Chapter 5

A study on corpus-based stopword lists in Indian language IR

5.1 Introduction

We explore and evaluate the effect of different stopword lists (non-corpus-based and corpus-based) in the IR tasks with different Indian languages such as Bengali, Marathi, Gujarati, Hindi and English. The issue is investigated from three viewpoints. Is there any difference in retrieval effectiveness between non-corpus-based and corpus-based stopword removal? Can a corpus-based stopword list improve retrieval effectiveness? If yes, to what extent? Does the length of a corpus-based stopword list affect the retrieval effectiveness in Indian languages? If yes, to what extent? Among different possible lengths of corpus-based stopword lists, which one should be chosen for IR? Is it similar across different Indian languages or different for different languages?

It is observed that a corpus-based stopword list provides a better MAP score than a non-corpus-based stopword list in different Indian languages. Among the different corpus-based stopword lists generated and experimented with, Zipf's law-based stopword list (idf-based one) provides the best retrieval effectiveness in different Indian languages. The aggregation1-based stopword list offers a better MAP score than the aggregation2-based list in Indian languages (details will follow). However, the aggregation2-based stopword list offers a better MAP score than the aggregation1-based list in English. The best performing idf-based stopword list improves MAP score by 5.43% in Bengali, 1.91% in Marathi, 5.4% in Gujarati, 1.5% in Hindi, and 2.12% in English, respectively, over their

baseline counterparts. The probabilistic retrieval models (BM25 and TF-IDF) perform best in different Indian languages. A smaller length of corpus-based stopword lists performs better than its larger length counterpart for all the Indian languages considered. The proposed schemes demonstrate that a stopword list can be heuristically generated in a language-independent statistical method and effectively used for IR tasks with retrieval effectiveness comparable to, or in some cases, even better than, non-corpus-based stopword lists.

5.2 Problem Formulation

During the literature survey, we found that stopwords are used as the syntactic structure of a sentence [146], but less to carry meaningful information. A stopword list typically comprises pronouns, conjunctions, adverbs, prepositions, and interjections from the linguistic point of view.

However, different methods have been proposed for generating and removing stopwords in low-resource languages. Some frequently used methods are:

- selection based on the probability distribution of a term
- selection based on dictionary or lexicon
- minimal DFA-based approach (as proposed by Fox [41])
- SMART[1] stopword list and
- stopword generation based on word statistics.

However, no language-agnostic corpus-specific stopword lists have been proposed and evaluated in the Indian language IR. This study proposes different corpus-based stopword lists in Indian languages by evaluating a term's statistical feature and entropy. We also consider some aggregation methods. The proposed stopword lists are evaluated in the IR domain.

[1]https://github.com/igorbrigadir/stopwords/blob/master/en/smart.txt

5.3 Methods Used

5.3.1 Statistical Method

George Kingsley Zipf [145] observed that the term's rank-frequency distribution could be fitted very closely by the following relation.

$$F(r) = \frac{C}{r^{\alpha}} \tag{5.1}$$

The above equation, known as Zipf's law, states that the frequency of a word (F) at rank r is inversely proportional to its rank when the words are arranged in decreasing order of their frequency of occurrence, α and C being two tuning parameters. During the literature study, we observe that the $\alpha = 1$ and $C = 0.1$ provide the best retrieval effectiveness in the standard corpus of English ([70], [78]). We thus tried $\alpha = 1$ and $C = 0.1$ for different Indian languages as well. Zipf's law inspires our first four-stopword-generation approaches. We propose Zipf's law-based stopword lists using the following steps.

Algorithm for Zipf's law-based stopword generation approach

- Calculate the term frequency of each term in a given corpus.
- Sort the terms in decreasing order of the term frequencies, i.e., the term with the highest term frequency at rank 1, the term with the second highest at rank 2, ..., and so on.
- Plot a graph of term frequency vs ranks. The graph is supposed to obey Zipf's law.
- Determine a threshold value. The words above the threshold value are considered stopwords for the collection.
- Use the above-generated stopword list for retrieval and note the MAP for a set of queries.

We propose stopword lists in different Indian languages following the above algorithm. Instead of using only the 'term frequency', we use other variants like normalized term frequency, inverse document frequency and normalized inverse document frequency (defined below) to get different stopword lists.

Term frequency (tf): It is given by the number of times a particular term (t) appears in a document (d) and represented by $tf_{t,d}$ or simply tf_{td}. Highly frequent terms are less informative and are considered stopwords in different languages. The statistics of a term are used in different fields of text analysis tasks such as text mining, information retrieval, and NLP-related areas.

