Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation

Manoj Kumar Chinnakotla[*], Sagar Ranadive, Om P. Damani and Pushpak Bhattacharyya

Department of Computer Science and Engineering, IIT Bombay, India

{manoj,sagar,damani,pb}@cse.iitb.ac.in

Abstract. In this paper, we present our Hindi to English and Marathi to English CLIR systems developed as part of our participation in the CLEF 2007 Ad-Hoc Bilingual task. We take a query translation based approach using bi-lingual dictionaries. Query words not found in the dictionary are transliterated using a simple rule based transliteration approach. The resultant transliteration is then compared with the unique words of the corpus to return the 'k' words most similar to the transliterated 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. Using the above approach, for Hindi, we achieve a Mean Average Precision (MAP) of 0.2366 using title and a MAP of 0.2952 using title and description. For Marathi, we achieve a MAP of 0.2163 using title.

1 Introduction

The World Wide Web (WWW), a rich source of information, is growing at an enormous rate. Although English still remains the dominant language on the web, global internet usage statistics reveal that the number of non-English internet users is steadily on the rise. Hence, making this huge repository of information, which is available in English, accessible to non-English users worldwide is an important challenge in recent times.

Cross-Lingual Information Retrieval (CLIR) systems allow the 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 the user in the 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. Using this, the user could single out the relevant documents for complete translation into the source language.

[*] Supported by a fellowship award from Infosys Technologies

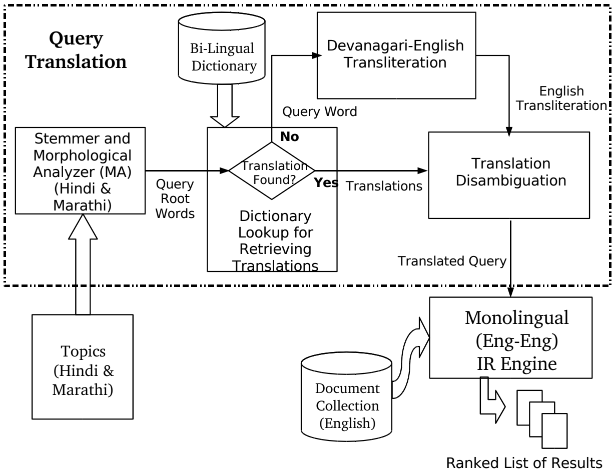

Query Translation
Bi-Lingual Dictionary
Devanagari-English Transliteration
English Transliteration
Query Word
No
Stemmer and Morphological Analyzer (MA) (Hindi & Marathi)
Query Root Words
Translation Found?
Yes
Translations
Translation Disambiguation
Dictionary Lookup for Retrieving Translations
Translated Query
Monolingual (Eng-Eng) IR Engine
Topics (Hindi & Marathi)
Document Collection (English)
Ranked List of Results

Fig. 1. CLIR System Architecture

Hindi is the official language of India along with English and according to *Ethnologue*[1], it is the fifth most spoken language in the world. *Marathi* is a widely spoken language in the state of Maharashtra. Both Hindi and Marathi use the "Devanagari" script.

In this paper, we describe our Hindi to English and Marathi to English CLIR approaches for the CLEF 2007 Ad-Hoc Bilingual task. The architecture of our CLIR system is shown in Figure 1. 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 Hindi to English and Marathi to English dictionaries created by Center for Indian Language Technologies (CFILT), IIT Bombay for query translation. The Hindi to English bi-lingual dictionary has around 115,571 entries and is also available online[2]. The Marathi to English bi-lingual has less coverage and has around 6110 entries.

Hindi and Marathi, 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 transliterated by the Devanagari to English transliteration module. The above module, based on a simple lookup table and index, returns top three English words from the corpus which are most similar to the source query word. Finally, the translation disambiguation module disambiguates the multiple translations/transliterations returned for the query and returns the most probable English translation of the

[1] `http://www.ethnologue.com`
[2] `http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/dict_search_user.php`

```
<num>10.2452/445-AH</num>
<title>प्रिन्स हैरी और नशीली दवाएं</title>
<desc>ऐसे दस्तवेज खोजिये जिनमे प्रिन्स हैरी द्वारा नशीली दवाएं ग्रहण किए जाने की
कोई रिपोर्ट हो</desc>
```

Table 1. A sample CLEF 2007 Hindi Topic: Number 445

original query. The translated query is fired against the monolingual IR engine to retrieve the final ranked list of documents as results.

