Published in Multilingual Information Access in South Asian Languages, 334-352, 2013
which should be used for any reference to this work

Information Retrieval with Hindi, Bengali, and Marathi Languages: Evaluation and Analysis

Jacques Savoy, Ljiljana Dolamic, and Mitra Akasereh

Computer Science Department,
University of Neuchatel,
Rue Emile Argand 11, 2000 Neuchatel, Switzerland
{Jacques.Savoy,Ljiljana.Dolamic,Mitra.Akasereh}@unine.ch

Abstract. Our first objective in participating in FIRE evaluation campaigns is to analyze the retrieval effectiveness of various indexing and search strategies when dealing with corpora written in Hindi, Bengali and Marathi languages. As a second goal, we have developed new and more aggressive stemming strategies for both Marathi and Hindi languages during this second campaign. We have compared their retrieval effectiveness with both light stemming strategy and *n*-gram language-independent approach. As another language-independent indexing strategy, we have evaluated the trunc-*n* method in which the indexing term is formed by considering only the first *n* letters of each word. To evaluate these solutions we have used various IR models including models derived from Divergence from Randomness (DFR), Language Model (LM) as well as Okapi, or the classical *tf idf* vector-processing approach.

For the three studied languages, our experiments tend to show that IR models derived from Divergence from Randomness (DFR) paradigm tend to produce the best overall results. For these languages, our various experiments demonstrate also that either an aggressive stemming procedure or the trunc-*n* indexing approach produces better retrieval effectiveness when compared to other word-based or *n*-gram language-independent approaches. Applying the Z-score as data fusion operator after a blind-query expansion tends also to improve the MAP of the merged run over the best single IR system.

Keywords: Hindi, Bengali and Marathi information retrieval, retrieval effectiveness with Indian Languages, FIRE evaluation campaign, automatic indexing.

1 Introduction

During the last ten years, the IR group at University of Neuchatel was involved in designing, implementing and evaluating various indexing and search strategies for various natural languages, including popular European [1], Far-East (e.g., Chinese, Japanese, and Korean) [2], as well as Indian languages [3]. This objective also includes bilingual IR (the topics are then written in one language, the retrieved documents in another) or multilingual IR systems (targeted information items are written

in different languages). In our participation in the second FIRE campaign (www.isical.ac.in/~fire/), our main motivation is to promote new tools and to improve existing ones for monolingual IR when facing with Hindi, Marathi and Bengali languages.

The rest of this paper is organized as follows: Section 2 presents an overview of the corpora used in the FIRE-2010 *ad hoc* track. Section 3 outlines the main aspects of various IR models used with these test-collections together with the stopword lists and stemming strategies we developed for these languages. Section 4 presents the evaluation carried out on the various query formulations, stemming and indexing strategies using various IR models for the Hindi, Bengali and Marathi corpora. Finally, Section 5 describes our official runs and their evaluation while Section 6 gives a general conclusion.

2 Overview of the Corpora

The corpora used in our experiments are based on newspaper articles extracted from four main sources, namely *Anandabazar Patrika* (2004-2007) for the Bengali language, *Maharashtratimes* & *Esakal* (2004-2007) for the Marathi corpus, and *Dainik Jagran* & *Amar Ujala* (2004-2007) for the Hindi collection. The latest newspaper forms a new source of information in the second FIRE 2010 *ad hoc* campaign. The encoding system used for both documents and topic formulations is UTF-8.

In order to obtain an overall picture of the three corpora, we have reposted some statistics in Table 1. In this table, we can see that the Hindi collection is the largest with around 1.3 GB of data. The number of documents is however similar between this language and the Bengali corpus. When inspecting the document length, the Hindi, Bengali and Marathi corpora show similar mean lengths. These values range from 300.7 for the Hindi corpus to 264.6 for the Marathi collection. It is interesting to note that even though the Marathi collection is the smallest (487 MB), it contains a larger number of distinct indexing terms (511,550) when compared to both the Hindi and Bengali corpora. This fact is certainly related to a more complex inflectional morphology for this language.

Based on the TREC model [4], each topic formulation consists of three logical sections, namely a brief title (denoted T), a one-sentence description (D), and a narrative part (N) used mainly to specify more precisely the relevance judgment. Available topics reflect a diversity of information needs having mostly a national coverage (e.g., Topic #86 "Privatization of the Mumbai and Delhi airports", Topic #106 "Ban on Taslima Nasreen's novel "Shame"", or Topic #119 "Taj Mahal controversy"). The real user information need behind the topic description is sometimes difficult to determine, at least based on the title of the topic formulation (e.g., Topic #89: "Involvement of Congress ministers in the oil-for-food scam").

In the bottom part of Table 1, we have indicated the number of relevant documents (label "#Rel. doc.") per topic, with the mean always being greater than the median (e.g., for the Marathi collection, the average number of relevant documents per query is 15.9, with the corresponding median being 10). These findings indicate that each

collection contains numerous queries, yet only a rather smaller number of relevant items are found. For each collection, 50 queries were created (numbered from #76 to #125), and then manually translated into the other languages, including also an English version. Relevant documents could not however be found for each request and each language. For the Marathi language, eleven topics (#95, #98, #103, #107, #108, #113, #117, #119, #120, #121, and #125) do not have any relevant item in the collection.

