

Evaluation of Hindi to English, Marathi to English and English to Hindi CLIR at FIRE 2008

Nilesh Padariya, Manoj Chinnakotla, Ajay Nagesh, Om P. Damani
Dept. of Computer Science and Engineering (CSE)
IIT Bombay
Mumbai, India

{nileshsp,manoj,damani}@cse.iitb.ac.in, ajaynagesh@it.iitb.ac.in

ABSTRACT

In this paper, we present the evaluation of our CLIR system performed as part of our participation in FIRE 2008. We participated in Hindi to English, Marathi to English, English to Hindi bilingual task and English, Hindi, Marathi monolingual task. We take a query translation based approach using bi-lingual dictionaries. Query words not found in the bi-lingual dictionary are *transliterated*. Since Devanagari is a phonetic script, for transliteration from Hindi/Marathi to English, we use a rule-based approach and for transliteration from English to Hindi, we use a segment based transliteration approach. The resultant transliteration/translation candidates for each query word are disambiguated using an iterative page-rank style algorithm which, based on term-term co-occurrence statistics, produces the most probable translation of the query.

In many Indian language documents the actual Indian word is used as-is without translation. For example, to describe *Kashmir Travel*, it is quite common to also use *Kashmir Yatra* in the documents i.e, the actual Hindi word transliterated in English. So, this motivated us to try a full transliteration of the source query without translation. We report the results of this experiment.

1. INTRODUCTION

The World Wide Web (WWW), a rich source of information, is growing at an enormous rate with an estimate of more than 11.5 billion pages by January 2005 [3]. One of the distinguishing features characterizing this growth is *multilinguality*. Although, English still continues to be the dominant language on the web, global internet usage statistics¹ reveal that the number of non-English internet users and content is steadily on the rise. Making this huge repository of information accessible, without any language barrier, to internet users worldwide has become an important challenge in recent times.

Cross-Lingual Information Retrieval (CLIR) systems aim to solve the above problem 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 aid the user in identification of relevant documents, each result in the final ranked list of documents is usually accompanied by an automati-

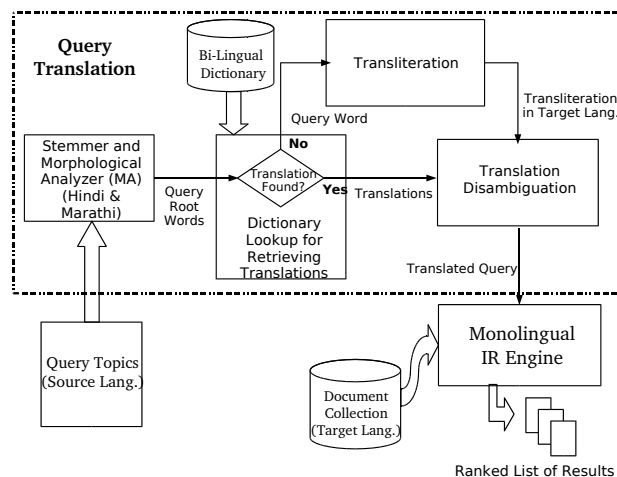


Figure 1: System Architecture of our CLIR System

cally generated short summary snippet in the source language. Later, the user can choose to completely translate a relevant document into the source language for accessing the required information.

Hindi is the official language of India along with English and according to *Ethnologue*², a well-known source for language statistics, it is the fifth most spoken language in the world. It is mainly spoken in the northern and central parts of India. *Marathi* is also one of the widely spoken languages in India especially in the state of Maharashtra. Both Hindi and Marathi use the “Devanagari” script and draw their vocabulary mainly from Sanskrit.