Normalised term frequency (ntf): Here, we normalize the term frequency in two ways. In the first variation, we normalize the frequency of each term t over the entire lexicon size and take its logarithm. Secondly, we normalize by summing up term frequency in a document (d) by the number of tokens 'n_d' in that document. Normalization reduces the frequency differences among different words. The mathematical formulae for these two normalisations are given by Eqn 5.2 and Eqn 5.3, respectively.

$$tf_{norm1}(t) = -\log \frac{tf_{td}}{m} \tag{5.2}$$

$$tf_{norm2}(t) = \frac{\sum_{d \in C} tf_{td}}{\sum_{d \in C} n_d} \tag{5.3}$$

where,

tf_{td} : term frequency of term t in the document d

m : total number of tokens in the lexicon file

n_{d} : total number of tokens in a particular document d

C_{d} : document collection or the set of documents in handed

Inverse document frequency (idf): The inverse document frequency captures the rarity of a term (t) in a given collection. In any document collection, a few terms occur frequently across the documents, but they are of very little importance. They are stopwords. So, we must lower the weight of frequently occurring terms. The rare or least frequent terms in the document set will likely to be more informative. Hence, inverse of the document frequency (the number of documents where a particular term occurs, represented as df) of a term (t) is given by dividing the total number of documents in the collection (N) by the df of the term t.

$$idf_t = \log \frac{N}{df_t} \tag{5.4}$$

Normalised idf (nidf): The most common form of idf weighting is the one used by Robertson and Jones ([110], [78]), which normalizes with respect to the number of documents not containing the term ($N - df_t$) and adds a constant of 0.5 to both numerator and denominator to moderate extreme values:

$$idf_{\text{norm}} = \log(\frac{N - df_t + 0.5}{df_t + 0.5}) \tag{5.5}$$

Terms having inverse document frequency values less than a chosen threshold are considered stopwords for a collection.

In the literature, no recommendations exist from the statistical methods on the length of the stopword lists applicable to different Indian languages. We, therefore, heuristically generate different lengths of stopword lists using different threshold values and experiment with such lists for each stopword generation approach. The best MAP score in each method is noted in Section 5.5. Also, the best MAP score obtained at a particular length of the stopword list is shown in Table 5.6.

5.3.2 Data Entropy Measure

According to Shannon [123], the measure of word randomness is called entropy. A word with a high probability of occurrence has less surprise element and hence contains low information - hence likely to be a stopword in the collection. On the contrary, a word with very little probability of occurrence will likely be informative. Entropy, which measures the average amount of information contained in a word across the documents in a given collection, can thus be used for stopword selection [146]. Suppose there are M distinct words and N documents altogether. Frequency of a word w_j $(j = 1, \ldots, M)$ in a document d_i $(i = 1, \ldots, N)$ is represented as d_{ij}. The probability of occurrence of a word w_j can be considered as the relative frequency of w_j with respect to the total number of words present in the document d_i and represented as p_{ij}. Hence, we measure the information contained in the word $H(w_j)$ by equation 5.6.

$$H(w_j) = \sum_{i=1}^{N} p_{ij} \cdot log(\frac{1}{p_{ij}}) \tag{5.6}$$

5.3.3 Term Variance-based Approach

Another attribute of a stopword is its homogeneous distribution across the documents. If a word is more spread in different documents, i.e., the term has higher variance across the documents, it is more likely to be a stopword. We measure the data spread by Equation 5.7.

$$S_t^2 = \frac{\sum_d (tf_{td} - \overline{tf_d})^2}{n-1} \tag{5.7}$$

where,

S_t^2 : sample variance

tf_{td} : term frequency of term t in document d

$\overline{tf_d}$: mean term frequency in document d

n : number of distinct terms in document d

5.3.4 Term-based Random Sampling Approach

In this approach, a term is considered a stopword based on its importance in a document. A term is more likely to be a stopword if it is less important. The Kullback-Leibler divergence measure is used to determine the importance of a term in the documents [25]. The weight $w(t)$ of a term t in a document using the Kullback-Leibler divergence measure is given by Equation 5.8.

$$w(t) = p_t \cdot \log \frac{p_t}{p_a} \tag{5.8}$$

where,

$$p_t = \frac{tf_t}{l_t} \tag{5.9}$$

$$p_a = \frac{z}{token_a} \tag{5.10}$$

tf_t : term frequency of term t in the sampled document set

l_t : sum of the length of the sampled document set in terms of number of terms

z : term frequency of term t in the collection

$token_a$: total number of tokens in the collection

The term-based random stopword generation approach aligns with the earlier works of Lo et al. [78]. Here, terms with lower $w(t)$ are considered stopwords. We propose stopword lists in different Indian languages in the following steps.

Algorithm for term-based random sampling approach

- Choose a random term (t) in the lexicon.
- Retrieve all the documents that contain the term (t).
- Assign a weight to each retrieved term (t) by applying the Kullback-Leibler divergence measure. The weight indicates the importance of a term in documents.
- Normalize each weight of a term (t) by its maximum weight.
- Rank the terms in ascending order of KL weights $w(t)$, and less importance of a term in documents is a stopword.
- Extract the top-ranked terms as stopwords in the collection.

We experiment with different lengths of stopword lists, and the best MAP score is noted in Section 5.5.

5.3.5 Aggregation Method

We propose two more stopword lists in different Indian languages based on two types of aggregation. First, (aggregation1 method), we aggregate three stopword lists, i.e., term frequency, inverse document frequency (Eqn. 5.4) and entropy-based (Eqn. 5.6). In the second method, aggregation2, we aggregate another three stopword lists, i.e., normalised term frequency (Eqn.5.3), variance (Eqn.5.7) and entropy-based (Eqn.5.6). We aggregate the top few ranked words in each stopword generation approach. During aggregation, we

observe that most stopwords are common in different stopword generation approaches. We apply an intersection and union method for aggregation to avoid common stopwords. Since there is no ideal length of a stopword list that can be used for different Indian languages, we heuristically experiment with different stopword lists. The best MAP score for a given stopword list is shown in Section 5.5.

Different corpus-based stopword lists are available on GitHub [2].