Table 1. FIRE Test-Collection Statistics

	Hindi	Bengali	Marathi
Size	1,300 MB	732 MB	487 MB
# documents	149, 481	123,047	99,357
# terms	230,578	249,215	511,550
Number of indexing terms per document			
Mean	300.7	291.88	264.6
Std dev.	337.13	180.62	188.96
Median	220	265	222
Max	6,998	2,928	5,077
Min	0	0	28
Topics			
Number	50	50	39
# Rel. doc.	913	510	621
Mean	18.26	10.2	15.9
Std dev.	15.3	6.6	18.5
Median	14	8	10
Max	74 (T #93)	29 (T #89)	72 (T #88)
Min	2 (T #78, #87)	2 (T #84)	1 (T #79, #102, #122, #124)

The largest number of relevant items is 74 for Topic #93 ("Relations between Congress and its allies") in the Hindi collection. On the other hand, and for the Marathi corpus only, Topic #79 ("Building roads between China and Mount Everest"), Topic #102 ("Pakistani cricketers involved in a doping scandal"), Topic #122 ("Sanjay Dutt's surrender") and Topic #124 ("Sale of illegal drugs in various Indian states") have only one relevant document.

3 IR Models and Stemming Strategies

3.1 IR Models

Instead of being limited to a single indexing and search strategy, our aim is to obtain a relatively large overview of the relative merits of different IR models. To achieve this, we have considered adopting different weighting schemes for the terms included in document or query representatives. These different IR schemes take account for term occurrence frequencies (denoted tf_{ij} for indexing term t_j in document D_i), as well as their inverse document frequency ($idf_j = \log(n/df_j)$ with n indicating the number of

documents in the corpus, and df_j the number of documents in which the term t_j occurs). To define the first IR model, we have normalized each indexing weight using the cosine in order to obtain the classical *tf idf* formulation.

In addition to this classical vector-space approach, we also considered probabilistic models such as Okapi (or BM25) [5] that also takes document length into account. As a second probabilistic approach we have implemented four variants of the DFR (*Divergence from Randomness*) paradigm proposed by Amati & van Rijsbergen [6]. In this framework, the indexing weight w_{ij} attached to term t_j in document D_i combines two information measures as follows:

$$w_{ij} = \mathrm{Inf}^1_{ij} \cdot \mathrm{Inf}^2_{ij} = -\log_2\left(\mathrm{Prob}^1_{ij}(tf)\right) \cdot \left(1 - \mathrm{Prob}^2_{ij}\right) \tag{1}$$

As a first model, we have implemented the PB2 scheme defined as follows:

$$\mathrm{Prob}^1_{ij} = \left[\frac{e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}}}{tf_{ij}!}\right] \quad \text{with} \quad \lambda_j = \frac{tc_j}{n} \tag{2}$$

$$\mathrm{Prob}^2_{ij} = 1 - \left[\frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)}\right] \quad \text{with} \quad tfn_{ij} = tf_{ij} \cdot \log_2\left[1 + \frac{c \cdot mean\ \ dl}{l_i}\right] \tag{3}$$

where tc_j indicates the number of occurrences of term t_j in the collection, l_i the length (number of indexing terms) of document D_i, *mean dl* the average document length, n the number of documents in the corpus, and c is a constant. Table 2 depicts the exact values of these parameters used in our experiments.

Table 2. Parameter Settings for the Various Test-Collections

	Okapi			DFR	
Language	*b*	K_1	*avdl*	*c*	*mean dl*
Hindi	0.55	1.2	300	1.5	300
Bengali	0.55	1.2	292	1.5	292
Marathi	0.75	1.2	265	1.5	265

For the GL2 model, the implementation of Prob^1_{ij} is shown in Equation 4, and Prob^2_{ij} in Formula 5.

$$\mathrm{Prob}^1_{ij} = 1 - \left[\frac{1}{1+\lambda_j}\right] \cdot \left[\frac{\lambda_j}{1+\lambda_j}\right]^{tfn_{ij}} \tag{4}$$

$$\mathrm{Prob}^2_{ij} = \frac{tfn_{ij}}{tfn_{ij} + 1} \tag{5}$$

For the PL2 model, the implementation was carried out using Formula 2 for $Prob^1_{ij}$ and Equation 5 for $Prob^2_{ij}$. Finally for the fourth model denoted I(n_e)C2, the implementation is based on the following two equations:

$$Inf^1_{ij} = tfn_{ij} \cdot \log\left[\frac{n+1}{n_e+0.5}\right] \quad \text{with} \quad n_e = n \cdot \left[1-\left(\frac{n-1}{n}\right)^{tc_j}\right] \tag{6}$$

$$\text{Prob}^2_{ij} = 1-\frac{tc_j+1}{df_j \cdot (tfn_{ij}+1)} \tag{7}$$

Finally we also considered an approach based on a statistical language model (LM) [7], [8], known as a non-parametric probabilistic model (the Okapi and DFR are viewed as parametric models). Probability estimates would thus not be based on any known distribution (e.g., as in Equation 2 or 4), but rather estimated directly based on the term occurrence frequencies in document D_i or the whole corpus C. Within this language model paradigm, various implementations and smoothing methods might be considered, although in this study we adopted a model proposed by Hiemstra [8], as described in Equation 8, combining an estimate based on the document (Prob[t_j|D_i]) and on the corpus (Prob[t_j|C]) combining using the Jelinek-Mercer smoothing [9] scheme.

$$\text{Prob}[D_i \mid Q] = \text{Prob}[D_i] \cdot \prod_{t_j \in Q} \left[\lambda_j \cdot \text{Prob}[t_j \mid D_i] + (1-\lambda_j) \cdot \text{Prob}[t_j \mid C]\right] \tag{8}$$

$$\text{Prob}[t_j \mid D_i] = tf_{ij}/l_i \quad \text{and} \quad \text{Prob}[t_j \mid C] = df_j/l_c \quad \text{with} \quad lc = \sum_k df_k$$

where λ_j is a smoothing factor (constant for all indexing terms t_j, and fixed at 0.35) and lc an estimate of the size of the corpus C.