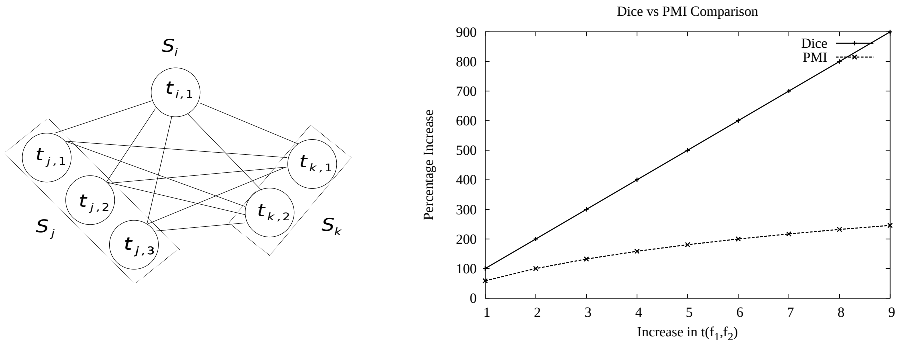
The organization of the paper is as follows: Section 2 presents the approach used for *Query Transliteration*. Section 3 explains the *Translation Disambiguation* module. Section 4 describes the experimental setup, discusses the results and also presents the error analysis. Finally, Section 5 concludes the paper highlighting some potential directions for future work.

2 Devanagari to English Transliteration

Many words of English origin like names of people, places and organizations, are likely to be used as part of the Hindi or Marathi query. Such words are usually not found in the Hindi to English and Marathi to English bi-lingual dictionaries. Table 1 presents an example Hindi topic from CLEF 2007. In the above topic, the word प्रिन्स हैरी is *Prince Harry* written in Devanagari. Such words need to be *transliterated* into English. We use a simple rule based approach which utilizes the corpus to identify the closest possible transliterations for a given Hindi/Marathi 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 are possible for each Devanagari letter. In our current work, we only use a single transliteration for each Devanagari letter. The English transliteration is produced by scanning a Devanagari word from left to right replacing each letter with its corresponding entry from the lookup table. The above approach produces many transliterations which are not valid English words. For example, for the word आस्ट्रेलियाई (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 indexed words in the corpus and choose the 'k' most similar indexed words in terms of *string edit distance*. For computing the string edit distance, we use the dynamic programming based implementation of *Levenshtein Distance* [1] metric.

Using the above technique, the top 3 closest transliterations for आस्ट्रेलियाई were *australian*, *australia* and *estrella*. Note that we pick the top 3 choices even if our preliminary transliteration is a valid English word. The final choice of transliteration for the source term is made by the translation disambiguation module based on the term-term co-occurrence statistics of the transliteration with translations/transliterations of other query terms.



(a) Co-occurrence Graph [2]

(b) Dice vs. PMI

Fig. 2. Translation Disambiguation: Co-occurrence Graph for Disambiguating Translations and Transliterations, Comparison of Dice Coefficient and PMI

3 Translation Disambiguation

Given the various translation and transliteration choices for the query, the Translation Disambiguation module, out of the various possible combinations, selects the *most probable* translation of the input query Q. The context within a query, although small, provides important clues for choosing the right translations/transliterations of a given query word. For example, for a query "नदी जल" (River Water), the translation for नदी is $\{$*river*$\}$ and the translations for जल are $\{$*water, to burn*$\}$. Here, based on the context, we can see that the choice of translation for the second word is *water* since the combination $\{$*river, water*$\}$ is more likely to co-occur in the corpus than $\{$*river, burn*$\}$.

Consider a query with three words $Q = \{s_i, s_j, s_k\}$. Let $tr(s_j) = \{t_{j,1}, t_{j,2}, \ldots, t_{j,l}\}$ denote the set of translations and transliteration choices corresponding to a given source word s_j where l is the number of translations found in dictionary for s_j. The set of possible translations for the entire query Q is $T = \{tr(s_i), tr(s_j), tr(s_k)\}$. As explained earlier, out of all possible combinations of translations, the most probable translation of query is the combination which has the maximum number of co-occurrences in the corpus. However, this approach is not only computationally expensive but may also run into data sparsity problem. Hence, we use a page-rank style iterative disambiguation algorithm proposed by Christof Monz *et. al.* [2] which examines pairs of terms to gather partial evidence for the likelihood of a translation in a given context.

3.1 Iterative Disambiguation Algorithm

Given a query Q and the translation set T, a co-occurrence graph is constructed as follows: the translation candidates of different query terms are linked together. But, no edges exist between different translation candidates of the same query term as shown in Figure 3 (a). In the above graph, $w^n(t|s_i)$ is the weight associated with node t at iteration n and denotes the probability of the candidate t being the right translation choice for the input query word s_i. A weight $l(t, t')$, is also assigned to each edge (t, t') which denotes the strength of relatedness between the words t and t'.

