IITB CFILT @ FIRE 2010: Discriminative Approach to IR

Manoj Chinnakotla, Vishal Vacchani, Shalini Gupta, Karthik Raman, Pushpak Bhattacharyya
Dept. of Computer Science and Engineering (CSE)
IIT Bombay
Mumbai, India
{manoj, vishalv, shalini, karthikr, pb}@cse.iitb.ac.in

ABSTRACT

In this paper, we describe our participation in the Adhoc Information Retrieval, Cross Lingual Information Retrieval (CLIR) tasks of FIRE 2010. We use a discriminative approach to IR. Besides the basic term based matching features proposed in literature, we include novel features which model certain crucial elements of relevance like named entities, document length and semantic distance between query and document. We also investigate the effect of including Pseudo-Relevance Feedback (PRF) based features in the above framework.

In CLIR, we use a query translation approach using bilingual dictionaries. We use an iterative query disambiguation algorithm for query disambiguation. In the disambiguation approach, in lieu of regular term-term co-occurrence measures proposed in literature, we use the similarity between terms in LSI space.

1. INTRODUCTION

In this paper, we describe our participation in the Adhoc IR and Cross Lingual Information Retrieval (CLIR) tasks of FIRE 2010. We participated in the Adhoc monolingual task for Hindi, Marathi and English languages and in CLIR task for Hindi-English and Marathi-English language pairs. We use a discriminative approach to IR. Besides the basic term based matching features proposed in literature, we include novel features which model certain crucial elements of relevance like named entities, document length and semantic distance between query and document. We also investigate the effect of including Pseudo-Relevance Feedback (PRF) based features in the above framework.

In CLIR, we use a query translation approach using bilingual dictionaries. We use an iterative query disambiguation algorithm for query disambiguation. In the disambiguation approach, in lieu of regular term-term co-occurrence measures proposed in literature, we use the similarity between terms in LSI space.

The paper is organized as follows: Section 2 gives a brief introduction to Discriminative Models for Information Retrieval. Section 3 describes in detail, the various features that we have used. Section 4 describes PRF. Section 5 describes the algorithm used for Cross Lingual Retrieval. We discuss our experiments and results obtained in Section 6 and Section 7.

1.1 Discriminative Models For Information Retrieval

Various approaches have been used for IR. These include, Vector Space Models, Generative approaches and Discriminative approaches.

Any combination of arbitrary features can be included in the discriminative framework. This allows exploring use of different features representing the importance of document.

Relative weights of these features can be learnt using discriminative models like Support Vector Machines (SVMs). [6] explores the applicability of discriminative classifiers for IR, Their experiments show that SVMs (discriminative models) are on par with the language models (generative models). [3] introduces the use of RankSVMs for Information Retrieval.

In our experiments, we explore the use of RankSVMs (discriminative models) for monolingual as well as cross lingual retrieval. We use the statistical (tf, idf) features, as described in [6] as the basic set of features. Along with this we have explored the use of features based on Co-ord factor, Document Length Normalization, NER features, Phrasal NE and LSI features.

2. RANK SVM FEATURES

2.1 Statistical Features

- Term Frequency
  This feature measures the total number of term matches between a given query, q and a given document, D.
- Normalized Term Frequency
  This feature measures total number of matches between the query and the document and normalizes it with respect to the document length.
- Inverse Document Frequency
  This measures the importance of a document in terms of IDF (importance) of each query term, which overlaps with the document content.
- Normalized Cumulative Frequency
  This measures the overall frequency of the overlapping terms (between a query and a document) in the whole corpus
- Normalized Term Frequency, weighted by IDF.
  This is similar to the normalized term frequency, except that each term is weighted by its importance, measured as IDF.
- Normalized Term Frequency, weighted by cumulative frequency.