3.2 Stopword Lists and Stemmers

During this evaluation campaign, our stopword lists for the Hindi and Bengali languages were the same as those used during our FIRE 2008 participation [3]. These stopword lists contain 165 Hindi terms and 114 Bengali words. During the second FIRE evaluation campaign, we have proposed a new stopword list for the Marathi language. This list, created in the same way as the stopword lists for other two languages [10], contains 99 terms. We may mention that compared to other Indo-European languages, these lists are rather short (e.g., the SMART system used a list of 471 words for the English language). We may certainly include additional words to speed up the query processing and to enhance the quality of the final ranked list of retrieved items. Recent studies tend to show that the elaboration of such stopword list may have a clear impact on the retrieval effectiveness of a search engine [11].

The light stemming procedures we have employed this year are the same as those used for the Indian languages during the first FIRE evaluation campaign. These stemming procedures remove inflectional suffixes attached to both nouns and

adjectives, while completely ignoring the verbal morphology of the underlying language. This reflects our belief and prior experiments with other languages [12] that nouns and adjectives are the Part-Of-Speech (POS) categories covering the most important part of the semantic content of both documents and queries. Moreover, including numerous verbal suffixes in a suffix-striping approach might hurt retrieval effectiveness, especially when we know that such stemmer does not consider the underlying POS or does not involve a complex morphological analysis. Finally, for the English language at least, we do not find that a deeper morphological analysis proposes a better retrieval effectiveness than simple stemmers [13], [14] approaches [15].

Finally, our participation in the FIRE evaluation campaigns was also motivated by our wish to promote and evaluate new and more aggressive stemming procedures for the Hindi, Marathi and Bengali languages. These procedures, apart from removing inflectional suffixes from nouns and adjectives, remove also some frequently used derivational suffixes found in the grammar of the corresponding language. In the web site `members.unine.ch/jacques.savoy/clef/` we can find the proposed stopword lists and various stemmers.

4 Evaluation and Analysis

To evaluate various indexing and search strategies, Section 4.1 exposes the evaluation measures and methodology used in our experiments. Section 4.2 presents the performance achieved by seven retrieval models based on three topic formulations (T, TD and TDN). Section 4.3 describes the performance that can be achieved by using three different stemming strategies (none, light or aggressive) while Section 4.4 reports the evaluation achieved by using three document and query representations (word-based, *n*-gram, and trunc-*n*). Section 4.5 analyzes some queries in an attempt to understand the impact of various stemming and indexing strategies as well as the effect of adding search terms to the current query. The last section evaluates the impact of two automatic blind-query expansion techniques to hopefully improve the retrieval effectiveness.

4.1 Evaluation Methodology

As a measure of retrieval effectiveness, we have adopted mean average precision (MAP) (computed by the TREC_EVAL software based on a maximum of 1,000 retrieved records). This performance measure has been used by all evaluation campaigns for around 20 years in order to objectively compare various IR models, particularly regarding their ability to retrieve relevant items (*ad hoc* tasks) [16]. Using this evaluation tool, some differences may occur in the values computed according to the official measure. The latter always takes 50 queries while in our presentation we did not account for queries having no relevant item, as for the Marathi collection owning only 39 queries. In the following tables, best performances under given conditions (same indexing scheme and same collection) are listed in bold type.

Using the mean as a measure of the system's performance signifies that we attached an equal importance to all queries. Comparisons between two IR strategies

will therefore not be based on a single query with respect to those available in the underlying test-collection or specifically created in order to demonstrate that a given IR approach must be rejected. Thus we believe that it is important to conduct experiments involving the largest possible number of observations (between 39 and 50 queries in our evaluations, depending on the language).

To statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology [17] showing very similar conclusion than the *t*-test but without requiring parametric assumption [18]. In our statistical tests, the null hypothesis H_0 stated that both retrieval schemes produce similar MAP performance. Such a null hypothesis would be accepted if two retrieval schemes returned statistically similar MAP, otherwise it must be rejected. Thus, in the experiments presented in this paper, statistically significant differences were detected by a two-sided test (significance level $\alpha = 5\%$).

4.2 Evaluation of Different Query Formulations

Based on the Hindi corpus, Table 3 shows the MAP obtained by seven IR models with three different query formulations (T, TD, and TDN) using a word-based and a light stemming approach. Tables 4 and 5 depict the same information for the Bengali and Marathi languages respectively. Across the three different corpora and three query formulations, we can see that the best IR model is usually I(n_e)C2, an implementation of the DFR family. We must recognize that the performance differences are not important when comparing this model with the DFR-PB2 scheme.

Table 3. MAP of Various IR Models (Light Stemming, Hindi Corpus, 50 queries)

	Mean Average Precision		
Light Stemming Model	Hindi T	Hindi TD	Hindi TDN
Okapi	0.3011	0.3717	0.4533 †
DFR-PB2	0.2851	0.3737	0.4630
DFR-GL2	0.2730 †	0.3524 †	0.4337 †
DFR-PL2	0.2843 †	0.3621	0.4430 †
DFR-I(n_e)C2	**0.3054**	**0.3836**	**0.4732**
LM	0.2533 †	0.3310 †	0.4223 †
tf idf	0.1427 †	0.1830 †	0.2367 †
Average	0.2635	0.3368	0.4179
Change % over T	base	+27.79%	+58.56%

To verify whether these performance differences are statistically significant, we compare the various IR schemes to the best performing model depicted in bold. We then marked with a cross ("†") the performance values depicting statistically significant differences. In this case, the classical *tf idf* vector-space model and the language model (LM) offer a performance level that is significantly lower for all languages and topic formulations. For the other models, the answer depends on the language and the

IR model. We can see however that the performance differences between the DFR-PB2 and DFR-I(n_e)C2 are never significant. With the Okapi model, we found just one significant performance difference (Hindi corpus and with TDN topic formulation). Finally we can observe mixed results when analyzing the retrieval effectiveness differences with the I(n_e)C2 and the two remaining DFR implementations, namely DFR-GL2 and DFR-PL2. These differences are usually significant with the Hindi corpus, and usually not when considering the Bengali or Marathi language.

