Kannada and Telugu Native Languages to English Cross Language Information Retrieval

Mallamma V Reddy, Dr. M. Hanumanthappa

Department of Computer Science and Applications,
Bangalore University, Bangalore, INDIA.

***Abstract*— One of the crucial challenges in cross lingual information retrieval is the retrieval of relevant information for a query expressed in as native language. While retrieval of relevant documents is slightly easier, analysing the relevance of the retrieved documents and the presentation of the results to the users are non-trivial tasks. To accomplish the above task, we present our Kannada English and Telugu English CLIR systems as part of Ad-Hoc Bilingual task. We take a query translation based approach using bi-lingual dictionaries. When a query words not found in the dictionary then the words are transliterated using a simple rule based approach which utilizes the corpus to return the 'k' closest English transliterations of the given Kannada/Telugu word. The resulting multiple translation/transliteration choices for each query word are disambiguated using an iterative page-rank style algorithm which, based on term-term co-occurrence statistics, produces the final translated query. Finally we conduct experiments on these translated query using a Kannada/Telugu document collection and a set of English queries to report the improvements, performance achieved for each task is to be presented and statistical analysis of these results are given.**

***Keywords*— Kannada-to-English, Telugu-to-English, Cross Language Information Retrieval, Query Translation.**

I. INTRODUCTION

Cross Language Information Retrieval (CLIR) can be defined as the process of retrieving information present in a language different from the language of the user's query. A typical CLIR scenario is shown in the Fig. 1 Where a user needs to retrieve documents from different Indian Languages using query in English. CLIR bridges the gap between information need (query) and the available content (documents).

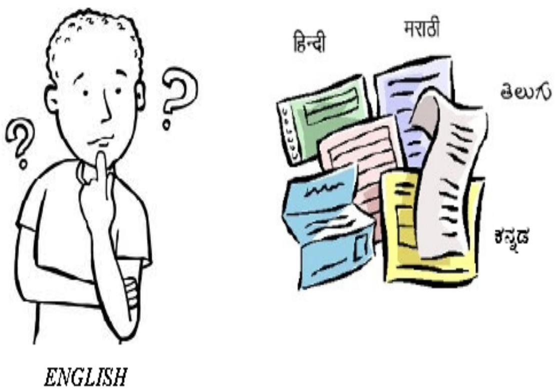

Fig. 1 CLIR scenario of a user trying to access information present in different languages [6]

The World Wide Web (WWW), a rich source of information, is growing at an enormous rate with an estimate of more than 29.7 billion pages on the World Wide Web as of February 2007. According to a survey conducted by Netcraft [is an Internet services company], English is still the dominant language on the web. However, global internet usage statistics reveal that the number of non-English internet users is steadily on the rise. Making this huge repository of information on the web, which is available in English, accessible to non-English internet users worldwide has become an important challenge in recent times.

The above problem is solved by Cross-Lingual Information Retrieval (CLIR) by allowing users to pose the query in a language (source language) which is different from the language (target language) of the documents that are searched. This enables users to express their information need in their native language while the CLIR system takes care of matching it appropriately with the relevant documents in the target language. To help in identification of relevant documents, each result in the final ranked list of documents is usually accompanied by an automatically generated short summary snippet in the source language. Later, the relevant documents could be completely translated into the source language.