Sr. No.	Feature Name	Description
1	Term Frequency	$\sum_{q_i \in Q \cap D} log(c(q_i, D))$
2	Normalized Term Frequency	$\sum_i log(1 + \frac{c(q_i, D)}{\|D\|})$
3	Inverse Document Frequency	$\sum_{q_i \in Q \cap D} log(idf(q_i))$
4	Normalized Cumulative Frequency	$\sum_{q_i \in Q \cap D} log(\frac{\|C\|}{c(q_i, C)})$
5	Normalized Term Frequency, weighted by IDF	$\sum_i log(1 + \frac{c(q_i, D)}{\|D\|} idf(q_i))$
6	Normalized Term Frequency, weighted by cumulative frequency	$\sum_i log(1 + \frac{c(q_i, D)}{\|D\|} \frac{\|C\|}{c(q_i, C)})$

Table 1: Statistical Features

This is similar to the above feature, replacing the weighting factor by the cumulative frequency, as described in Normalized Cumulative Frequency Feature.

Table 1 expresses each of these features precisely.

2.2 Co-ord Factor

For a given query, q and document, D, Co-ord is a score factor based on how many of the query terms are found in the specified document. Typically, a document that contains more of the query's terms will receive a higher score than another document with fewer query terms. This measure is one of the factors in Lucene Ranking.

2.3 Document Length Normalization

Retrieval algorithms are usually biased towards shorter documents [9]. To remove this bias, we have added this feature (pivoted normalization, as described in [9]) as the smoothing factor.

2.4 Named Entity Features

We observed that most of the important words in the queries were named entities. We felt that, in a given document, the presence of query words which are named entities, is a strong sign of an important document. We capture the intuition through this feature. This feature calculates sum of term frequency of named entity query words in document. Sometimes complete multi-word Named Entity is not present in the document at consecutive locations. We, therefore, use phrasal NEs. We find out if all the NE words in a query are present within a context window in the document.

E.g. If the query is "Laloo Prasad Yadav", all words may not be found consecutively in the document. However, the presence of all 3 words in a context window of size 50, is a strong indicator of an important document.

We use the score provided by Lucene's Phrase Query Scoring.

2.5 LSI Feature

This feature captures the LSI based similarity between query and document. This was implemented using the "Semantic Vectors Package" [10]. Instead of word-based match between query and document, as captured by the "Term Frequency" feature, this feature measures the concept-based match between query and document.

3. PSEUDO RELEVANCE FEEDBACK

Pseudo Relevance Feedback (PRF) is a method of automatic local analysis where retrieval performance is expected to improve through query expansion by adding terms from top ranking documents. An initial retrieval is conducted returning a set of documents. The top *n* retrieved documents from this set are then assumed to be the most relevant documents, and the query is reformulated by expanding it using words that are found to be of importance in these documents. PRF has shown improved IR performance, but there is a risk of query drift in applying PRF. We use language modeling based Model Based Feedback (MBF) approach [11] for performing PRF.

<topics>
<top lang="hi">
<num> 76 </num>
<title> *gurjaron aur meena samuday ke beech sangharsh*</title>

Table 2: FIRE 2010 Topic Number 76

4. CROSS LINGUAL IR

Our CLIR system is based on a dictionary based approach to query translation. After removing the stops words from the query, the query words are passed to the stemmer. We use Hindi and Marathi stemmers developed at CFILT, IIT Bombay. The stemmed query words are translated using Hindi and Marathi dictionary. We use Hindi dictionary available at the site "www.cfilt.iitb.ac.in" for Hindi-English translation containing 131750 entries. Marathi-English bilingual dictionary contains 31845 entries.

4.1 Query Transliteration

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 2 presents a sample Hindi topic from FIRE 2010. In the above topic, the word is "Gurjars" 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 Segment Based Transliteration approach for Devanagari to English transliteration. The current accuracy of the system is 71.8% at a rank of 5 [1].

4.2 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. The main principle for disambiguation is that the correct translation of query terms should co-occur in target language documents and incorrect translation should tend not to co-occur. Given the possible target equivalents of two source terms, we can infer the most likely translations by looking at the pattern of co-occurrence for

Language	No. of Documents	No. of Tokens	No. of unique terms	Avg. Doc. Length
English	125586	33259913	191769	264.84
Hindi	95216	38541048	146984	404.77
Marathi	99275	25568312	609155	257.55

