are used for Transliteration as shown in Figs. 2 and 3.

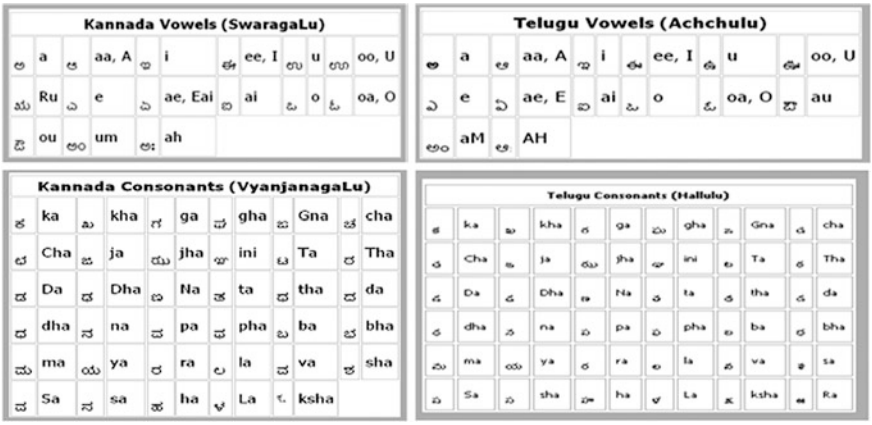

Fig. 2 English–Kannada/Telugu character mapping

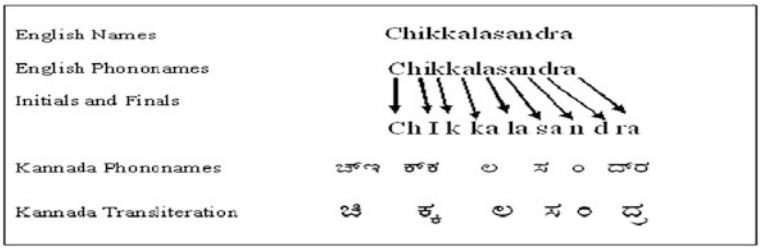

Fig. 3 English–Kannada name transliteration

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.
- **Pronounceable**: The resulting characters have reasonable pronunciations in the target script.
- **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 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 character or two English characters to single Kannada/Telugu characters. Algorithm followed for making dictionary is as follows:

```
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 ame_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
```

4 Translation

Machine translation [11, 12] systems that produce translations between only two particular languages are called bilingual systems and those that produce translations for any given pair of languages are called multilingual systems. Multilingual systems may be either uni-directional or bi-directional. The ideal aim of machine translation systems is to produce the best possible translation without human assistance. Query translation module with Bilingual Dictionary is depicted in Fig. 4.

Kannada and Telugu, like other Indian languages, are morphologically rich. Therefore, we stem the query words before looking up their entries in the bilingual 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 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:

- **Declinable words** (namapada): Morphology of declinable words shown in Fig. 5, as in many Dravidian languages is fairly simple compared to verbs.

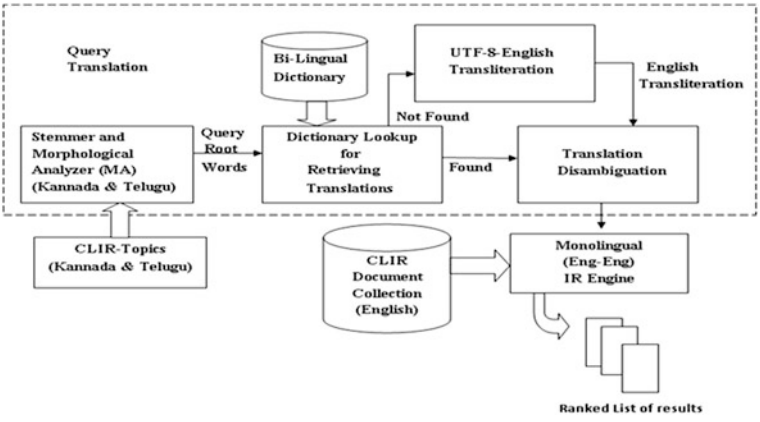

Fig. 4 Query based translation module

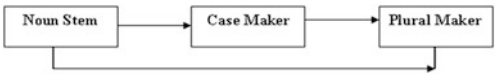

Fig. 5 Formal grammar for Kannada nouns

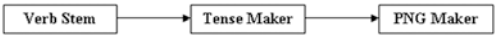

Fig. 6 A formal grammar for Kannada verbs

Kannada words are of three genders- masculine, feminine and neutral. Declinable and Conjugable words have two numbers- singular and plural.