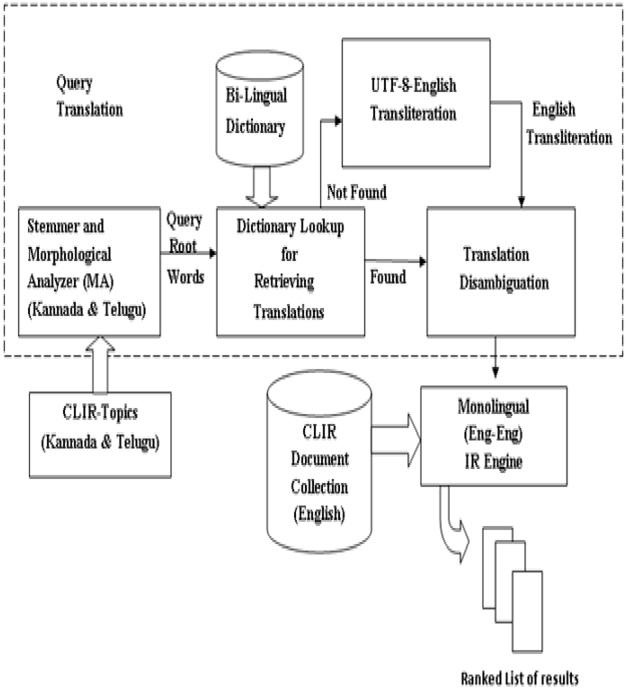
Query Translation
Bi-Lingual Dictionary
UTF-8-English Transliteration
English Transliteration
Not Found
Stemmer and Morphological Analyzer (MA) (Kannada & Telugu)
Query Root Words
Dictionary Lookup for Retrieving Translations
Found
Translation Disambiguation
CLIR-Topics (Kannada & Telugu)
CLIR Document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

Fig. 2 System *Architecture of our CLIR System*

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 [2]. 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" code and draw their vocabulary mainly from Sanskrit.

In this paper, we describe our Kannada English and Telugu English CLIR approaches for the Ad-Hoc Bilingual task. We also present our approach for the English-English Ad-Hoc Monolingual task. The organization of the paper is as follows: Section 2 explains the architecture of our CLIR system. Section 3 describes the algorithm used for English-English monolingual retrieval. Section 4 presents the approach used for Query Transliteration. Section 5 explains the Translation Disambiguation module. Section 6 describes the experiments. Finally, Section 7 concludes the paper highlighting some potential directions for future work.

II. SYSTEM ARCHITECTURE

We use a Query Translation 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 Kannada→English and Telugu→English dictionaries created by BUBShabdasagar for query translation this is depicted in Fig. 2. 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 6110 entries.

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. Algorithm 1 clearly depicts the entire flow of our system.

```
Algorithm1. Query Translation Approach
1:  Remove all the stop words from query
2:  Stem the query words to find the root words
3:  for stem_i Є stems of query words do
        4:  Retrieve all the possible translations from bilingual
        dictionary
5:     if list is empty then
6:     Transliterate the word using to produce    candidate
transliterations
7:     end if
8:   end for
9:  Disambiguate the various translation/transliteration candidates
for each word
10:  Submit the final translated English query to
        English →English Monolingual IR     Engine.
```

III. ENGLISH ENGLISH MONOLINGUAL

We used the standard Okapi BM25 Model [3] for English English monolingual retrieval. Given a query Q, containing keywords $q_1 \ldots q_n$, the BM25 score of a document D is:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D)+k_1 \cdot \left(1-b+b \cdot \frac{|D|}{avgdl}\right)}, \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where $f(q_i, D)$ is q_i's term frequency in the document D, $|D|$ is the length of the document D in words, k_1 & b are free parameters to be set, and *avgdl* is the average document length of documents in corpus. In our current experiments, we set the value of $k_1 = 2.0$ and $b = 0.75$. IDF (q_i) is the IDF (inverse document frequency) weight of the query term q_i. It is usually computed as:

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \qquad \ldots\ldots(2)$$

Where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i.

IV. UTF-8 ENGLISH TRANSLITERATION

Many proper nouns of English like names of people, places and organizations, used as part of the Kannada or Telugu query, are not likely to be present in the Kannada→English and Telugu→English bi-lingual dictionaries. Fig. 3 presents an example Kannada topic CLIR Record Number 199.