In this paper, we describe our Hindi to English, Marathi to English and English to Hindi CLIR approaches for the FIRE 2008 Ad-Hoc Bilingual task. Besides, we describe the algorithms used for the Ad-Hoc Monolingual task. We also present our experiment where we tried to explore the effectiveness of *transliterating the whole query instead of translation* for Indian Languages. The organization of the paper is as follows: Section 2, explains the architecture of our CLIR system. Section 3 describes the algorithms used for monolingual retrieval. Section 4 outlines the approaches used for *Query Transliteration*. Section 5 explains the *Translation Disambiguation* module. Section 6 discusses the motivation and idea of transliterating the whole query without translation. Section 7 describes the experiments and discusses the

¹<http://www.internetworldstats.com/stats7.htm>

²<http://www.ethnologue.com>

Algorithm 1 Query Translation Approach

```
1: Remove all the stop words from query
2: Stem the query words to find the root words
3: for  $stem_i \in$  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 query to Monolingual IR En-
    gine
```

results. Finally, Section 8 concludes the paper.

2. SYSTEM ARCHITECTURE

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.

Hindi and Marathi, 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 transliteration module. Finally, the translation disambiguation module disambiguates the multiple translations/transliterations returned for each word and returns the most probable translation of the entire query to the monolingual IR engine. Algorithm 1 clearly depicts the entire flow of our system.

3. MONOLINGUAL RETRIEVAL

We used the standard Okapi BM25 Model [5] for English, Hindi and Marathi monolingual retrieval. Given a keyword query $Q = \{q_1, q_2, \dots, q_n\}$ and document D , the BM25 score of the document D is as follows:

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

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

where $f(q_i, D)$ is the term frequency of q_i in D , $|D|$ is length of document D , k_1 & b are free parameters to be set, $avgdl$ is the average length of document in corpus, N is the total no. of documents in collection, $n(q_i)$ is the number of documents containing q_i . In our current experiments, we set the value of $k_1 = 1.2$ and $b = 0.75$.

4. QUERY TRANSLITERATION

<top lang="hi"> <num> 26 </num> <title>सिंगूर और नंदीग्राम भूमि विवाद</title>

Table 1: FIRE 2008 Topic Number 26

Many *proper nouns* of English like names of people, places and organizations, used as part of the source language query, are not likely to be present in the bi-lingual dictionaries. Table 1 presents a sample Hindi topic from FIRE 2008.

In the above topic, the word “सिंगूर” is “Singur” (Place Name) written in Devanagari. Such words are to be transliterated to generate their equivalent spellings in the target language. Since Hindi and Marathi use Devanagari which is a phonetic script, we use a *Rule Based Transliteration* approach for Devanagari to English transliteration. The current accuracy of the system is 80% at a rank of 5.

For English to Hindi transliteration, we use *Substring-Based Transliteration* [4, 8] approach which uses the english substrings as transliteration units instead of just letters.

5. 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 kept *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. For example, for a query “नदी जल”, 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 it is more likely to co-occur with *river*.

Assuming we have a query with three terms, s_1, s_2, s_3 , 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 computationally expensive but may also run into data sparsity problem. We use a page-rank style iterative disambiguation algorithm proposed by Christof Monz *et. al.* [6] which examines pairs of terms to gather partial evidence for the likelihood of a translation in a given context.

5.1 Iterative Disambiguation Algorithm

Consider three words s_i, s_j, s_k , as shown in Figure 2, with multiple translations. Let their translations be denoted as $\{\{t_{i,1}\}, \{t_{j,1}, t_{j,2}, t_{j,3}\}, \{t_{k,1}, t_{k,2}\}\}$. 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.