Table 4. MAP of Various IR Models (Light Stemming, Bengali Corpus, 50 queries)

	Mean Average Precision		
Light Stemming Model	Bengali T	Bengali TD	Bengali TDN
Okapi	0.3527	0.4256	0.4925
DFR-PB2	**0.3586**	**0.4405**	0.4980
DFR-GL2	0.3350	0.4081	0.4697 †
DFR-PL2	0.3434	0.4221 †	0.4872
DFR-I(n_e)C2	0.3543	0.4383	**0.5026**
LM	0.3102 †	0.3946 †	0.4720 †
tf idf	0.1750 †	0.2061 †	0.2456 †
Average	0.3185	0.3908	0.4525
Change % over T	base	+22.70%	+42.10%

Table 5. MAP of Various IR Models (Light Stemming, Marathi Corpus, 39 queries)

	Mean Average Precision		
Light Stemming Model	Marathi T	Marathi TD	Marathi TDN
Okapi	0.2986	0.3446	0.3855
DFR-PB2	0.2921	0.3247	0.3910
DFR-GL2	0.2884	0.3336	0.3745
DFR-PL2	0.2849	0.3177 †	0.3658 †
DFR-I(n_e)C2	**0.3075**	**0.3502**	**0.4137**
LM	0.2892 †	0.3185 †	0.3824 †
tf idf	0.2024 †	0.2286 †	0.2535 †
Average	0.2804	0.3168	0.3666
Change % over T	base	+12.98%	+30.73%

In the bottom part of these three tables, we can find under the label "Average" the average MAP across the seven IR models. The last line shows the relative percentage of variation obtained when compared to the short (T) query formulation.

From these tables we can see that enlarging the query formulation (from T to TD, and from TD to TDN) brings the improvement of the retrieval effectiveness for all three languages in question. This improvement is less for the Marathi language (see Table 5) than for both the Hindi (see Table 3) and Bengali corpus (see Table 4). The

longest topic formulation improve the MAP from 30.73% for the Marathi language to 58.56% for the Hindi collection, showing clearly the need of more search terms in order to perform an effective search.

When applying our statistical test with the performance achieved under the Title-only topic formulation as baseline, we always found statistically significant differences with either TD or TDN topic formulation. Thus when the user is able to enlarge the query, the retrieval effectiveness is significantly improved.

4.3 Evaluation of Various Stemming Strategies

In the previous section we compared different topic formulations using the light stemmer. In order to investigate whether others stemming strategies may improve the retrieval effectiveness, we need to consider a more aggressive stemmer on the one hand and, on the other, ignoring this word normalization process (no stemming). For the Hindi language, Table 6 depicts the MAP obtained by various IR models using these three different stemming approaches (TD query formulation). Similar conclusions can be obtained with T or TDN query formulations.

Table 6. MAP of Various Stemming Strategies, TD queries (Hindi Corpus)

	Mean Average Precision		
TD Model	Hindi no stemmer	Hindi light	Hindi aggressive
Okapi	0.3835	0.3717	0.3986
DFR-PB2	0.3908	0.3737	0.3875
DFR-GL2	0.3627	0.3524	0.3684
DFR-PL2	0.3740	0.3621	0.3873
DFR-I(n_e)C2	**0.3917**	**0.3836**	**0.4067**
LM	0.3481	0.3310 †	0.3523
tf idf	0.1975	0.1830 †	0.1833 †
Average	0.3498	0.3368	0.3549
Change %	base	-3.17%	+1.46%

Using the same condition, MAP obtained with the Bengali corpus is reported in Table 7 and Table 8 depicted the same information with the Marathi language. While for the Hindi the light stemming hurts MAP in mean (-3.17% compared to "no stemmer" scheme), for the Bengali and Marathi languages this indexing scheme tends to produce, in mean, better retrieval effectiveness compared to an indexing strategy ignoring the stemming stage. On the other hand, the aggressive stemming results in an improvement in MAP for all three languages, the performance difference comparing to no stemming being rather small for Hindi language (e.g., +1.46% in mean over 7 IR models) while being much more important for both Bengali and Marathi languages (e.g., +19.57% and +33.76% respectively).

When applying our statistical test to verify whether these performance differences are statistically significant, we used the performance achieved without any stemming normalization as baseline (values shown under the label "no stemmer"). We marked with a cross ("†") the performance values depicting statistically significant differences.

Table 7. MAP of Various Stemming Strategies, TD queries (Bengali Corpus)

	Mean Average Precision		
TD Model	Bengali no stemmer	Bengali light	Bengali aggressive
Okapi	0.3640	0.4256 †	**0.4446** †
DFR-PB2	0.3629	**0.4405** †	0.4366 †
DFR-GL2	0.3498	0.4081 †	0.4405 †
DFR-PL2	0.3550	0.4221 †	0.4311 †
DFR-I(n_e)C2	**0.3673**	0.4383 †	0.4443 †
LM	0.3402	0.3946 †	0.4078 †
tf idf	0.2179	0.2061	0.2136
Average	0.3367	0.3908	0.4026
Change %	base	16.05%	19.57%

Table 8. MAP of Various Stemming Strategies, TD queries (Marathi Corpus)

