Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007

Manoj Kumar Chinnakotla, Sagar Ranadive, Pushpak Bhattacharyya and Om P. Damani
Department of CSE
IIT Bombay
Mumbai, India
{manoj,sagar,pb,damani}@cse.iitb.ac.in

Abstract

In this paper, we present our Hindi→English and Marathi→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 approach which utilizes the corpus to return the 'k' closest English transliterations of the given Hindi/Marathi 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 in title which is 61.36% of monolingual performance and a MAP of 0.2952 in title and description which is 67.06% of monolingual performance. For Marathi, we achieve a MAP of 0.2163 in title which is 56.09% of monolingual performance.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Hindi-to-English, Marathi-to-English, Cross Language Information Retrieval, Query Translation

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 [4]. According to a survey conducted by Online Computer Library Center (OCLC)[1], English is still the dominant language on the web. However, global internet usage statistics[2] 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.

[1] http://www.oclc.org/research/projects/archive/wcp/stats/intnl.htm

[2] http://www.internetworldstats.com/stats7.htm

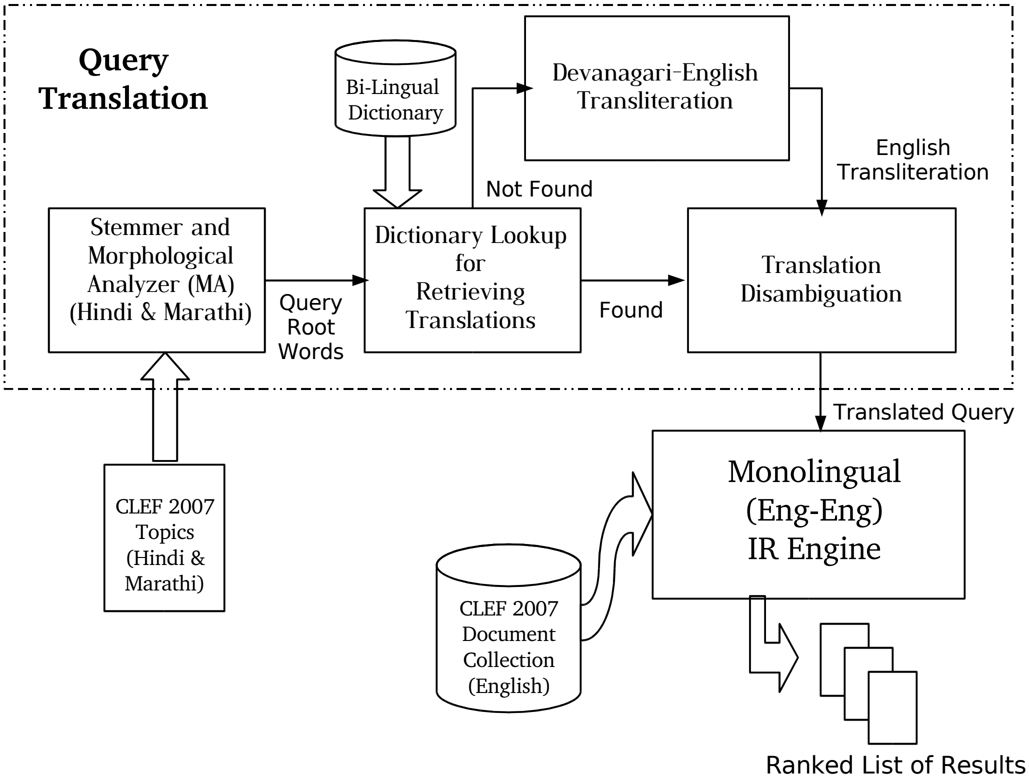
Query Translation
Bi-Lingual Dictionary
Devanagari-English Transliteration
English Transliteration
Not Found
Stemmer and Morphological Analyzer (MA) (Hindi & Marathi)
Query Root Words
Dictionary Lookup for Retrieving Translations
Found
Translation Disambiguation
Translated Query
Monolingual (Eng-Eng) IR Engine
CLEF 2007 Topics (Hindi & Marathi)
CLEF 2007 Document Collection (English)
Ranked List of Results

Figure 1: System Architecture of our CLIR System

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 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.

Hindi is the official language of India along with English and according to *Ethnologue*[3], 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→English and Marathi→English CLIR approaches for the CLEF 2007 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 and discusses the results. Finally, Section 7 concludes the paper highlighting some potential directions for future work.