<num>199</num>
<word>ಮೂಡಲಮನೆ ಮತ್ತು ಸೂರ್ಯೋದಯ<\word>

Fig. 3 CLIR Record Number 199

In the above topic, the word ಮೂಡಲಮನೆ is "MUdalamane" written in UTF-8/Western Windows Encoding, Such words are to be transliterated to English. There are many standard formats possible for Devanagari English transliteration viz. ITRANS, IAST, ISO 15919, etc. but they all use small and capital letters, and diacritic characters to distinguish letters uniquely and do not give the actual English word found in the corpus.

We use a simple rule based approach which utilizes the corpus to identify the closest possible transliterations for a given Kannada/Telugu word. We create a lookup table which gives the roman letter transliteration for each Devanagari letter. Since English is not a phonetic language, multiple transliterations [5] are possible for each Devanagari letter. In our current work, we only use the most frequent transliteration. A Devanagari word is scanned from left to right replacing each letter with its corresponding entry from the lookup table. For e.g. a word ಗಂಗೋತ್ರಿ is transliterated as shown in Table 1.

TABLE I
TRANSLITERATION EXAMPLE

Input Letter	Output String
ಗ	ga
ಂ	gan
ಗಂ	ganga
ೋ	gango
ತ್ರಿ	gangotri

The above approach produces many transliterations which are not valid English words. For example, for the word "[illegible]" (Australian), the transliteration based on the above approach will be "astreliyai" which is not a valid word in English? Hence, instead of directly using the transliteration output, we compare it with the unique words in the corpus and choose 'k' words most similar to it in terms of string edit distance. For computing the string edit distance, we use the dynamic programming based implementation of Levenshtein Distance [1] metric which is the minimum number of operations required to transform the source string into the target string. The operations considered are insertion, deletion or substitution of a single character.

Using the above technique, the top 3 closest transliterations for "[illegible]" were "australian", "australia" and "estrella". Note that we pick the top 3 choices even if our preliminary transliteration is a valid English word and found in the corpus. The exact choice of transliteration is decided by the translation disambiguation module based on the term-term co-occurrence statistics of a transliteration with translations/transliterations of other query terms.

V. Translation Disambiguation

Given the various translation and transliteration choices for each word in the query, the aim of the Translation Disambiguation module is to choose the most probable translation of the input query Q. In word sense disambiguation, the sense of a word is inferred based on the company it keeps i.e. based on the words with which it co-occurs. Similarly, the words in a query, although less in number, provide important clues for choosing the right translations/transliterations.

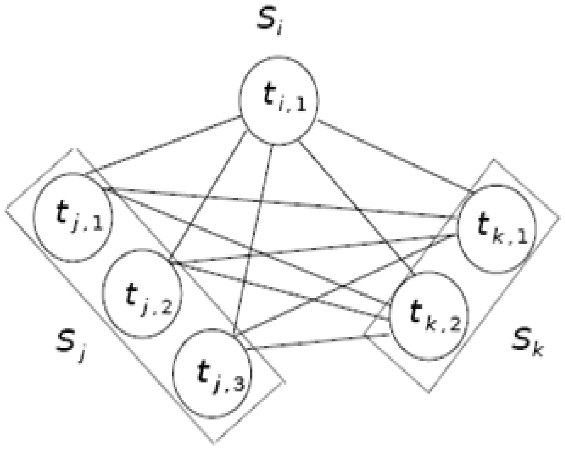

Fig. 4 Co-occurrence Network for Disambiguating Translations/Transliterations

For example, for a query "[illegible] [illegible]". the translation for [illegible] is {green} and the translations for [illegible] are {hang, designing}. Here, based on the context, we can see that the choice of translation for the second word is water since it is more likely to co-occur with river.

Assuming we have a query with three terms, s1, s2, s3, each with different possible translations/transliterations, the most probable translation of query is the combination which has the maximum number of occurrences in the corpus. However, this approach is not only computation ally expensive but also run into data sparsity problem. We use a page-rank style iterative disambiguation algorithm [4] which examines pairs of terms to gather partial evidence for the likelihood of a translation in a given context.