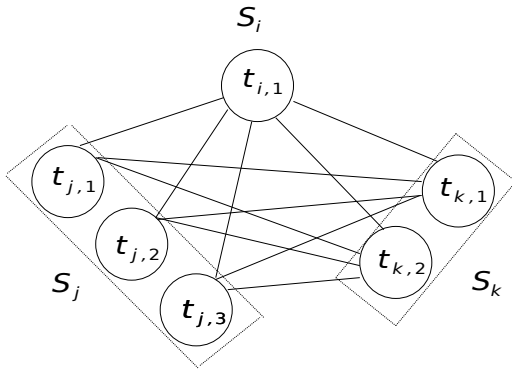


Figure 2: Co-occurrence Network for Disambiguating Translations/Transliterations [6]

Initialization step:

$$w^0(t|s_i) = \frac{1}{|tr(s_i)|} \quad (3)$$

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:

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

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:

$$w^n(t|s_i) = \frac{w^n(t|s_i)}{\sum_{m=1}^{|tr(s_i)|} w^n(t_{i,m}|s_i)} \quad (5)$$

Stpdf 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 *Dice Coefficient* as it has been earlier shown to perform better [2, 6].

Symbol	Explanation
s_i	Source word
$tr(s_i)$	Set of translations for word s_i
t	Translation candidate, $t \in tr(s_i)$
$w(t s_i)$	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 2: Mathematical symbols involved in translation disambiguation

	English	Hindi	Marathi
Documents	125586	95125	99359
Terms	33258828	38835238	26080361
Unique Terms	191619	192351	333361
Avg. Doc Length	264	408	262

Table 3: FIRE 2008 Document Collection Details

Dice Coefficient (DC) is defined as follows:

$$l(t, t') = DC(t, t') = \frac{2 * freq(t, t')}{freq(t) + freq(t')} \quad (6)$$

6. ONLY QUERY TRANSLITERATION

Named Entities (NEs) are a crucial part of any query. In the sample query shown in Table 1, if the words सिंगूर (“Singur”) and नंदीग्राम (“Nandigram”) are properly transliterated then a large part of the results will be relevant even if the other words are not translated since the places are uniquely identified with the dispute. Moreover, many a times, in Indian language documents the actual Indian word is used as-is without translation. For example, to describe *Kashmir Travel*, it is quite common to also use *Kashmir Yatra* in the documents i.e, the actual Hindi word transliterated in English. So, this motivated us to try the following experiments:

- Transliterate the whole query without translation
- Transliterate only NEs and ignore rest

In the first case, the translated query for a Hindi query meaning *Kashmir Travel* would be *Kashmir Yatra* and in the second case, it would be only *Kashmir*. We discuss the results of our experiments in the next section.

7. EXPERIMENTS AND RESULTS

The details of the FIRE 2008 document collection are given in Table 3. We used *Trec Terrier* [7] as the monolingual English IR engine. We used the standard implementation of Okapi BM25 in Trec Terrier for our runs. The documents were indexed after stemming and stop-word removal. The topic set consisted of 50 topics each in Hindi, Marathi and English. We used the Hindi and Marathi stemmers and morphological analyzers developed at CFILT, IIT Bombay for stemming the topic words. For English, we use the standard Porter steamer. Due to time constraints, we only submitted the Title runs for all the tasks. The details of the runs which we submitted are given in Table 4.

We use the following standard evaluation measures [9]: Mean Average Precision (MAP), R-Precision, Precision at 5, 10 and 20 documents (P@5, P@10 and P@20) and Recall. As is commonly reported in CLIR evaluation, we also report the percentage with respect to monolingual retrieval for each performance figure. The overall results are tabulated in Table 5. The corresponding precision-recall curves are shown in Figures 4, 5 and 3.

7.1 Discussion

In Monolingual runs, we observed a Mean Average Precision (MAP) of 0.4429, 0.1670 and 0.2963 for English, Hindi and Marathi respectively. The Marathi and Hindi scores were relatively low due to issues in stemming. In Bilingual runs, for Hindi to English, we observed a MAP of 0.3349