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

$$w^0(t|s_i) = \frac{1}{|tr(s_i)|} \tag{1}$$

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:

$$w^n(t|s_i) = w^{n-1}(t|s_i) + \sum_{t' \in inlink(t)} l(t, t') * w^{n-1}(t'|s) \tag{2}$$

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:

$$w^n(t|s_i) = \frac{w^n(t|s_i)}{\sum_{m=1}^{|tr(s_i)|} w^n(t_m|s_i)} \tag{3}$$

Steps 2 and 3 are repeated iteratively till they converge approximately. 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 (vertices), could be measured using various functions. In this work, we use two such functions: *Dice Coefficient (DC)* and *Point-wise Mutual Information (PMI)*.

PMI [3] is defined as follows:

$$l(t, t') \ = \ PMI(t, t') \ = \ log_2 \, \frac{p(t, t')}{p(t) * p(t')} \tag{4}$$

where $p(t, t')$ is the joint probability of t and t' *i.e.* the probability of finding the terms t and t' together, in a given context, in the corpus. $p(t)$ and $p(t')$ are the marginal probabilities of t and t' respectively *i.e.* the probability of finding these terms in the entire corpus.

Title Only

Run Desc.	MAP	R-Precision	P@5	P@10	P@20	Recall
EN-MONO-TITLE	0.3856	0.3820	0.5440	0.4560	0.3910	81.40%
IITB_HINDI_TITLE_DICE	0.2366	0.2468	0.3120	0.2920	0.2700	72.58%
	(61.36%)	(64.60%)	(57.35%)	(64.03%)	(69.05%)	(89.16%)
IITB_HINDI_TITLE_PMI	0.2089	0.2229	0.2800	0.2640	0.2390	68.53%
	(54.17%)	(58.35%)	(51.47%)	(57.89%)	(61.12%)	(84.19%)
IITB_MAR_TITLE_DICE	0.2163	0.2371	0.3200	0.2960	0.2510	62.44%
	(56.09%)	(62.07%)	(58.82%)	(64.91%)	(64.19%)	(76.70%)
IITB_MAR_TITLE_PMI	0.1935	0.2121	0.3240	0.2680	0.2280	54.07%
	(50.18%)	(55.52%)	(59.56%)	(58.77%)	(58.31%)	(66.42%)
Title + Description						
EN-MONO-TITLE+DESC	0.4402	0.4330	0.5960	0.5040	0.4270	87.67%
IITB_HINDI_TITLEDESC_DICE	0.2952	0.3081	0.3880	0.3560	0.3150	76.55%
	(67.06%)	(71.15%)	(65.10%)	(70.63%)	(73.77%)	(87.32%)
IITB_HINDI_TITLEDESC_PMI	0.2645	0.2719	0.3760	0.3500	0.2950	72.76%
	(60.08%)	(62.79%)	(63.09%)	(69.44%)	(69.09%)	(82.99%)

Table 2. CLEF 2007 Ad-Hoc Monolingual and Bilingual Overall Results (Percentage of monolingual performance given in brackets below the actual numbers)

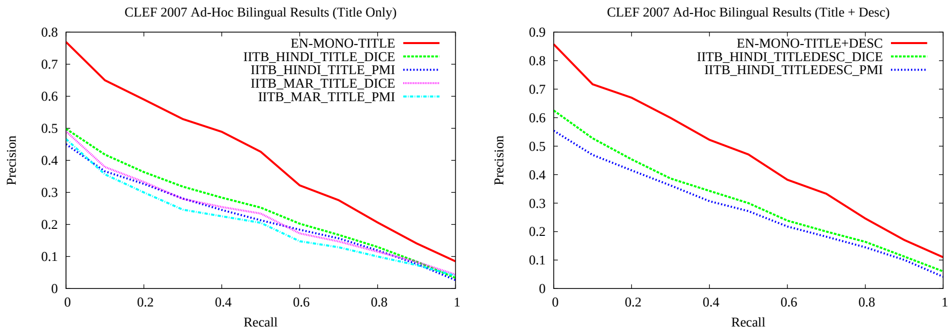


Fig. 3. CLEF 2007 Ad-Hoc Monolingual and Bilingual Precision-Recall Curves

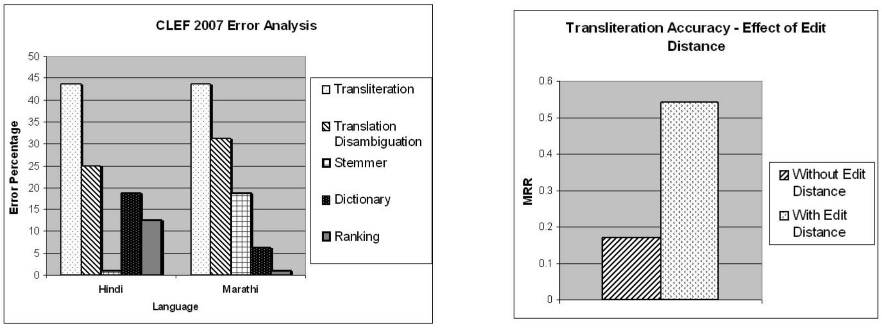
DC is defined as follows:

$$l(t,t') = DC(t,t') = \frac{2 * freq(t,t')}{freq(t) + freq(t')} \qquad (5)$$

where $freq(t,t')$, $freq(t)$ and $freq(t')$ are the combined and individual frequency of occurrence of terms t and t' respectively. For computing $freq(t,t')$, which is needed for both the measures, we consider co-occurrences at the document level.