5.4 Experimental Setup

We propose different stopword lists on several standard test collections using different stopword generation approaches described in Section 5.3 and then evaluate their effectiveness in the IR domain through the following steps.

1. From the corpus, generate different stopword lists by the above stopword generation approaches described in Section 5.3.

2. Manually remove proper nouns for named entities (such as names of persons, objects, places, etc.) which are not likely to be a stopword.

3. Perform retrieval experiments using different retrieval models and calculate MAP.

4. Compare the retrieval results among different stopword evaluation approaches (none, non-corpus-based, Fox-based [41], SMART[3], tf, idf, ntf, nidf, term-based random sampling, aggregation).

We compare the MAP scores of different stopword generation approaches with no stopword removal (None) as the baseline. The non-corpus-based stopword list is extracted from GitHub [4] for different Indian languages. Here, we explore the effect of different non-corpus-based and corpus-based stopword removal in different Indian languages IR from the following point of view.

[RQ 1.3] Is there any difference in retrieval effectiveness between non-corpus-based and corpus-based stopword removal? In this experiment, we evaluate the effect of

[2]https://github.com/cse-iitbhu/Corpus-based-stopword-list
[3]https://github.com/igorbrigadir/stopwords/blob/master/en/smart.txt
[4]https://github.com/cse-iitbhu/Non-corpus-based-stopword-list

different non-corpus-based and corpus-based stopword removal in the Indian language IR.

[RQ 1.4] Can a corpus-based stopword list improve retrieval effectiveness? If yes, to what extent? Among the different corpus-based stopword lists, which one performs the best in the IR setting? We experiment with different corpus-based stopword lists based on word statistics. Further, we suggest a stopword generation approach that performs the best in the Indian language IR.

[RQ 1.5] Does the length of the corpus-based stopword list affect retrieval effectiveness in Indian languages? If yes, to what extent? Among different possible lengths of corpus-based stopword lists, which one should be chosen for IR? Is it similar across different Indian languages or different for different languages? We could not find a single particular stopword length that can be used for different Indian languages. However, we heuristically evaluated different stopword lengths, and the best MAP score is noted in Section 5.5.

5.5 Evaluation

To address the research questions (RQs) described in the above section, we detail the results and observations below.

5.5.1 Effect of non-corpus-based and corpus-based stopword list on retrieval

In the first set of experiments, we evaluate the effect of different non-corpus-based and corpus-based stopword removal in different Indian languages IR. In Bengali, the MAP scores of different non-corpus-based and corpus-based stopword removal are shown in Table 5.1. We also conduct similar experiments in the other four Indian languages. In Marathi, Gujarati, Hindi and English languages, the MAP scores of different non-corpus-based and corpus-based stopword removal are shown in Table 5.2, 5.3, 5.4 and 5.5 respectively. In these tables, the best MAP score by a retrieval model is shown in boldface. It is observed that both (non-corpus-based and corpus-based) stopword removal improves retrieval effectiveness in different Indian languages IR. In Marathi and Hindi, the non-corpus-based stopword list reduces MAP score in different retrieval models, whereas the

corpus-based stopword list improves MAP score. We also observe that non-corpus-based stopword removal provides poorer retrieval effectiveness than corpus-based stopword removal in Indian languages IR. However, both (non-corpus-based and corpus-based) stopword lists provide similar MAP scores in English. A non-corpus-based stopword list is independent of the collection in hand and comprises a wide range of vocabulary. Using that stopword list forces an IR system to miss some of the potentially important words, eventually reduces the effectiveness of the system. On closer observation, we find that the corpus-based stopword list outperforms the non-corpus-based list in different Indian languages (Bengali, Marathi, Gujarati and Hindi) IR. For example, the impact of the non-corpus-based stopword removal is minimal in Bengali (less than 2%) and Gujarati (less than 4%), but corpus-based stopword removal improves MAP score by more than 5%.

We also observe that the MAP scores achieved in different Indian languages are quite comparable with the MAP scores reported by Ghosh and Bhattacharya [44]. Among the retrieval models we experimented with, the probabilistic retrieval models (BM25 and TF-IDF) provide the best retrieval effectiveness. In DFR-based models (DLH, PL2 and IFB2), the effectiveness of PL2 is relatively low. The effectiveness of the language model is moderate in different Indian languages. We observe that the stopword removal improves MAP score in different retrieval models across the Indian languages.

Table 5.1: MAP scores for different methods of stopword evaluation in the Bengali (50 T queries)

Retrieval Model	None	Noncorpus based	Fox based	tf	idf	ntf	nidf	Term R.S.	Entropy	Agg1	Agg2
BM25	**0.2063**	**0.2059**	**0.2051**	**0.2139**	**0.2164**	**0.2139**	**0.2132**	**0.2073**	0.2074	**0.2138**	**0.2073**
TF-IDF	0.2058	0.2055	0.2051	0.2131	0.2157	0.2132	0.2131	0.2071	**0.2075**	0.2134	0.2069
DLH	0.1922	0.1974	0.1958	0.1999	0.2028	0.2001	0.2029	0.194	0.1936	0.2008	0.1944
PL2	0.1443	0.1494	0.1456	0.1508	0.1514	0.1507	0.1505	0.145	0.1454	0.1503	0.1451
IFB2	0.1727	0.1757	0.1732	0.1788	0.1806	0.1787	0.1793	0.1737	0.1739	0.1787	0.1737
LM	0.1879	0.1918	0.1903	0.1943	0.1971	0.1942	0.196	0.1891	0.1887	0.1953	0.1891
Mean	0.184	0.187	0.185	0.191	0.194	0.191	0.192	0.186	0.186	0.192	0.186
% Change		+1.63%	+0.05%	+3.8%	+5.43%	+3.8%	+4.34%	+1.08%	+1.08%	+4.34%	1.08%