S.No.	Description	Run ID
1	English-English Monolingual	EN-MONO-TITLE
2	Hindi-Hindi Monolingual	HI-MONO-TITLE
3	Marathi-Marathi Monolingual	MR-MONO-TITLE
4	Hindi-English Bilingual Title with DC	IITB_HINDI_ENG_TITLE_DICE
5	Marathi-English Bilingual Title with DC	IITB_MAR_ENG_TITLE_DICE
6	English-Hindi Bilingual Title with DC	IITB_ENG_HINDI_TITLE_DICE
7	Hindi-English Bilingual Title: Entire Query Transliteration	IITB_HINDI_ENG_TITLE_TRANSLIT
8	Hindi-English Bilingual Title: Only NE Transliteration	IITB_HINDI_ENG_TITLE_NETRANSLIT
9	Marathi-English Bilingual Title: Entire Query Transliteration	IITB_MAR_ENG_TITLE_TRANSLIT
10	Marathi-English Bilingual Title: Only NE Transliteration	IITB_MAR_ENG_TITLE_NETRANSLIT

Table 4: Details of Monolingual and Bilingual Runs Submitted

Title Only						
Run Desc.	MAP	R-Precision	P@5	P@10	P@20	Recall
EN-MONO-TITLE	0.4429	0.4589	0.6320	0.6180	0.5680	89.75%
HI-MONO-TITLE	0.1670	0.1833	0.3280	0.3020	0.2800	59.22%
MR-MONO-TITLE	0.2963	0.2985	0.3640	0.3100	0.2640	85.29%
IITB_HINDI_ENG_TITLE_DICE	0.3349	0.3436	0.4760	0.4560	0.4290	78.11%
	(75.61%)	(74.87%)				(87.03%)
IITB_MAR_ENG_TITLE_DICE	0.2852	0.3017	0.4680	0.4460	0.3970	73.85%
	(64.39%)	(65.74%)				(82.28%)
IITB_ENG_HINDI_TITLE_DICE	0.1061	0.1286	0.2000	0.1820	0.1730	34.10%
	(63.53%)	(70.15%)				(57.58%)
Transliteration Experiment Runs						
IITB_HINDI_ENG_TITLE_TRANSLIT	0.1583	0.1653	0.2040	0.2220	0.2030	40.75%
	(35.74%)	(36.02%)				(45.40%)
IITB_HINDI_ENG_TITLE_NETRANSLIT	0.1572	0.1650	0.1960	0.2220	0.2010	39.90%
	(35.49%)	(35.95%)				(44.45%)
IITB_MAR_ENG_TITLE_TRANSLIT	0.1132	0.1204	0.1280	0.1480	0.1500	33.50%
	(25.55%)	(26.23%)				(37.32%)
IITB_MAR_ENG_TITLE_NETRANSLIT	0.1111	0.1204	0.1280	0.1340	0.1420	34.16%
	(25.08%)	(26.23%)				(38.06%)

Table 5: FIRE 2008 Overall Results (Percentage of monolingual performance given in brackets below the actual numbers)

which is 75.61% of monolingual performance. In Marathi to English, the MAP was 0.2852 which is 64.39% of monolingual performance. For English to Hindi, the MAP was 0.1061 and it was low due to stemming issues with Hindi mentioned earlier.

In the transliteration experiment, we notice that the results of transliterating the whole query vs. transliterating the named-entities are almost identical. The results for Hindi to English are better with a MAP of 0.1583 and 0.1572 which is 35.74% and 35.49% of monolingual performance respectively. The Marathi to English MAP is 0.1132 and 0.1111 which is 25.55% and 25.08% of monolingual performance respectively.

8. CONCLUSION

We discussed the evaluation of our Hindi to English, Marathi to English and English to Hindi CLIR systems performed as part of our participation in FIRE 2008 Ad-Hoc Bilingual and Monolingual tasks. Our approach is based on query translation using bi-lingual dictionaries. For Hindi and Marathi to English, transliteration of OOV words is done using a simple rule based approach and in case of English to Hindi, we use substring-based transliteration technique. Disambiguating the various translations/transliterations is performed us-