Consider three words s_i, s_j, s_k, as shown in Fig. 4 with multiple translations. Let their translations be denoted as {{t_i, 1}, {t_j.1, tj.2, tj.3}, {t_k.1, t.2}}. Given this, a co-occurrence network is constructed as follows: the translation candidates of different query terms are linked together. But, no links exist between different translation candidates of a query term. In the above graph, a weight w ($t|s_i$), is associated to each node t which denotes the probability of the candidate being the right translation choice for the input query Q. A weight, l (t, t'), is also associated to each edge (t, t') which denotes the association measure between the words t and t'.

Initially, all the translation candidates are assumed to be equally likely.

- *Initialization step*

$$w^{0}(t|S_{i}) - \frac{1}{|t_r(S_i)|} \tag{3}$$

TABLE II
Mathematical Symbols Involved In Translation Disambiguation

Symbol	Explanation
S_i	Source word
$t_r(S_i)$	Set of translations for word S_i
t	Translation candidate, t Є $t_r(S_i)$
$w(t\|S_i)$	weight of node t, where S_i is the source word
l(t, t')	Weight of link between nodes t and t'
$t_{i,m}$	m^{th} translation of i^{th} source word

TABLE III
Details of Document Collection

Number of Terms	14,000
Number of Unique Terms	5000
Average Document Length	98

After initialization, each node weight is iteratively updated using the weights of nodes linked to it and the weight of link connecting them.

- *Iteration step*

$$w^{n}(t|s_{i}) = w^{n-1}(t|s_{i}) + \sum_{t' \in inlink(t)} l(t,t') * w^{n-1}(t'|s) \tag{4}$$

Where s is the corresponding source word for translation candidate t and inlink(t) is the set of translation candidates that are linked to t. After each node weight is updated, the weights are normalized to ensure they all sum to one.

- *Normalization step*

$$w^{n}(t|s_{i}) = \frac{w^{n}(t|s_{i})}{\sum_{m=1}^{|t_r(S_i)|} w^{n}(t_{i,m}|s_i)} \tag{5}$$

Steps 4 and 5 are repeated iteratively till convergence. Finally, the two most probable translations for each source word are chosen as candidate translations.

- *Link-weights computation*

The link weight, which is meant to capture the association strength between the two words (nodes), could be measured using various functions. In our current work, we use two such functions: Dice Coefficient and Point-wise Mutual Information (PMI).

Point-wise Mutual Information (PMI) [7] is defined as follows:

$$I(t,t') = PMI(t,t') = log_2 \frac{p(t,t')}{p(t)*p(t')} \qquad (6)$$

Where $p(t,t')$ is the joint probability of t and t . p(t) and p(t‘) are the marginal probabilities of t and t respectively. If the two terms are highly related then their joint probability will be higher when compared to the pro duct of their marginals. Therefore, their PMI will in turn be higher. The joint probability $p(t,t')$is computed by considering the co-occurrence of the terms t and t and dividing it with all possible term combinations. The marginal probability p(t) is the probability of finding the term independently in the entire corpus.

$$p(t,t') = \frac{freq(t,t')}{avgdl*avgdl} \qquad (7)$$

$$p(t) = \frac{freq(t)}{N} \qquad (8)$$

Where $freq(t,t')$is the number of times t and t' co-occur in the entire corpus, freq(t) is the number of times t occurs in the corpus, N is the number of words in the entire corpus, avgdl is the average document length.

Dice Coefficient (DC) is defined as follows:

$$I(t,t') = DC(t,t') = \frac{2*freq(t,t')}{freq(t)+freq(t')} \qquad (9)$$

As we can see, similar to PMI, Dice Coefficient also tries to capture the degree of relatedness between terms only using a different ratio.

We used the standard implementation of Okapi MB25 in Trec for our runs. The dictionaries were indexed after stemming (using Porter stemmer) and stop word removal. The dictionary consists of 5,000 words each in Kannada and Telugu. We used the Kannada and Telugu stemmers and morphological analysers for stemming the words description and RunID is shown in Table 4 and 5.