	Mean Average Precision		
TD Model	Marathi no stemmer	Marathi light	Marathi aggressive
Okapi	0.2872	0.3446 †	0.3958 †‡
DFR-PB2	0.2947	0.3247	0.3928 †‡
DFR-GL2	0.2937	0.3335 †	0.3829 †
DFR-PL2	0.2767	0.3177 †	0.3891 †‡
DFR-I(n_e)C2	**0.2976**	**0.3502** †	**0.4118** †‡
LM	0.2907	0.3185 †	0.3824 †‡
tf idf	0.2023	0.2286	0.2435
Average	0.2775	0.3168	0.3712
Change %	base	14.17%	33.76%

With the Hindi language (see Table 6), the performance differences between the three stemming approaches are usually not significant, except with the classical *tf idf* vector-space model. As depicted in Table 7, the Bengali corpus presents a different situation where all performance differences are significant, except with the *tf idf* model. With the Marathi language (see Table 8), we found a similar conclusion demonstrating that applying a stemmer tends to improve the retrieval effectiveness with language owning a more complex inflectional morphology than English or Hindi.

Finally, we want to analyze the differences in retrieval effectiveness when applying the light stemmer (baseline) with the aggressive stemming approach. Each statistically

significant difference is marked with a double cross ("‡") in Table 6 to 8. As we can see, for both the Hindi and Bengali languages, we detect no significant difference between the light and a more aggressive stemmer. For the Marathi corpus however, the aggressive stemming scheme tends to produce significantly better results as depicted in Table 8.

4.4 Evaluation of Various Indexing Strategies

Finally we will compare different document and query representation strategies. Instead of being limited to a word-based surrogate, we want to evaluate the effectiveness of two language-independent indexing strategies, namely 4-gram [19] and trunc-4 [20]. For the Hindi (see Table 9), Bengali (see Table 10), and Marathi language (see Table 11), we have computed the MAP of these indexing approaches and compared them to the word-based indexing scheme with the aggressive stemmer and using TD query formulation.

Table 9. MAP of Various Indexing Strategies, TD queries (Hindi Corpus)

	Mean Average Precision		
TD Model	Hindi aggressive	Hindi 4-gram	Hindi trunc-4
Okapi	0.3986	0.3674	0.3770
DFR-PB2	0.3875	0.3704	0.3687
DFR-GL2	0.3684	0.3479	0.3569
DFR-PL2	0.3873	0.3554	0.3719
DFR-I(n_e)C2	**0.4067**	**0.3841**	**0.3780**
LM	0.3523	0.3366	0.3378
tf idf	0.1833	0.1837	0.1844
Average	0.3549	0.3351	0.3392
Change %	base	-5.58%	-4.40%

Table 10. MAP of Various Indexing Strategies, TD queries (Bengali Corpus)

	Mean Average Precision		
TD Model	Bengali aggressive	Bengali 4-gram	Bengali trunc-4
Okapi	**0.4446**	0.3803 †	**0.4522**
DFR-PB2	0.4366	0.3875 †	0.4395
DFR-GL2	0.4405	0.3740 †	0.4260
DFR-PL2	0.4311	0.3841	0.4377
DFR-I(n_e)C2	0.4443	**0.3876** †	0.4493
LM	0.4078	0.3557	0.4063
tf idf	0.2136	0.2143	0.1921
Average	0.4026	0.3548	0.4004
Change %	base	-11.89%	-0.55%

Results depicted in these tables tend to indicate that both language-independent indexing strategies result in similar performance levels when facing with the Hindi language (e.g., in average, -5.58% for 4-gram, and -4.40% for trunc-4, see Table 9). Applying our statistical test, all performance differences compared to the word-based with an aggressive stemmer are not significant. When comparing with either the light or no stemmer, we reached the same conclusion: no significant performance differences for all IR models with the Hindi corpus.

For both Bengali and Marathi language, results depicted respectively in Table 10 and 11 tend to indicate that trunc-4 language-independent indexing strategy result in similar retrieval effectiveness, in average, when compared to a word-based indexing scheme with an aggressive stemmer (e.g., -0.55% for the Bengali, +3.18% for the Marathi corpus). We can also find that the trunc-n tends to produce better retrieval results than the n-gram scheme that is also more complex to implement and require more query and indexing processing time. The statistical tests can however detect significant performance differences only when comparing word-based (aggressive stemmer) with 4-gram with the Bengali corpus (indicated by a cross "†" in Table 10).

Table 11. MAP of Various Indexing Strategies, TD queries (Marathi Corpus)

	Mean Average Precision		
TD Model	Marathi aggressive	Marathi 4-gram	Marathi trunc-4
Okapi	0.3958	0.3525	0.4161
DFR-PB2	0.3928	0.3329	0.4191
DFR-GL2	0.3830	0.3653	0.3881
DFR-PL2	0.3891	0.3440	0.3972
DFR-I(n_e)C2	**0.4118**	**0.3744**	**0.4347**
LM	0.3824	0.3592	0.3920
tf idf	0.2435	0.2204	0.2337
Average	0.3712	0.3355	0.3830
Change %	base	-9.61%	3.18%

When using a light stemmer as baseline and with the Bengali language, the performance differences are usually significant with the 4-gram approach, and not significant with the trunc-4 scheme. For Marathi (see Table 11) and using the light stemmer as baseline, no significant difference can be found with the 4-gram indexing scheme, but when compared with the trunc-4, the retrieval effectiveness differences are always significant, except with the *tf idf* model.