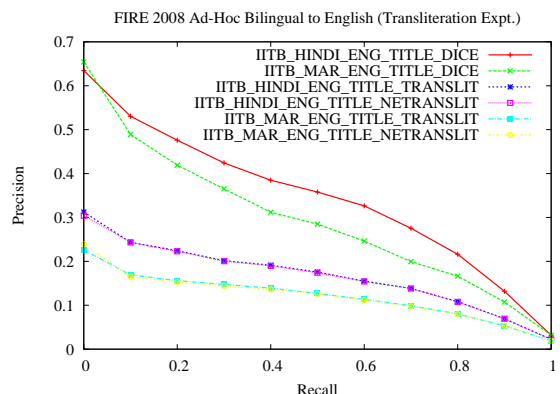


Figure 3: FIRE 2008 Transliteration Expt. Precision-Recall Curves

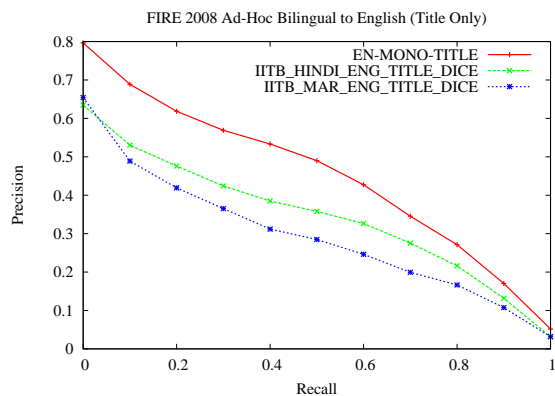


Figure 4: FIRE 2008 Ad-Hoc Monolingual and Bilingual P-R Curves: English

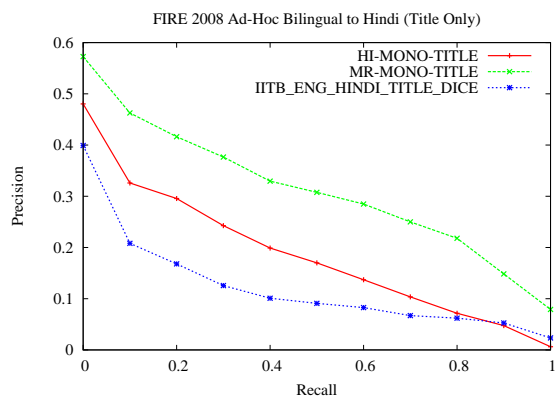


Figure 5: FIRE 2008 Ad-Hoc Monolingual and Bilingual P-R Curves: Hindi

ing an iterative page-rank style algorithm which is based on term-term co-occurrence statistics. We also experimented with the idea of transliterating the whole query without translation and observe that we can achieve reasonable accuracy in many cases. This could form the basis of further investigation to automatically mine translation dictionaries from query results while starting with only with a transliteration engine.

9. ACKNOWLEDGEMENTS

The second author is supported by a Fellowship Award from Infosys Technologies Limited, India.

10. REFERENCES

- [1] N. Bertoldi and M. Federico. Statistical models for monolingual and bilingual information retrieval. *Inf. Retr.*, 7(1-2):53–72, 2004.
- [2] M. K. Chinnakotla, S. Ranadive, O. P. Damani, and P. Bhattacharyya. Hindi to english and marathi to english cross language information retrieval evaluation.

Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, pages 111–118, 2008.

- [3] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903, New York, NY, USA, 2005. ACM Press.
- [4] F. Huang. Cluster-specific named entity transliteration. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 435–442, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
- [5] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments (parts 1& 2). *Information Processing and Management*, 36(6):779–840, 2000.
- [6] C. Monz and B. 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.
- [7] I. Ounis, G. Amati, P. V., B. He, C. Macdonald, and Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research (ECIR 2005)*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer, 2005.
- [8] S. Tarek and K. Grzegorz. Substring-based transliteration. In *Proceedings of ACL*, 2007.
- [9] R. B. Yates and B. R. Neto. *Modern Information Retrieval*. Pearson Education, 2005.