TABLE IV
CLIR EVALUATION FOR RUN DESCRIPTION

Run ID	Description
EK	English Queries and Kannada Documents
ET	English Queries and Telugu Documents

TABLE V
DETAILS OF RUNS SUBMITTED

SI. No.	*Description*	*Run ID*
1.	English-English Monolingual	EN-MONO-WORD
2.	Kannada-English Bilingual Word with DC	BUBShabdasagar Kannada Word DICE
3.	Kannada-English Bilingual Word with PMI	BUBShabdasagar Kannada Word PMI
4.	Telugu-English Bilingual Word with DC	BUBShabdasagar Telugu Word DICE
5.	Telugu-English Bilingual Word with PMI	BUBShabdasagar Telugu Word PMI

VI. EXPERIMENTS AND RESULTS

We use the following standard measures for evaluation [8]: Mean Average Precision (MAP), R-Precision, and Precision at 5, 10 and 20 do cuments (P@5, P@10 and P@20) and Recall.

Since different systems may be using different monolingual retrieval algorithms, to facilitate comparison, we also report the percentage with respect to monolingual retrieval for each performance figure. The overall results are tabulated in Table 6.

For Kannada, we overcome a Mean Average Precision of (MAP) 0.3356 in Word which is 65.13% of monolingual performance. For Telugu, we get a MAP of 0.2256 in word which is 57.10% of monolingual performance. The recall levels in Kannada are 73.58% for word runs which is 89.79% of monolingual. The recall levels in Telugu are 64.23% in word run which is 77.10% of monolingual.

TABLE VI
MONOLINGUAL AND BILINGUAL OVERALL RESULTS (PERCENTAGE OF MONOLINGUAL PERFORMANCE GIVEN IN BRACKETS BELOW THE ACTUAL NUMBERS)

WORD ONLY						
Run Desc	MAP	R-precision	p@5	p@10	p@20	Recall
EN-MONO-WORD	0.4556	0.4520	0.5540	0.4565	0.4910	82.40%
KANNADA_WORD_DICE	0.3356 (65.13%)	0.3466 (69.35%)	0.4120 (59.33%)	0.3010 (64.30%)	0.2900 (69.89%)	73.58% (89.79)
KANNADA_WORD_PMI	0.2169 (54.29%)	0.2323 (59.10%)	0.2911 (51.96%)	0.2690 (57.98%)	0.2910 (62.07%)	69.42% (85.15%)
TELUGU_WORD_DICE	0.2256 (57.10%)	0.2410 (63.00%)	0.3300 (59.10%)	0.3010 (65.96%)	0.2725 (67.16%)	64.23% (77.10%)
TELUGU_WORD_PMI	0.2010 (51.01%)	0.2238 (56.69%)	0.3732 (60.13%)	0.2821 (59.20%)	0.2369 (58.93%)	56.32% (68.25%)

VII. CONCLUSIONS

We presented our Kannada→English and Telugu→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. It makes use of the corpus to return the 'K' closest possible English transliterations of a given Kannada/Telugu word. Disambiguating the various translations/transliterations is performed using an iterative page-rank style algorithm. Further sentence and phrase translation is to be carried out as part of Ad-Hoc bilingual task for English to Kannada, English toTelugu.

REFERENCES

[1] Gusfield. Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge University Press, 1997.

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

[3] Karen Sparck Jones, Steve Walker, and Stephen E. Robertson. *A probabilistic model of information retrieval: development and comparative experiments* (parts 1& 2). Information Processing and Management, 36(6):779–840, 2000.

[4] Christof Monz and Bonnie J. Dorr. *Iterative translation disambiguation for cross-language information retrieval.* In SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 520–527, New York, NY, USA, 2005. ACM Press.

[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] Doulas W.Oard. *A comparative study of query and document translation for cross language information retrieval*. In AMTA 98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup, pages 472-483,London, UK,1998.Springer-Verlag.

[7] Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley-Interscience, New York, NY, USA, 1991.