In this dissertation, the best baseline MAP scores achieved in the FIRE dataset with Title (T) query formulation in Bengali, Marathi, Gujarati, and Hindi are 0.2063, 0.2840, 0.2788, and 0.4439, respectively. To the best of our knowledge, the previous best MAP scores reported by Akasereh Mitra and Jacques Savoy [5], Ljiljana Dolamic and Jacques

Table 5.2: MAP scores for different methods of stopword evaluation in the Marathi (39 T queries)

Retrieval Model	None	Non-corpus based	Fox based	tf	idf	ntf	nidf	Term R.S.	Entropy	Agg1	Agg2
BM25	0.2833	0.2742	0.2771	0.2885	0.2882	0.2884	0.2884	0.2863	0.2864	0.2862	**0.2866**
TF-IDF	**0.284**	**0.2743**	**0.2773**	**0.289**	**0.2888**	**0.2888**	**0.2888**	**0.2864**	**0.2866**	**0.2863**	0.2859
DLH	0.2717	0.2653	0.2686	0.2779	0.2786	0.2759	0.2757	0.2712	0.2712	0.2746	0.2713
PL2	0.21	0.2092	0.2096	0.2145	0.2194	0.2175	0.2183	0.2119	0.2116	0.2166	0.2118
IFB2	0.2528	0.2365	0.2376	0.2559	0.2566	0.2552	0.2546	0.2536	0.2539	0.2548	0.2534
LM	0.2692	0.2647	0.2655	0.2689	0.2686	0.268	0.2682	0.2695	0.2691	0.2658	0.2693
Mean	0.261	0.254	0.255	0.265	0.266	0.265	0.265	0.261	0.261	0.264	0.261
% Change		-2.68%	-2.29%	+1.53%	+1.91%	+1.53%	+1.53%	0.7%	0.7%	+1.14%	0.7%

Table 5.3: MAP scores for different methods of stopword evaluation in the Gujarati (50 T queries)

Retrieval Model	None	Noncorpus based	Non-corpus based	tf	idf	ntf	nidf	Term R.S.	Entropy	Agg1	Agg2
BM25	**0.2788**	**0.2769**	**0.2856**	**0.2881**	**0.2886**	**0.2911**	**0.2887**	**0.2802**	**0.2801**	0.2862	**0.2797**
TF-IDF	0.2744	0.2716	0.2835	0.2861	0.2854	0.2876	0.2853	0.2742	0.2747	**0.2878**	0.2742
DLH	0.2636	0.2673	0.2756	0.2809	0.2812	0.2820	0.2810	0.2666	0.2677	0.2796	0.2672
PL2	0.2272	0.2334	0.2324	0.2395	0.2397	0.2390	0.2394	0.2263	0.2264	0.2420	0.2262
IFB2	0.2530	0.2562	0.2646	0.2698	0.268	0.2674	0.2683	0.2525	0.2549	0.2677	0.2547
LM	0.2589	0.2631	0.2718	0.2705	0.2772	0.2733	0.277	0.2605	0.2627	0.2741	0.2626
Mean	0.259	0.261	0.268	0.272	0.273	0.273	.273	0.260	0.261	0.272	0.260
% Change		+0.7%	+3.47%	+5.01%	+5.4%	+5.4%	+5.4%	+0.39%	+0.7%	+5.01%	+0.7%

Savoy [36], Jacques Savoy et al. [120], and Ghosh and Bhattacharya [44] using the same FIRE dataset in Bengali, Marathi, Gujarati and Hindi languages are 0.2465, 0.2587, 0.2818, 0.4695 respectively. We also found that in Bengali and Hindi languages, the FIRE text collection statistics used by the authors Ljiljana Dolamic and Jacques Savoy [36], Jacques Savoy et al. [120] are pretty different from the FIRE text collection used in the dissertation. However, the text collection used by Ghosh and Bhattacharya [44] is the same as the dissertation used. From the above MAP scores, we find that in Bengali, Gujarati and Hindi, the MAP scores reported in the dissertation are quite low, but in Marathi comparatively better. The differences are, however, not significant. Ghosh and Bhattacharya [44] used the Indri tool kit for indexing and retrieval, while we used the Terrier retrieval system. They also used the language model for their best retrieval scores while we obtained the best scores in BM25. These differences are due to the usage of different retrieval systems (Terrier vs Indri) and models with different parameter settings, details of which for the other system are not available to us.