4 Experiments and Results

We used *Trec Terrier* [4] as the monolingual English IR engine and Okapi BM25 as the ranking algorithm. The details of the topics and document set are given



(a) Percentage of errors module-wise

(b) Effect of edit distance

Fig. 4. CLEF 2007 Analysis of Results

in [6]. The documents were indexed after stemming (using Porter Stemmer) and stop-word removal. We used the Hindi and Marathi stemmers and morphological analyzers developed at CFILT, IIT Bombay for stemming the topic words. For each of the Title and Title + Description runs, we tried DC and PMI for calculating the link weight. This gave rise to four runs for Hindi. For Marathi, due to resource constraints, we could not carry out the Title + Description run and only did the Title run.

We use the following standard measures [5] for evaluation: Mean Average Precision (MAP), R-Precision, Precision at 5, 10 and 20 documents and Recall. We also report the percentage of monolingual English retrieval achieved for each performance figure. The overall results are tabulated in Table 2 and the corresponding precision-recall curves appear in Figure 3.

4.1 Discussion

In agreement with the results reported by Christof Monz *et. al.* [2], we observe that, as an association measure, DC consistently performs better than PMI. One reason for this behavior is that DC, when compared to PMI which uses a logarithmic function, is more sensitive to slight variations in frequency counts. Figure 3 (b) depicts this phenomenon where we vary the joint frequency count $f(t_i, t_j)$, keeping the individual term frequencies $f(t_i), f(t_j)$ constant.

The output of the transliteration module is a list of transliterations ranked by edit distance. We evaluated its accuracy on the CLEF 2007 topic words which had to be actually transliterated. We used the standard *Mean Reciprocal Rank (MRR)* metric for evaluation which is defined as: $MRR = \sum_{i=1}^{N} \frac{1}{Rank(i)}$ where $Rank(i)$ is the rank of the correct transliteration in the ranked list. We observe that the simple rule based transliteration works quite well with an MRR of 0.543 *i.e.* on an average it outputs the correct translation at rank 2. The addition of edit distance module drastically improves the accuracy as shown in Fig. 4 (b).

4.2 Error Analysis

We performed an error analysis of all the queries. We categorized these errors based on the modules in which the errors occurred. A graph depicting the percentage error contributions by various modules for each language is shown in Figure 4 (a).

For both Hindi and Marathi, the largest error contribution is due to *Devanagari to English Transliteration*. Since, we only use a single grapheme mapping, it is difficult to capture different spelling variations of a Devanagari word in English. For instance, while transliterating the word “क्वीन” (Queen), the correct mapping for the letter ‘क’ is ‘qa’. However, since we only have a single mapping, ‘क’ is mapped to ‘ka’ and hence it doesn’t get rightly transliterated into *Queen*. The next major source of error is the *Translation Disambiguation* module. Since we have considered document-level co-occurrence, many unrelated words also usually co-occur with the given word due to which the DC/PMI score increases. Other important sources of error were language specific resources like Stemmer and Bi-lingual dictionaries.

5 Conclusion

We presented our Hindi to English and Marathi to English CLIR systems developed for the CLEF 2007 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 Hindi/Marathi word. Disambiguating the various translations/transliterations is performed using an iterative page-rank style algorithm which is based on term-term co-occurrence statistics.

Based on the current experience, we plan to explore the following directions in future: In transliteration, instead of a single rule for each letter, multiple rules could be considered. Calculating the joint frequency count at a more finer level like sentence or n-gram window instead of document-level. To improve ranking, the terms in the final translated query could be augmented with weights.

References

1. Gusfield, D.: Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge University Press (1997)
2. Monz, C., Dorr, B.J.: Iterative translation disambiguation for cross-language information retrieval. In: SIGIR ’05: New York, USA, ACM Press (2005) 520–527
3. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley-Interscience, New York, NY, USA (1991)
4. Ounis, I., Amati, G., V., P., He, B., Macdonald, C., Johnson: Terrier Information Retrieval Platform. In: ECIR ’05. Volume 3408 of LNCS., Springer (2005) 517–519
5. Yates, R.B., Neto, B.R.: Modern Information Retrieval. Pearson Education (2005)
6. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 Ad Hoc Track Overview LNCS., Springer (2007)