- **Verbs** (kriyapada) or Conjugable words: The verb is much more complex than the nouns. There are three persons namely first, second and third person. Tense of verbs as shown in Fig. 6. is past, present or future. Aspect may be simple, continuous or perfect. Verbs 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.
- **Uninflected words (avyaya)**: Uninflected words may be classified as adverbs, postpositions, conjunctions and interjections. Some of the example 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.

- **Morphological analysis and generation**: Morphological analysis [13] determines the word form such as inflections, tense, number, part of speech, etc. shown in "Table 1" and Fig. 7. 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 1 Different cases and their corresponding and few inflections of a verb stem

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

Fig. 7 Characteristic suffixes for nouns 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

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 Fig. 8. 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 gives information concerning morphological properties of the words it analyses.

In this section we are going to describe about the new algorithm which is developed for morphological analyzer [13] and generator. The main advantage for this algorithm is simple and accurate.

```
Algorithm
1: Get the word to be analyzed.
2: find entered word is found in the Root Dict.
3: If the word is found in the Dict, stop; Else
```

Fig. 8 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

```
4: Separate any suffix from the right hand side
5: If any suffix is present in the word, then check the
availability of the suffix in the dictionary. Then
6: Remove the suffix present, Then re-initialize the word
without identified suffix, Go to Step 2.
7: Repeat until the Dictionary finds the root/stem word.
8: Store the English root/stem word in a variable and then
get the corresponding Kannada word from the bilingual
dictionary
9: Check what all grammatical features does the English word
have given and then generate the corresponding features for
the Kannada word
10: Exit.
```

- **Dictionary based approach**: Dictionary based translation [14] is basically translation with a help of a bi-lingual dictionary. Only translation words with high coherence scores will be selected for the translation of the query as shown in Fig. 9. Query translation 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.
- **Rule-Based Approach**: This approach 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.

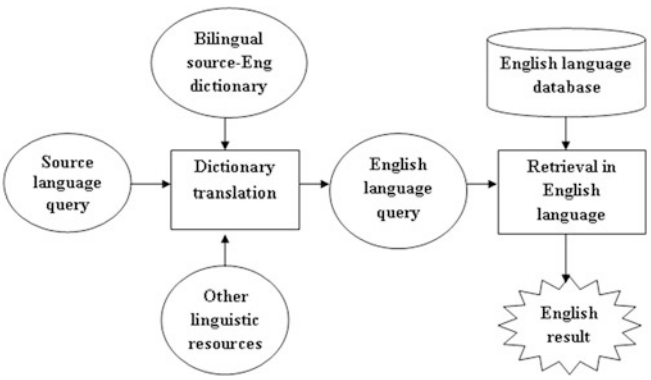

Fig. 9 Dictionary based method for query translation

4.3 The Selection of Word Translation

Normally in CLIR words that are not included in phrases are translated word-by-word shown in Fig. 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 language query words, we select for each of them the best translation word such that it co-occurs most often with other translation words in destination language documents. For example as shown in Fig. 10.

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 $\{a_1 \ldots a_n\}$, 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 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 and works based on term similarity. The EMMI weighting measure has been successfully used to estimate the term similarity in [7]. 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

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

where

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

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

Fig. 10 Word-by-word translation

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).

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

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, is shown in Eq. 4.

5 Experimental Setup

We use machine-readable bi-lingual Kannada → English and Telugu → English dictionaries created by BUBShabdasagar. The Kannada → English bi-lingual dictionary has around 14,000 English entries and 40,000 Kannada entries. The Telugu → English bi-lingual has relatively less coverage and has around 6,110 entries. CLIR Tool [15] is developed by using the ASP.NET as front end and Database as back end. We have trained the systems with corpus size of 200, 500 and 1,000 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 Figs. 11 and 12. The comparative results are shown in Figs. 13 and 14.