Table 5.4: MAP scores for different methods of stopword evaluation in the Hindi (50 T queries)

Retrieval Model	None	Noncorpus based	Fox based	tf	idf	ntf	nidf	Term R.S.	Entropy	Agg1	Agg2
BM25	0.4429	0.4418	0.4426	0.4427	0.4427	0.4392	0.4427	0.4437	0.4445	0.4444	0.4438
TF-IDF	**0.4439**	**0.4425**	**0.4439**	**0.4438**	**0.4437**	**0.441**	**0.444**	**0.4449**	**0.4455**	**0.4455**	**0.4449**
DLH	0.4129	0.4154	0.4178	0.4194	0.4206	0.4170	0.4180	0.4112	0.4145	0.4212	0.4111
PL2	0.3448	0.3744	0.3566	0.4011	0.3544	0.4004	0.4028	0.3645	0.3649	0.405	0.3942
IFB2	0.3579	0.3490	0.3713	0.3546	0.3694	0.3557	0.3553	0.3620	0.3611	0.3554	0.3503
LM	0.3951	0.3635	0.3818	0.3677	0.4015	0.3646	0.3682	0.3944	0.3943	0.3724	0.3842
Mean	0.399	0.397	0.402	0.404	0.405	0.402	0.405	0.403	0.404	0.407	0.404
% Change		-0.5%	+0.7%	+1.25%	+1.5%	+0.7%	+1.5%	+1%	+1.25%	+2.0%	+1.25%

Table 5.5: MAP scores for different methods of stopword evaluation in the English (50 T queries)

Retrieval Model	None	Smart stopword	Noncorpus based	tf	idf	ntf	nidf	Term R.S.	Entropy	Agg1	Agg2
BM25	0.3148	0.3158	0.3163	0.3162	0.3173	0.3156	0.3149	0.3162	0.3164	0.3161	0.3161
TF-IDF	**0.316**	**0.3176**	**0.3183**	**0.3187**	**0.3187**	**0.3167**	**0.3186**	**0.3185**	**0.3189**	**0.3182**	**0.3186**
DLH	0.2833	0.2897	0.2901	0.2862	0.2892	0.2865	0.2876	0.2888	0.2888	0.2876	0.2887
PL2	0.2496	0.2607	0.2583	0.2556	0.2562	0.2535	0.2552	0.2589	0.2577	0.2566	0.2574
IFB2	0.2691	0.2781	0.2765	0.2742	0.2735	0.2724	0.2741	0.2760	0.2765	0.2746	0.276
LM	0.2621	0.2689	0.2684	0.2661	0.2679	0.2648	0.2665	0.2687	0.2684	0.2666	0.2678
Mean	0.282	0.288	0.287	0.286	0.287	0.284	0.286	0.287	0.287	0.286	0.287
% Change		+2.29%	+2.12%	+1.41%	+2.12%	+1.03%	+1.47%	+2.12%	+2.12%	+1.41%	+2.12%

To get more insights into the effect of stopwords, we perform a query-by-query analysis. Here, we consider the best MAP score obtained by a stopword generation approach. We examine Zipf's law-based (idf-based) stopword list in Bengali, Marathi, and Gujarati languages. Similarly, we examine the aggregation1 and aggregation2 stopword lists in Hindi and English. On closer observation, we find that the stopword removal improves the average precision score in 32 topics and reduces the average precision score in 13 topics in Bengali. For example, in Topic 46, 'Bill and Melinda Gates Foundation's philanthropic activities in India', stopword removal improves the average precision score by 1.73%. A per query level analysis for the Bengali is shown in Fig 5.1. We also perform query-by-query analysis in Marathi, Gujarati, Hindi and English. In these languages, stopword removal improves the average precision score in 20, 31, 25, and 27 topics and reduces the average precision score in 11, 15, 25 and 23 topics, respectively. A per query level analysis for the Marathi, Gujarati, Hindi and English are shown in Fig 5.2, 5.3, 5.4 and 5.5.