4.5 Some Query-by-Query Analysis

To obtain a better understanding of effects associated with different stemming and indexing strategies, we analyzed a few Hindi queries. As a first example we can inspect Topic #108 "Greater Nagaland" (owning 13 relevant items). This query performs poorly with Title-only topic formulation, achieving an average precision (AP)

of 0.0963 with word-based indexing (light stemmer). The performance increases when considering TD topic formulation (AP = 0.7393) or with TDN topic formulation (AP = 0.8516). The internal representation of this query is limited to one term ("वृहत्तर" or "greater") when using the Title-only topic formulation, and this ineffective surrogate cannot retrieve pertinent information. When including related terms from the descriptive or narrative section (such as the name "एनएससीएन"), the overall performance increases.

To analyze the effect of an aggressive stemming procedure in representing the query, we can look at Topic #77 "Attacks by Hezbollah guerrillas" (with eight relevant items). When using a word-based indexing scheme ignoring the stemming normalization, the average precision is rather low (0.1063). Using a light stemmer, we degrade the retrieval effectiveness for this query (AP = 0.0340). After employing an aggressive stemmer, the performance increases to 0.3238. The main reason is related to the word "Hezbollah" ("हिजबुल्लाह") that cannot retrieve many relevant documents owning the related expression "हिजबु" ("Hezbo"). This second term can be found only after applying the aggressive stemmer, thus offering more matching possibilities between the query and other relevant articles.

To illustrate the difference between word-based and *n*-gram indexing strategies, we analyze Topic #107 "Furore over the release of a CD containing anti-Muslim sentiments in Uttar Pradesh" (owning ten relevant items). Using a word-based representation with a light stemmer, we can achieve an AP of 0.6491, but using a 4-gram scheme the performance is clearly lower (AP = 0.0692). In the relevant articles, we can usually find the state name "Uttar Pradesh" and "Muslim". With the 4-gram indexing scheme, many non-relevant items appear in the top of the ranked list due to the fact that they have the term "प्रदेशा" ("Pradesh", meaning *state*) instead of "प्रदेश" (the same form but without an additional letter at the end). With word-based, the search system can discriminate between the correct answers and the irrelevant ones. This capability is less accurate with the 4-gram representation, and thus the performance decreases.

4.6 Pseudo-relevance Feedback

Previous experiments with different languages and corpora tend to indicate that pseudo-relevance feedback (PRF or blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we have adopted Rocchio's approach [21]) with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add m terms extracted from the k best ranked documents from the original query (see Table 12). From our previous experiments we learned that this type of blind query expansion strategy does not always work well. More particularly, we believe that including terms occurring frequently in the corpus (because they also appear in the top-ranked documents) may introduce more noise, and thus be an ineffective means of discriminating between relevant and non-relevant items [22]. Consequently we also chose to apply our *idf*-based query expansion model [23].

Using the Rocchio's method, Table 12 shows that the retrieval performance after applying a pseudo-relevance feedback approach can be improved for both the Hindi and Marathi corpus, not with the Bengali. The best result for the Hindi language indicates an enhancement of +16.8% (from 0.4067 to 0.4750), while for the Marathi we

were able to increase the MAP of +5.8% (from 0.4118 to 0.4359). To verify whether these performance differences are statistically significant, we applied our statistical test with the performance before blind-query expansion as baseline (values shown in the third row). We marked with a cross ("†") the MAP values depicting a statistically significant difference. As depicted in Table 12, only a few parameter settings were able to achieve a significant performance difference over the baseline.

Table 12. MAP of Different Blind-Query Expansions, Rocchio's method, TD queries

	Mean Average Precision		
TD Model	Hindi aggressive	Bengali aggressive	Marathi aggressive
DFR-I(n_e)C2	0.4067	**0.4443**	0.4118
10 docs / 10 terms	0.4465	0.3729 †	0.4224
10 docs / 30 terms	0.4688	0.4085	0.4303
10 docs / 50 terms	0.4714 †	0.4060	**0.4359**
10 docs / 70 terms	0.4729 †	0.4137	0.4347
10 docs / 100 terms	**0.4750** †	0.4131	0.4453
15 docs / 10 terms	0.4327	0.3966 †	0.4041
15 docs / 30 terms	0.4691 †	0.4007	0.4162
15 docs / 50 terms	0.4647	0.3965	0.4195
15 docs / 70 terms	0.4642	0.4016	0.4192
15 docs / 100 terms	0.4608	0.4014	0.4193

Table 13. MAP of Different Blind-Query Expansions, *idf*-based method, TD queries

	Mean Average Precision		
TD Model	Hindi aggresive	Bengali aggressive	Marathi aggressive
DFR-I(n_e)C2	0.4067	0.4443	0.4118
10 docs / 10 terms	0.4333	0.4270	0.4328
10 docs / 30 terms	0.4721 †	0.4416	0.4490
10 docs / 50 terms	0.4766 †	0.4544	0.4551
10 docs / 70 terms	**0.4829** †	0.4430	0.4495
10 docs / 100 terms	0.4760 †	0.4422	0.4550
15 docs / 10 terms	0.4123	0.4133	0.4281
15 docs / 30 terms	0.4695	**0.4578**	0.4577
15 docs / 50 terms	0.4801 †	0.4539	**0.4605**
15 docs / 70 terms	0.4703 †	0.4550	0.4473
15 docs / 100 terms	0.4633 †	0.4510	0.4406

Based on our *idf*-based blind-query expansion [23], Table 13 reports the results using different parameter settings across the three languages. The best improvement was obtained with the Hindi corpus (+18.7%, from 0.4067 to 0.4829) followed by the Marathi language (+11.8%, from 0.4118 to 0.4605). For the Bengali, the enhancement was smaller (+3%, from 0.4443 to 0.4578). Compared to the results obtained

with the Rocchio's method (see Table 12), the *idf*-based approach performs better. When applying our statistical test to verify whether these performance differences are statistically significant, we select the performance before blind-query expansion as baseline (MAP values shown in the third row). We marked with a cross ("†") the retrieval effectiveness values depicting a statistically significant difference. As depicted in Table 13, such a significant differences occurs usually only for the Hindi corpus.