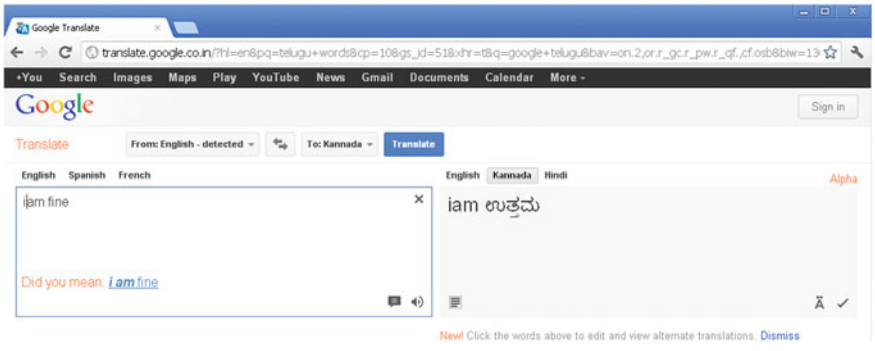

Fig. 11 Google translation

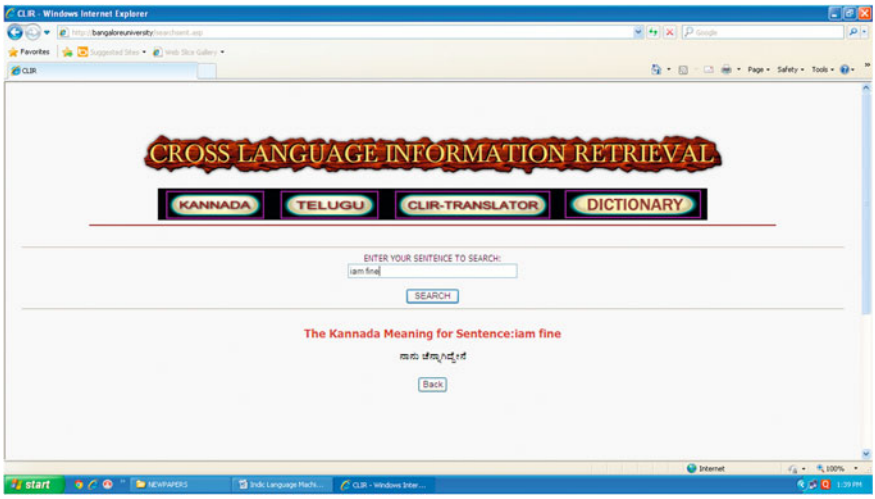

Fig. 12 CLIR translation

5.1 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:

$$WordPrecision\,(WP) = \frac{number\ of\ correctly\ extracted\ word}{number\ of\ extracted\ words} \tag{5}$$

$$WordRecall\,(WR) = \frac{number\ of\ correctly\ extracted\ Words}{number\ of\ correct\ words} \tag{6}$$

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 Fig 15.