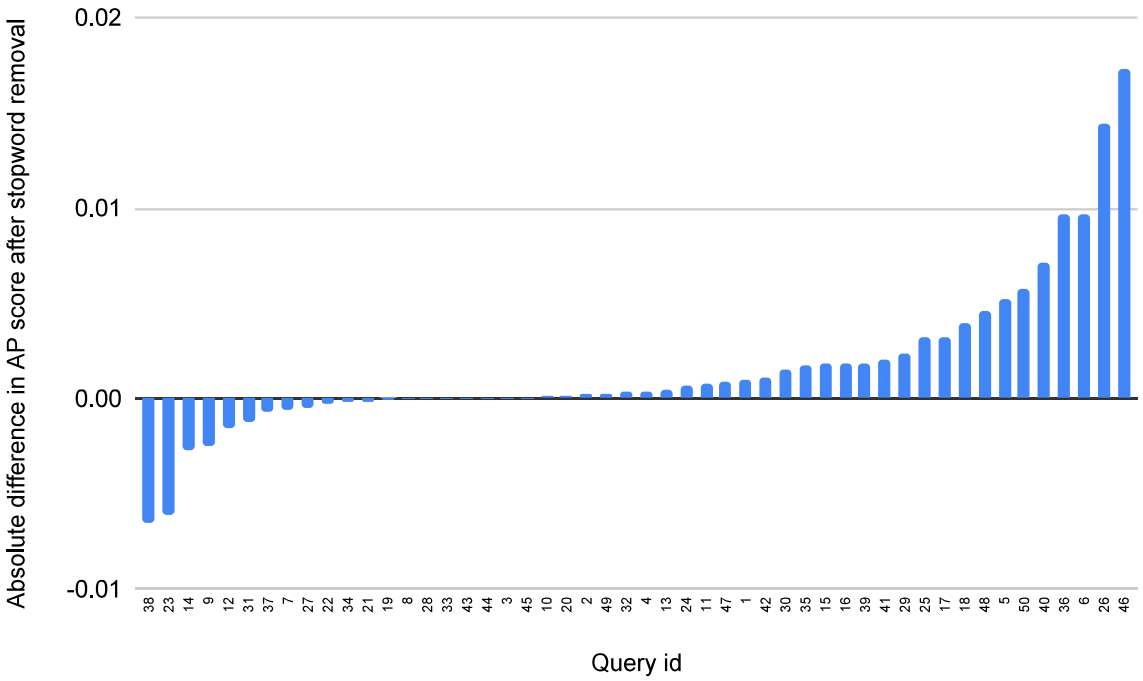

Figure 5.1: A query by query evaluation in the Bengali by BM25 model

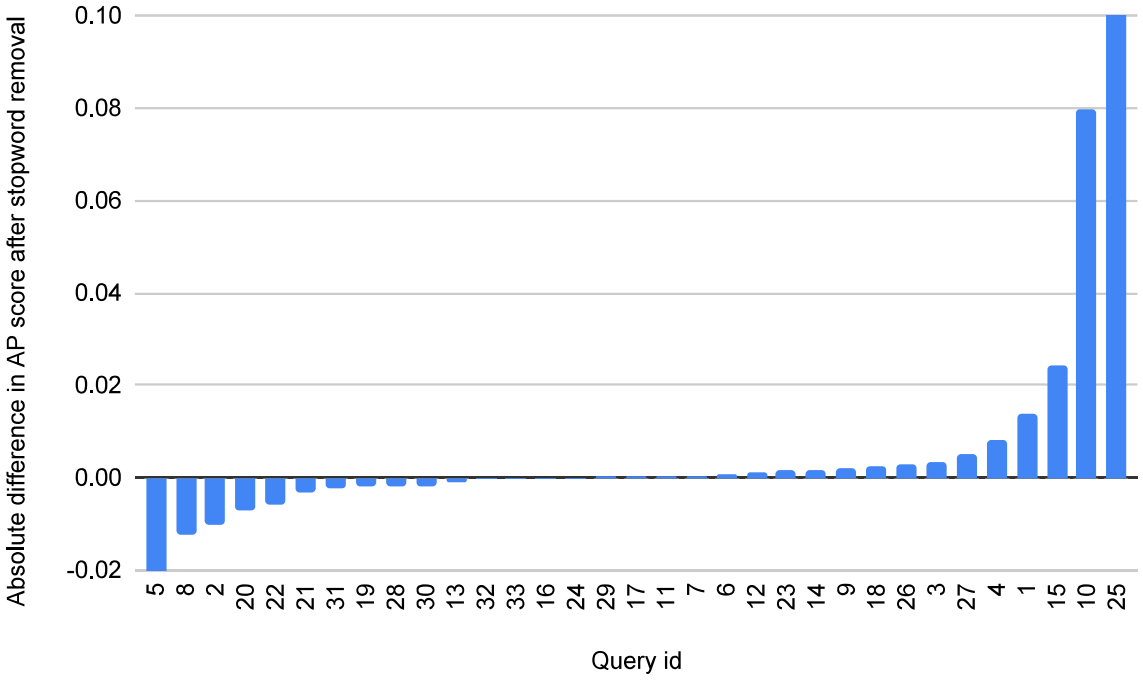

Figure 5.2: A query by query evaluation in the Marathi by TF-IDF model

5.5.2 Effect of corpus-based stopwords on retrieval

In the second set of experiments, we propose corpus-based stopword lists in different Indian languages and evaluate their effectiveness in the IR domain. Different stopword