5 Official Results

Table 14 shows the exact specifications of our 6 official monolingual runs for the Hindi *ad hoc* monolingual evaluation task. These runs are based on three probabilistic models (Okapi, DFR and statistical language model (LM)). In each case, we then applied a pseudo-relevance feedback stage, and finally we merged the three individual ranked lists into a fused common list based on the Z-score merging strategy [1]. Table 15 lists the same information for Bengali, showing our 6 official submissions while Table 16 reports our official experiments for the Marathi language.

To propose effective search strategies, we selected three IR probabilistic models and enlarged the query by adding 20 to 150 terms retrieved from the 3 to 10 best-ranked articles contained in the Hindi (see Table 14), Bengali (see Table 15) or Marathi (see Table 16) collection. In the last column of Table 16 we have given in brackets the official results taking into account all 50 available topics.

Table 14. Description and Mean Average Precision (MAP) for our Official Hindi Monolingual Runs

	Index	Model	PRF	MAP	MAP
1	trunc-4	PL2	3 / 20	0.4440	Z-score
TD	4-gram	LM	10 / 50	0.3741	**0.4904**
	aggressive	I(n_e)C2	10 / 70	0.4904	
2	trunc-4	Okapi	10 / 100	0.4201	Z-score
TD	4-gram	PL2	5 / 50	0.4338	0.4836
	aggressive	LM	3 / 20	0.3936	
3	light	PB2	10 / 20	0.4504	Z-score
TD	aggressive	PL2	10 / 50	0.4431	0.4686
	trunc-4	PB2	3 / 20	0.4027	
4	light	PL2	10 / 20	0.4546	Z-score
TD	trunc-4	PL2	3 / 20	0.4440	0.4879
	4-gram	PL2	5 / 100	0.4477	
5	trunc-4	PB2	10 / 50	0.5193	Z-score
TDN	aggressive	PL2	5 / 50	0.5142	0.5339
	light	Okapi	10 / 100	0.4747	
6	4-gram	LM	5 / 50	0.4468	Z-score
TDN	trunc-4	Okapi	3 / 100	0.4942	**0.5467**
	aggressive	I(n_e)C2	3 / 50	0.4981	

Table 15. Description and Mean Average Precision (MAP) for our Official Bengali Monolingual Runs

	Index	Model	PRF	MAP	MAP
1	trunc-4	PL2	10 / 70	0.4679	Z-score
TD	4-gram	LM	10 / 70	0.4100	0.4646
	aggressive	PB2	5 / 50	0.4352	
2	light	I(n_e)C2	10 / 70	0.4327	Z-score
TD	aggressive	PL2	5 / 20	0.4590	0.4731
	trunc-4	LM	10 / 50	0.4234	
3	4-gram	Okapi	5 / 150	0.3899	Z-score
TD	trunc-4	PL2	10 / 70	0.4679	0.4684
	aggressive	PB2	5 / 50	0.4352	
4	trunc-4	Okapi	5 / 100	0.4660	Z-score
TD	aggressive	PB2	10 / 50	0.4576	**0.4862**
5	trunc-4	PB2	5 / 20	0.4866	Z-score
TDN	4-gram	Okapi	5 / 150	0.4594	0.5329
	aggressive	PL2	10 / 70	0.5021	
6	light	PL2	10 / 70	0.5165	Z-score
TDN	trunc-4	PB2	5 / 20	0.4866	**0.5438**
	4-gram	GL2	5 / 50	0.4659	

Table 16. Description and Mean Average Precision (MAP) for our Official Marathi Monolingual Runs

	Index	Model	PRF	MAP	MAP
1	trunc-4	GL2	5 / 20	0.4437	Z-score
TD	4-gram	I(n_e)C2	3 / 50	0.4273	**0.5009**
	aggressive	PB2	10 / 50	0.4541	(0.3907)
2	trunc-4	LM	10 / 20	0.4718	Z-score
TD	4-gram	Okapi	5 / 50	0.4197	0.4897
	aggressive	PL2	10 / 50	0.4610	(0.3820)
3	trunc-4	LM	10 / 20	0.4718	Z-score
TD	aggressive	PB2	10 / 50	0.4541	0.4817
	light	Okapi	10 / 20	0.4079	(0.3757)
4	4-gram	I(n_e)C2	3 / 50	0.4273	Z-score
TD	light	GL2	5 / 70	0.4113	0.4885
	trunc-4	PL2	10 / 70	0.4790	(0.3810)
5	trunc-4	Okapi	3 / 70	0.4474	Z-score
TDN	aggressive	LM	10 / 70	0.5412	**0.5355**
	light	LM	10 / 100	0.4729	(0.4177)
6	trunc-4	LM	10 / 20	0.4910	Z-score
TDN	4-gram	Okapi	5 / 150	0.4182	0.5126
	aggressive	PL2	5 / 50	0.4878	(0.3998)

For the Hindi, Marathi and Bengali corpora, we have submitted four runs with TD formulation and two additional runs with the longest TDN query formulation in order to enhance the quality of the final pool. All runs were fully automated using our

stopword lists and different word-based and language-independent indexing strategies. Furthermore, in order to improve the overall retrieval effectiveness, we may consider different merging strategies [24], [25]. In all cases the same Z-score data fusion approach (see details in [1]) was applied.