[3] `http://www.ethnologue.com`

Algorithm 1 Query Translation Approach

```
Remove all the stop words from query
Stem the query words to find the root words
for $stem_i \in$ stems of query words do
    Retrieve all the possible translations from bilingual dictionary
    if list is empty then
        Transliterate the word using to produce candidate transliterations
    end if
end for
Disambiguate the various translation/transliteration candidates for each word
Submit the final translated English query to English→English Monolingual IR Engine
```

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→English and Marathi→English dictionaries created by Center for Indian Language Technologies (CFILT)[4], IIT Bombay for query translation. The Hindi→English bi-lingual dictionary has around 1,15,571 entries and is also available online[5]. The Marathi→English bi-lingual has relatively 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 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 Devanagari→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.

3 English→English Monolingual

We used the standard Okapi BM25 Model [6] for English→English monolingual retrieval. Given a keyword query $Q = \{q_1, q_2, \ldots, q_n\}$ and document D, the BM25 score of the document D is as follows:

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

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

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] http://www.cfilt.iitb.ac.in

[5] 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>

Table 1: CLEF 2007 Topic Number 445

4 Devanagari to English Transliteration

Many *proper nouns* of English like names of people, places and organizations, used as part of the Hindi or Marathi query, are not likely to be present in the Hindi→English and Marathi→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 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 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 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 गंगोत्री is transliterated as shown in Table 2.

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 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* [5] 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 "आस्ट्रेलियाई" 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.

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

Input Letter	Output String
ग	ga
ं	gan
ग	ganga
ओ	gango
त्री	gangotri

Table 2: Transliteration Example

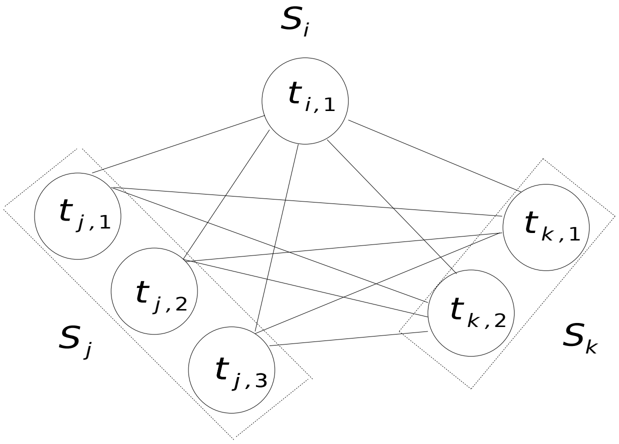

Figure 2: Co-occurrence Network for Disambiguating Translations/Transliterations [7]

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. 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.* [?] 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.

Initialization step:

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

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

Table 3: Mathematical symbols involved in translation disambiguation

Number of Documents	135153
Number of Terms	13362159
Number of Unique Terms	126034
Average Document Length	98

Table 4: Details of LA Times 2002 Collection

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:

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

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) [3] is defined as follows:

$$l(t,t') \;=\; PMI(t,t') \;=\; log_2 \frac{p(t,t')}{p(t) * p(t')} \tag{6}$$

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 product 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.

$$p(t,t') \;=\; \frac{freq(t,t')}{avgdl \times avgdl} \tag{7}$$

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

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:

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

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

S.No.	Description	Run ID
1	English-English Monolingual	EN-MONO-TITLE
2	Hindi-English Bilingual Title with DC	IITB_HINDI_TITLE_DICE
3	Hindi-English Bilingual Title with PMI	IITB_HINDI_TITLE_PMI
4	Marathi-English Bilingual Title with DC	IITB_MAR_TITLE_DICE
5	Marathi-English Bilingual Title with PMI	IITB_MAR_TITLE_PMI
6	English-English Monolingual Title+Desc	EN-MONO-TITLE+DESC
7	Hindi-English Bilingual Title+Desc with DC	IITB_HINDI_TITLEDESC_DICE
8	Hindi-English Bilingual Title+Desc with PMI	IITB_HINDI_TITLEDESC_PMI

Table 5: Details of Runs Submitted

6 Experiments and Results

The CLEF 2007 document collection for Ad-Hoc Bilingual Task consisted of a collection of articles from LA Times that appeared in the year 2002. The details of the target document collection is given in Table 4. We used *Trec Terrier* [8] 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 (using Porter Stemmer) and stop-word removal. The topic set consisted of 50 topics each in Hindi and Marathi. 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 Dice Coefficient and PMI for calculating the link weight. This gave rise to four runs for Hindi. For Marathi, due to resource constraints, we could not submit the Title + Description run. The details of the runs which we submitted are given in Table 5.