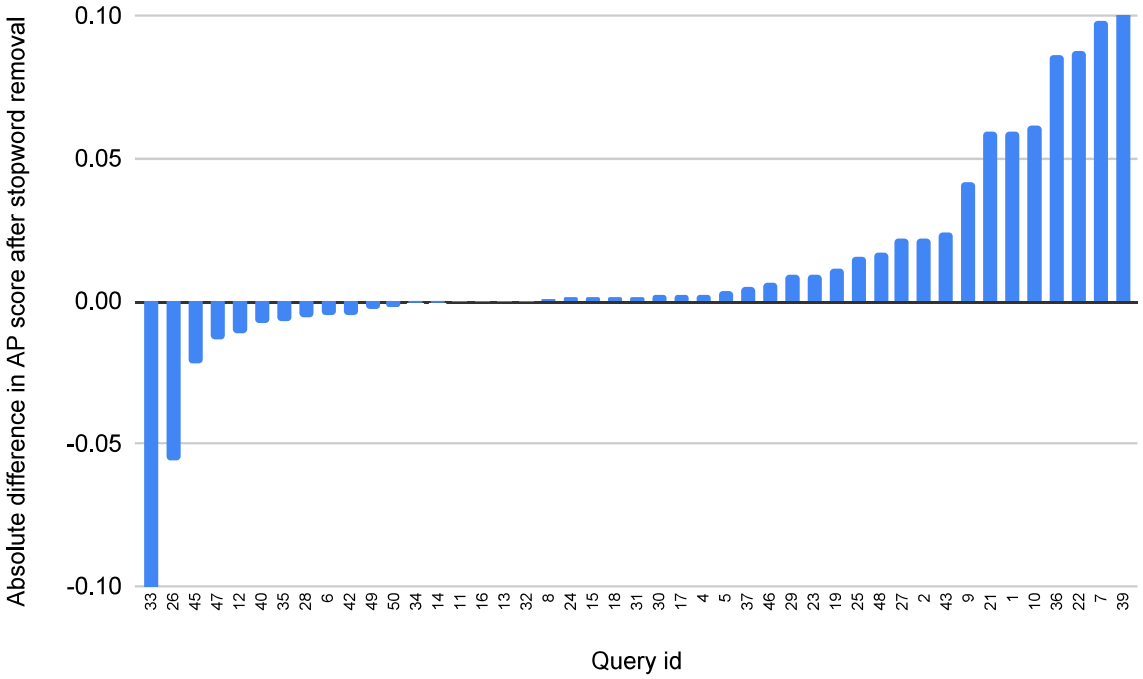

Figure 5.3: A query by query evaluation in the Gujarati by BM25 model

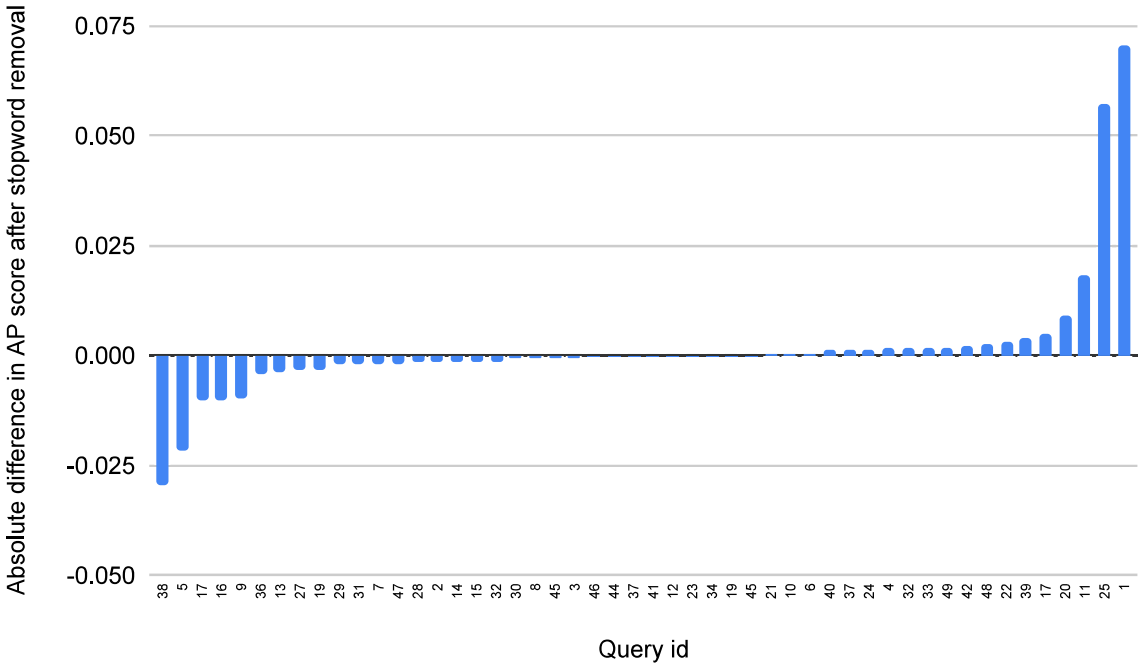

Figure 5.4: A query by query evaluation in the Hindi by TF-IDF model

lists are generated by applying the methodology described in section 5.3. The evaluation of different corpus-based stopword removal in different Indian languages is shown in