6 Conclusion

The results achieved in FIRE 2010 evaluation campaign confirm the retrieval effectiveness of models derived from *Divergence from Randomness* (DFR) paradigm. Implementations of the DFR-I(n_e)C2 or DFR-PB2 tend to produce high MAP when facing different test-collections, in this case Hindi, Marathi, and Bengali collections. Moreover, the effectiveness of these models proves to be independent of underlying indexing strategy or query formulation. After applying a statistical test, we can conclude that the retrieval effectiveness differences were always significant when comparing the best result (DFR-I(n_e)C2 or DFR-PB2) with either the *tf idf* or LM approach. Usually, other implementations of the DFR family (DFR-GL2, DFR-PL2) tend to achieve lower performance levels for which the differences are however usually not statistically significant.

For all three languages studied we have found that enlarging the topic formulation from T to TD, or from TD to TDN (and of course from T to TDN) will improve retrieval effectiveness (up to 58% in mean, over 7 models when comparing T to TDN query formulation for Hindi collection). When enlarging the query from the Title-only topic description, performance differences were always statistically significant.

For each language and based on our experiments we have reached following conclusions regarding usage of various indexing strategies. For the Hindi language, all stemming strategies produce similar levels of performance and the differences were usually not significant. The light stemming or language-independent indexing strategies resulted in lower MAP when compared to no stemming approach. However, incorporating the aggressive stemming brings slight but not significant improvement in MAP.

For the Marathi language our experiments tend to show that an aggressive stemming approach performs significantly better than a light stemmer. This last approach is better than no stemming. While light stemming and 4-gram approaches result in comparable performances, aggressive stemming or trunc-4 brings a clear improvement in MAP for this language presenting more complex morphology.

For the Bengali language, usage of a light or aggressive stemming generates significantly better results than an indexing scheme ignoring the stemming normalization. The performance differences between the light and the aggressive stemmer are not significant. For this language, a trunc-4 language-independent approach or a word-based with an aggressive stemmer result in usually significant better performance levels than a 4-gram indexing scheme.

Of course we do not have the needed resources to investigate all possible and pertinent research questions dealing with the Hindi, Bengali, and Marathi languages. We must also mention that the elaboration of test-collections with other languages

families (e.g., Dravidian languages such as Telugu or Tamil) is, from our point of view, an important task in order to have a better understanding of the underlying problems dealing with the automatic processing of various Indian languages.

Acknowledgement. The authors would like to thank the FIRE-2010 task organizers for their efforts in developing various Indian language test-collections. This research was supported in part by the Swiss National Science Foundation under Grant #200020-129535/1.

References

1. Savoy, J.: Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. IR Journal 7, 121–148 (2004)
2. Savoy, J.: Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. ACM - Transactions on Asian Languages Information Processing 4, 163–189 (2005)
3. Dolamic, L., Savoy, J.: UniNE at FIRE 2008: Hindi, Marathi and Bengali IR. FIRE 2008 Working Notes (2008)
4. Voorhees, E.M., Harman, D.K. (eds.): TREC. Experiment and Evaluation in Information Retrieval. The MIT Press, Cambridge (2005)
5. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. Information Processing & Management 36, 95–108 (2002)
6. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems 20, 357–389 (2002)
7. Hiemstra, D.: Using Language Models for Information Retrieval. Ph.D. Thesis (2000)
8. Hiemstra, D.: Term-Specific Smoothing for the Language Modeling Approach to Information Retrieval. In: Proceedings of ACM-SIGIR, pp. 35–41. The ACM Press (2002)
9. Zhai, C., Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Information Retrieval. ACM Transactions on Information Systems 22, 179–214 (2004)
10. Fox, C.: A Stop List for General Text. ACM-SIGIR Forum 24, 19–35 (1990)
11. Dolamic, L., Savoy, J.: When Stopword Lists Make the Difference. Journal of the American Society for Information Sciences and Technology 61, 200–203 (2010)
12. Savoy, J.: Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages. In: Proceedings of ACM-SAC, pp. 1031–1035. The ACM Press (2006)
13. Harman, D.K.: How Effective is Suffixing? Journal of the American Society for Information Science 42, 7–15 (1991)
14. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14, 130–137 (1980)
15. Fautsch, C., Savoy, J.: Algorithmic Stemmers or Morphological Analysis: An Evaluation. Journal of the American Society for Information Sciences and Technology 60, 1616–1624 (2009)
16. Buckley, C., Voorhees, E.M.: Retrieval System Evaluation. In: Voorhees, E.M., Harman, D.K. (eds.) TREC. Experiment and Evaluation in Information Retrieval, pp. 53–75. The MIT Press, Cambridge (2005)
17. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management 33(4), 495-512

18. Abdou, S., Savoy, J.: Statistical and Comparative Evaluation of Various Indexing and Search Models. In: Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (eds.) AIRS 2006. LNCS, vol. 4182, pp. 362–373. Springer, Heidelberg (2006)
19. McNamee, P., Mayfield, J.: Character N-gram Tokenization for European Language Text Retrieval. IR Journal 7, 73–97 (2004)
20. McNamee, P., Nicholas, C., Mayfield, J.: Addressing Morphological Variation in Alphabetic Languages. In: Proceedings of ACM-SIGIR 2009, pp. 75–82. The ACM Press (2009)
21. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: Proceedings of TREC-4, pp. 25–48. NIST Publication #500-236, Gaithersburg (1996)
22. Peat, H.J., Willett, P.: The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems. Journal of the American Society for Information Science 42, 378–383 (1991)
23. Abdou, S., Savoy, J.: Searching in Medline: Stemming, Query Expansion, and Manual Indexing Evaluation. Information Processing & Management 44, 781–789 (2008)
24. Vogt, C.C., Cottrell, G.W.: Fusion via a Linear Combination of Scores. IR Journal 1, 151–173 (1999)
25. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: Proceedings of TREC-2, pp. 243–249. NIST Publication #500-215, Gaithersburg (1994)