Table 3: Details of FIRE 2010 Corpus

S.No.	Description	Run ID
1	EN Monolingual Title without Feedback	IITBCFILT_EN_MONO_TITLE_BASIC
2	EN Monolingual Title with Feedback	IITBCFILT_EN_MONO_TITLE_FEEDBACK
3	EN Monolingual Title+Desc without Feedback	IITBCFILT_EN_MONO_TITLE+DESC_BASIC
4	EN Monolingual Title+Desc with Feedback	IITBCFILT_EN_MONO_TITLE+DESC_FEEDBACK
5	HI Monolingual Title without Feedback	IITBCFILT_HI_MONO_TITLE_BASIC
6	HI Monolingual Title+Desc without Feedback	IITBCFILT_HI_MONO_TITLE+DESC_BASIC
7	HI Monolingual Title+Desc with Feedback	IITBCFILT_HI_MONO_TITLE+DESC_FEEDBACK
8	MR Monolingual Title without Feedback	IITBCFILT_MR_MONO_TITLE_BASIC
9	MR Monolingual Title with Feedback	IITBCFILT_MR_MONO_TITLE_FEEDBACK
10	MR Monolingual Title+Desc without Feedback	IITBCFILT_MR_MONO_TITLE+DESC_BASIC
11	MR Monolingual Title+Desc with Feedback	IITBCFILT_MR_MONO_TITLE+DESC_FEEDBACK
12	HI-EN Bilingual Title without Feedback	IITBCFILT_HI_EN_TITLE_BASIC
13	HI-EN Bilingual Title with Feedback	IITBCFILT_HI_EN_TITLE_FEEDBACK
14	MR-EN Bilingual Title without Feedback	IITBCFILT_MR_EN_TITLE_BASIC
15	MR-EN Bilingual Title with Feedback	IITBCFILT_MR_EN_TITLE_FEEDBACK

Table 4: Details of Monolingual and Bilingual Runs Submitted

each possible pairs of definitions. For example, for a query "*nadi jal*", the translation for "*nadi*" is {river} and the translations for "*jal*" are {water, to burn}. Here, based on the context, we can see that "water" is a better translation for the second word "*jal*", 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. We use a page-rank style iterative disambiguation algorithm proposed by Christof Monz et. al. [5] which examines pairs of terms to gather partial evidence for the likelihood of a translation in a given context.

The link weight, which is meant to capture the association strength between the two words (nodes), could be measured using various functions. In FIRE-2008 [8], we used Dice-Coefficient as a link-weight. In FIRE-2010, we use Latent Semantic Indexing (LSI) [2] based similarity measure between two terms.

Latent Semantic Indexing (LSI) is an indexing method that uses a mathematical technique called Singular Value Decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. We use this feature as a link weight in Iterative disambiguation algorithm.

5. EXPERIMENTS AND RESULTS

The details of the FIRE 2010 document collection are given in Table 3. We used TREC Terrier [7] as the monolingual English IR engine. As explained in sections 1 and 2, we use Discriminative IR frame work for all our runs. The documents were indexed after stemming and stop-word removal. We use FIRE-2008 topic set consisting of 50 topics each, in Hindi, Marathi and English along with their relevance judgment as training and development data set for our experiments.

We use 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 stemmer. In Discriminative IR framework, we experiment with SVM and Rank-SVM. We observed that Rank-SVM always performs better than SVM. Hence, we use Rank-SVM in our experiments. We submitted the Title(T), and Title + Description (T+D) runs for all the tasks. The details of the runs which we submitted are given in Table 4. Results of monolingual runs are shown in the Table 5 and cross-lingual runs are given in the Table 6.

We use the following standard evaluation measures [4]: Mean Average Precision (MAP), Precision at 5, 10 and 20 documents (P@5, P@10 and P@20) and Recall. The results shown are the results of the submitted runs. We observed that there were some errors in the runs. Those runs have been marked with an asterisk (*) in Table 5

6. DISCUSSION

We introduced four new features, NE feature, Co-ord factor, Normalization and LSI based similarity. We describe the results obtained after adding these features to the Basic Feature Set (Statistical Features) for different languages. The table 7 shows the MAP score with different features for different languages. These scores represent the MAP values obtained by using only "Title" as the query.