We use the following standard measures for evaluation [9]: Mean Average Precision (MAP), R-Precision, Precision at 5, 10 and 20 documents (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. The corresponding precision-recall curves are shown in Figure 3.

For Hindi, we achieve a Mean Average Precision (MAP) of 0.2366 in title which is 61.36% of monolingual performance and a MAP of 0.2952 in title and description which is 67.06% of monolingual performance. For Marathi, we achieve a MAP of 0.2163 in title which is 56.09% of monolingual performance. The recall levels in Hindi are 72.58% for title runs which is 89.16% of monolingual and 76.55% for title and description run which is 87.32% of monolingual. The recall levels in Marathi are 62.44% in title run which is 76.70% of monolingual.

6.1 Discussion

In the title runs, we observe better performance in Hindi than Marathi. One of the reasons for the above is that the Marathi Morphological Analyzer (MA) is still under development. Hence, many words were not properly stemmed due to which the correct translations/transliterations could not be retrieved. Dice Coefficient consistently performs better than PMI. This result needs to be further investigated.

7 Conclusion

We presented our Hindi→English and Marathi→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

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_TITLEDESC_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 6: CLEF 2007 Ad-Hoc Monolingual and Bilingual Overall Results (Percentage of monolingual performance given in brackets below the actual numbers)

performed using an iterative page-rank style algorithm which is based on term-term co-occurrence statistics.

The bi-lingual dictionaries available with us also have Parts-Of-Speech (POS) information for each word. POS tagging the input query may help in reducing the ambiguity since translations of only matching POS will be retrieved. As part of future work, we plan to investigate the above idea in more detail. Besides, we plan to explore alternate string matching measures which are based on phonetic similarity for retrieving 'k' best transliterations from corpus. Finally, we would like to study the effect of varying 'k' on disambiguation.

8 Acknowledgements

The first author is supported by a Fellowship Award from Infosys Technologies Limited, India. We would like to thank our project linguists Jaya Madam, Gajananji, Gauri, Sushant and Subodh for their help. We would also like to thank Manish Shrivastav for his help with Stemmer and Morphological Analyzer.

,

References

[**1**] Nicola Bertoldi and Marcello Federico. Statistical models for monolingual and bilingual information retrieval. *Inf. Retr.*, 7(1-2):53–72, 2004.

[**2**] Martin Braschler and Carol Peters. Cross-language evaluation forum: Objectives, results, achievements. *Inf. Retr.*, 7(1-2):7–31, 2004.

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

[**4**] 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.

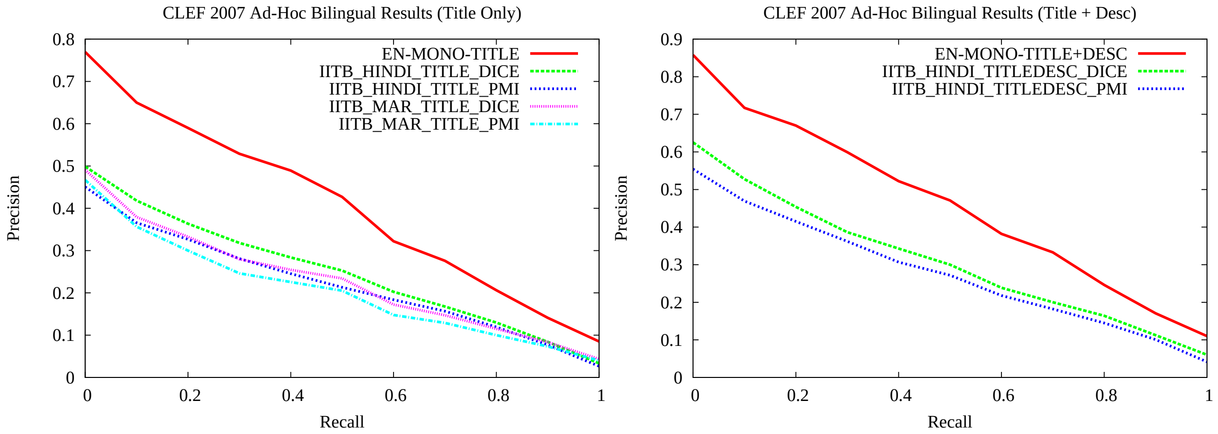

Figure 3: CLEF 2007 Ad-Hoc Monolingual and Bilingual Precision-Recall Curves

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

[**6**] 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.

[**7**] 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.

[**8**] I. Ounis, G. Amati, Plachouras 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.

[**9**] Ricardo Baeza Yates and Berthier Ribeiro Neto. *Modern Information Retrieval.* Pearson Education, 2005.