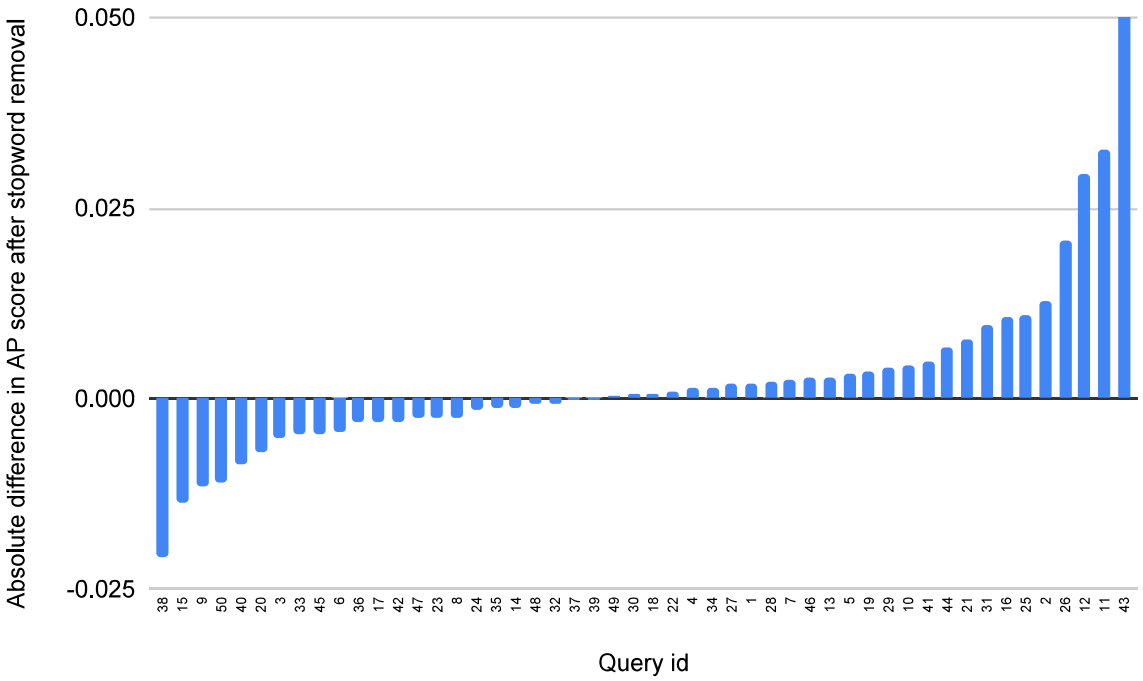

Figure 5.5: A query by query evaluation in the English by TF-IDF model

Tables 5.1, 5.2, 5.3, 5.4 and 5.5. We observe that Zipf's law-based stopword removal techniques outperform other stopword removal techniques in Indian languages IR. Among the different corpus-based stopword lists, Zipf's law-based stopword removal (specifically, the idf-based) provides the best retrieval effectiveness across the Indian languages. The idf-based stopword removal improves MAP score by more than 5% in Bengali and Gujarati. Similarly, in Marathi, Hindi and English languages, the idf-based stopword removal improves MAP score by 1.91%, 1.5% and 2.12%, respectively, compared to their baseline approaches. The term-based random sampling, aggregation2 (probability term frequency, entropy, and variance), and entropy-based stopword removal provide feeble improvements (less than 2%) in terms of MAP scores in the Indian languages considered (Bengali, Marathi, Gujarati and Hindi). However, in English, the stopword removal improves a MAP score by more than 2%. The aggregation1 method (term frequency, inverse document frequency, and entropy-based) stopword removal improves a MAP score by more than 4% in Bengali and Gujarati. In contrast, in Marathi, Hindi and English, stopword removal improves a MAP score by less than 2%. In summary, we find that a smaller length of a corpus-based stopword list improves retrieval effectiveness in Indian languages IR.

5.5.3 Length of stopword lists

In the third set of experiments, we evaluate the minimum length of the stopword list that can be used to improve the effectiveness of an IR system. Stopword removal improves retrieval effectiveness in different languages, but we could not find any recommendations on the length of stopword lists in the literature. A short stopword list accommodates only a few less-important words, leading to increased length of the dictionary and total size of postings. On the other hand, a long stopword list can affect the effectiveness of an IR system as some potentially important terms are missed. We thus heuristically experiment with different lengths of stopword lists, viz. at length $50, 100, \ldots, 400$ and evaluate MAP at these points. The best MAP score among the scores obtained at different lengths of stopword lists in different Indian languages is shown in Table 5.6. A short stopword list provides the best retrieval effectiveness in Bengali, Hindi and English. However, a moderate-length stopword list performs best in Marathi and Gujarati. No single particular stopword length can be suggested for different Indian languages. Hence, different Indian languages require different lengths of stopword lists to improve the effectiveness of a system. We find that morphologically rich Indian languages require a larger stopword list length than English. The stopword length varies from one language to another. Languages like Marathi and Gujarati require larger stopword lengths than Bengali and Hindi. In summary, we find that a smaller length of a corpus-based stopword list outperforms that of a non-corpus-based list in Indian languages IR.

Table 5.6: Length of stopword list used for evaluation in Indian languages

	Non corpus-based	Non corpus-based	tf	idf	ntf	nidf	Term R.S.	Entropy	Agg1	Agg2
Bengali	398	119	150	150	150	150	150	150	300	200
Marathi	216	99	200	250	200	250	150	150	300	150
Gujarati	210	1240	200	250	200	250	200	200	300	200
Hindi	225	163	150	200	150	150	150	150	200	150
English	571	733	150	150	150	150	150	150	150	150

5.6 Summary

Stopword removal is an essential pre-processing step in IR. The above experiments show that stopword removal improves retrieval effectiveness in different Indian languages IR. On closer observation, we find that the non-corpus-based stopword lists provide poorer effectiveness compared to the corpus-based ones in the Indian language IR. However, both (non-corpus-based and corpus-based) provide similar MAP scores in English. Moreover, we observe that a smaller corpus-based stopword list offers a better MAP score than a larger length of non-corpus-based stopword lists for all four Indian languages studied. Among the different corpus-based stopword lists, Zipf's law-based stopword list (specifically, the idf-based) provides the best MAP score across the Indian languages. Other techniques (term-based random sampling, entropy-based, and aggregation2-based stopword list) provide poor retrieval effectiveness. However, the aggregation1-based stopword list provides a comparable MAP score to Zipf-based ones in all Indian languages. For English, the aggregation2-based stopword list offers the best retrieval effectiveness. When evaluating different retrieval models, the probabilistic retrieval models (BM25 and TF-IDF) exhibit the best retrieval effectiveness; however, the PL2 model performs poorly. We also observe that the shorter length of the corpus-based stopword list provides comparable retrieval effectiveness with relatively longer non-corpus-based stopword lists across a set of Indian languages. Hence, we argue that a corpus-based stopword list is a good technique for Indian language retrieval tasks.