In English, we observe that the LSI feature helps in improving the performance. However, the NE and Document Length Normalization features are not useful. In Hindi, we observe that Co-ord factor and document length normaliza-

Run ID	MAP	P@5	P@10	P@20	Recall
IITBCFILT_EN_MONO_TITLE_BASIC	0.3803	0.4280	0.3540	0.2850	0.9832
IITBCFILT_EN_MONO_TITLE_FEEDBACK*	0.3779	0.4080	0.3740	0.2860	0.9908
IITBCFILT_EN_MONO_TITLE+DESC_BASIC*	0.3803	0.4280	0.3540	0.2850	0.9832
IITBCFILT_EN_MONO_TITLE+DESC_FEEDBACK*	0.3779	0.4080	0.3740	0.2860	0.9908
IITBCFILT_HI_MONO_TITLE_BASIC	0.2384	0.3360	0.2760	0.2200	0.8175
IITBCFILT_HI_MONO_TITLE+DESC_BASIC	0.3317	0.4440	0.3680	0.2940	0.8962
IITBCFILT_HI_MONO_TITLE+DESC_FEEDBACK*	0.0001	0.0000	0.0000	0.0000	0.0164
IITBCFILT_MR_MONO_TITLE_BASIC	0.2135	0.2560	0.2380	0.1850	0.8019
IITBCFILT_MR_MONO_TITLE_FEEDBACK*	0.0021	0.0040	0.0100	0.0060	0.0660
IITBCFILT_MR_MONO_TITLE+DESC_BASIC	0.2399	0.2800	0.2720	0.2280	0.9275
IITBCFILT_MR_MONO_TITLE+DESC_FEEDBACK*	0.0064	0.0200	0.0220	0.0150	0.1046

Table 5: FIRE 2010 Monolingual Results

Run ID	MAP	P@5	P@10	P@20	Recall
IITBCFILT_HI_EN_TITLE_BASIC	0.2848	0.2960	0.2580	0.2130	0.9081
IITBCFILT_HI_EN_TITLE_FEEDBACK	0.3365	0.3560	0.3160	0.2510	0.9142
IITBCFILT_MR_EN_TITLE_BASIC	0.2515	0.2776	0.2286	0.1878	0.7850
IITBCFILT_MR_EN_TITLE_FEEDBACK	0.2771	0.2939	0.2510	0.2031	0.8411

Table 6: FIRE 2010 Cross Lingual Results

Language	Basic	Basic + NE	Basic + LSI	Basic + Norm	Basic + Co-ord
English	0.3715	0.3759	0.3818	0.3689	0.3707
Hindi	0.2247	0.1900	NA	0.2284	0.2338
Marathi	0.2141	0.2142	NA	0.2128	0.2148

Table 7: FIRE 2010 Cross Lingual Results

tion improve the results. Finally, in Marathi the extra features do not improve performance over basic term features. Initial analysis into the failure of these features reveal that many of these problems are due to inaccuracies of the stemmer. However, a thorough error-analysis needs to be done to confirm the above hypothesis.

7. CONCLUSION

We presented the evaluation of our Hindi-English, Marathi-English CLIR systems performed as part of our participation in FIRE 2010 Ad-Hoc Bilingual task. In CLIR, we used a LSI based term-term similarity metric for query disambiguation. We also discussed the monolingual evaluation of our system for English, Hindi and Marathi languages using a discriminative approach to IR. Besides the regular term based features, we experiment with novel features like NEs, LSI based similarity features, document length normalization, co-ord factor, pseudo-relevance feedback and studied their impact. Results show improvement in some languages but they are not consistent and a detailed error analysis is required to understand the behaviour of each of these features. We plan to take up this work in future.

8. ACKNOWLEDGEMENTS

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

9. REFERENCES

[1] M. K. Chinnakotla and O. P. Damani. Character sequence modeling for transliteration. In *ICON 2009*, IITB, India, 2009.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[3] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA, 2002. ACM.

[4] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, UK, 2008.

[5] 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.

[6] R. Nallapati. Discriminative models for information retrieval, 2004.

[7] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[8] N. Padariya, M. Chinnakotla, A. Nagesh, and O. P. Damani. Evaluation of hindi to english, marathi to english and english to hindi clir at fire 2008. In *FIRE 2008*, IITB, India, 2008.

[9] A. Singhal, C. Buckley, M. Mitra, and A. Mitra. Pivoted document length normalization. pages 21–29.

ACM Press, 1996.

[10] D. Widdows and K. Ferraro. Semantic vectors: a scalable open source package and online technology management application. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[11] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410, New York, NY, USA, 2001. ACM.