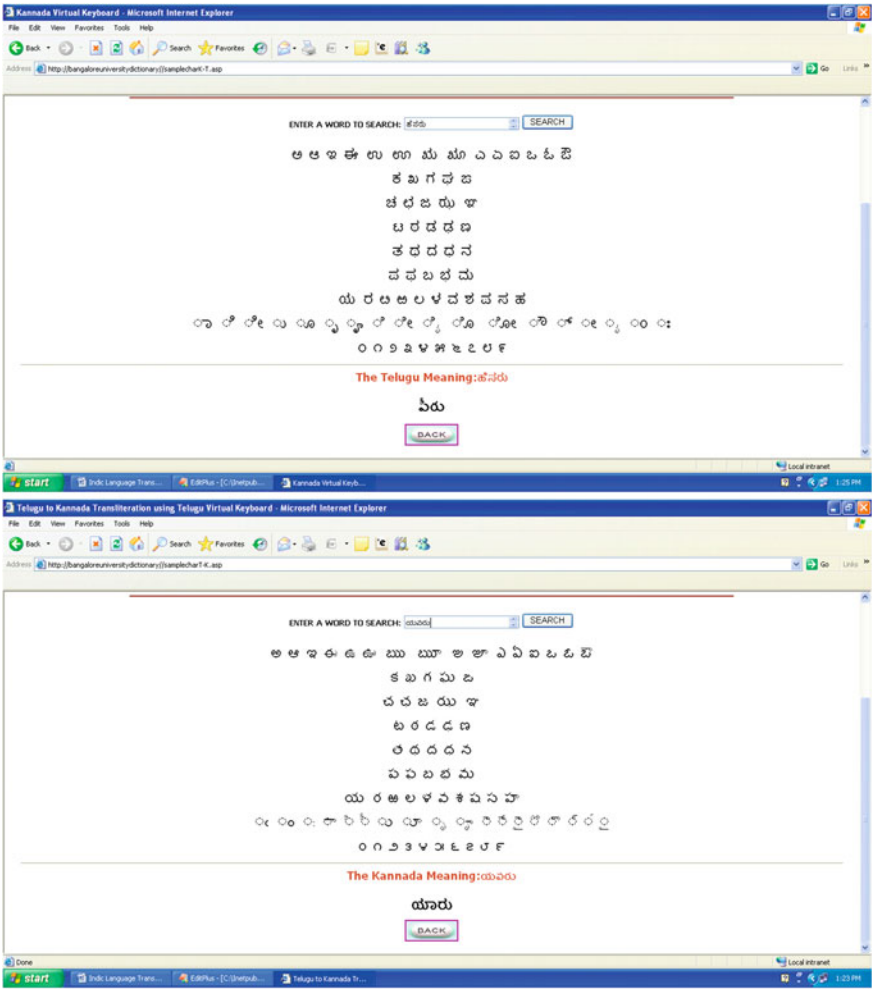

Fig. 13 Sample results for word

Fig. 14 Sample results for sentences

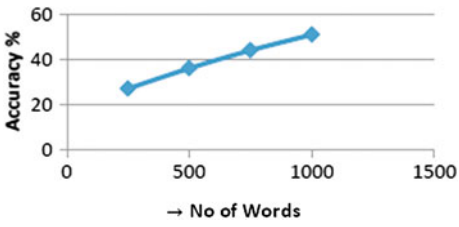

Fig. 15 Performance graph

6 Conclusion and Future Work

In this paper, we presented our Kannada → English and Telugu → English CLIR system developed for the Ad-Hoc bilingual Task. Separate text lines of English, Kannada and Telugu documents from a trilingual document are presented for Natural Language Identification. 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 challenges in CLIR 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.

Acknowledgments 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. I 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. Konchady M (2008) Text mining application programming, 3rd edn. Charles River Media, Boston
2. Homiedan AH (2010) Machine translation. http://faculty.ksu.edu.sa/homiedan/Publications/Machine%20Translation.pdf. Accessed 4 Sep 2011
3. The Karnataka Official Language Act. Official website of Department of Parliamentary Affairs and Legislation. Government of Karnataka. Retrieved 29 July 2012
4. http://en.wikipedia.org/wiki/Telugu_language Retrieved 29 July 2012
5. http://www.ssec.wisc.edu/~tomw/java/unicode.html#x0C80. Retrieved 29 July 2012
6. Botha G, Zimu V, Barnard E (2007) Text-based language identification for South African languages. Published in South African Institute Of Electrical Engineers vol 98 (4)
7. Sibun P, Reynar JC (1996) Language identification: examining the issues. http://www.cis.upenn.edu/~nenkova/Courses/cis430/languageIdentification.pdf. Accessed on 10 Jan 2012

8. Vatanen T, Väyrynen JJ, Virpioja S (2010) Language identification of short text segments with N-gram models
9. Ahmed B, Cha SH, Tappert C (2004) Language identification from text using N-gram based cumulative frequency addition. In: Proceedings of Student/Faculty Research Day, CSIS, Pace University
10. Pingali P, Varma V (2006) Hindi and Telugu to English cross language information retrieval at CLEF 2006, 20–22 Sep, Alicante, Spain
11. Kereto S, Wongchaisuwat C, Poovarawan Y (1993) Machine translation research and development. In: Proceedings of the Symposium on Natural Language processing in Thailand, pp. 167–195
12. Knowles F (1982) The pivotal role of the dictionaries in a machine translation system. In: Lawson V (ed) Practical experience of machine translation. North-Holland, Amsterdam
13. Ritchie G, Whitelock (eds) (1985) The lexicon. pp. 225–256
14. Ballesteros L, Croft WB (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proceedings of ACM SIGIR Conference 20: 84–91
15. Reddy MV, Hanumanthappa M (2009) CLIR Project (English to Kannada and Telugu). Available at http://bangaloreuniversitydictionary//menu.asp