Chapter 4

Effect of stopwords in Indian language IR

4.1 Introduction

We explore and evaluate the effect of stopwords on retrieval effectiveness in different Indian languages such as Marathi, Bengali, Gujarati, Hindi and Sanskrit. The issue was investigated from two viewpoints. Is there any impact of non-corpus-based stopword removal on Indian languages (if yes, to what extent)? Is there any relationship between stopwords and average document length from a retrieval perspective? It is observed that stopword removal generally improves the mean average precision (MAP) score compared to when it is not done. We also study the effect of stopwords on retrieval effectiveness over document length. The impact of stopwords is generally low in short documents compared to their long counterparts across the five Indian languages.

4.2 Problem Formulation

Stopwords are the most frequently used words that carry little or zero information from the novelty aspect; however, they are very much needed to connect different important words and form a meaningful sentence. We downloaded the stopword lists from GitHub [1], that are language-specific but independent of any document collection of the language in hand (non-corpus-based). The stopword list primarily comprises pronouns, conjunctions,

[1] https://github.com/cse-iitbhu/Non-corpus-based-stopword-list

adverbs, prepositions, and interjections from a linguistics point of view. Sometimes, some verbs and adverbs are also treated as stopwords.

In this chapter, we seek answers to the following research questions (RQs).

1. *[RQ 1.1] At the gross level, is there any impact of non-corpus-based stopword removal on Indian languages IR?*

   Here, we see the effect of stopwords on retrieval effectiveness in different Indian languages by not removing the stopwords (not using any stopword list during indexing) and then removing them.

2. *[RQ 1.2] Do stopwords have any relationship with average document length from the perspective of retrieval effectiveness? In other words, how does retrieval effectiveness change with the number of stopwords and average document length?*

4.3 Experimental Setup

We study the effect of stopwords in the traditional IR framework. We use Terrier [2] retrieval system to index and retrieve the document collections in its entirety.

However, to conduct experiments related to **RQ 1.2**, we divide each corpus into short and long documents so that each part contains almost equal number of documents. In each subpart, we evaluate the effect of stopwords on retrieval effectiveness in different languages.

The queries used are in TREC format, each representing an information need having three subparts: a terse set of words called a title (T), a brief description (D) of the information need and a detailed narration (N) of the need explaining what qualifies a document to be relevant and what does not. We use each topic's three parts (TDN) from the topic set.

All experiments are conducted on a laptop with a core i3 processor and 8 GB RAM.

[2] http://terrier.org/

4.4 Evaluation

To address the RQs discussed above, we experimented with and evaluated different Indian languages in the following way.

4.4.1 Effect of stopword removal on retrieval

In the first set of experiments, we investigate the impact of non-corpus-based stopwords on retrieval effectiveness in different Indian languages. We claim that stopword removal changes the retrieval effectiveness: it can increase or decrease the retrieval effectiveness. For Marathi, the MAP, R-prec and P@10 evaluated without stopword removal and with stopword removal are shown in Table 4.1. It is observed above that in most retrieval models, the MAP and P@10 scores increase after stopword removal. However, the R-prec scores decrease after stopword removal. We conduct similar experiments for the other four languages: Bengali, Gujarati, Hindi and Sanskrit, as shown in Table 4.2, 4.3, 4.4 and 4.5, respectively. In all the tables, the best retrieval score by a given retrieval model is shown in italics.

Table 4.1: MAP, R-prec and P@10 in Marathi (39 TDN queries)

	Without stopword removal			With stopword removal		
Retrieval model	MAP	R-prec	P@10	MAP	R-prec	P@10
BM25	0.3232	0.32	*0.3769*	*0.3258*	0.3159	*0.3744*
TF-IDF	*0.324*	*0.321*	0.3718	0.3233	*0.3195*	0.3728
In_expC2	0.2603	0.2559	0.3205	0.2638	0.2561	0.3282
In_expB2	0.2803	0.278	0.3462	0.2816	0.2738	0.3487
InL2	0.2802	0.2761	0.3487	0.2824	0.2718	0.3513
Hiem_LM	0.2571	0.2544	0.3256	0.2578	0.2492	0.3231
Mean	0.2875	0.2842	0.3483	0.2891	0.281	0.3497
% Change				+.56%	−1.1%	+.42%

For most retrieval models, the MAP, R-prec and P@10 scores increase after stopword removal. We also observe that in Marathi and Bengali languages, the P@10 values are pretty high, which signifies that more relevant documents are retrieved at early ranks. In Gujarati, the MAP, R-prec and P@10 values are similar. In Sanskrit, the P@10 values are pretty low. There are two reasons for this: a) the number of relevant documents in Sanskrit is less compared to that in other languages, as the Sanskrit dataset is the

Table 4.2: MAP, R-prec and *P*@10 in Bengali (50 TDN queries)

	Without stopword removal			With stopword removal		
Retrieval model	MAP	R-prec	*P*@10	MAP	R-prec	*P*@10
BM25	0.2668	*0.3047*	*0.452*	*0.271*	*0.3112*	0.448
TF-IDF	*0.267*	0.302	0.448	0.2706	0.3093	*0.45*
In_expC2	0.2188	0.2638	0.398	0.2206	0.2614	0.396
In_expB2	0.2385	0.2831	0.412	0.2388	0.2832	0.414
InL2	0.2434	0.2865	0.43	0.239	0.2873	0.432
Hiem_LM	0.2057	0.2518	0.39	0.2109	0.2558	0.392
Mean	0.24	0.282	0.4217	0.2418	0.2847	0.422
% Change				+.74%	+.96%	+0.07%

Table 4.3: MAP, R-prec and *P*@10 in Gujarati (46 TDN queries)

	Without stopword removal			With stopword removal		
Retrieval model	MAP	R-prec	*P*@10	MAP	R-prec	*P*@10
BM25	0.302	*0.3134*	*0.3*	*0.3132*	*0.3252*	*0.3087*
TF-IDF	0.2957	0.3059	0.2935	0.3107	0.3225	0.3
In_expC2	0.2892	0.3022	0.2957	0.2958	0.3059	0.2978
In_expB2	*0.3068*	0.3091	0.2957	0.3086	0.3164	0.3022
InL2	0.3013	0.3001	0.287	0.2999	0.297	0.3065
Hiem_LM	0.2079	0.2318	0.237	0.2201	0.2425	0.2391
Mean	0.2838	0.2938	0.2848	0.2914	0.3016	0.2924
% Change				+2.66%	+2.67%	+2.65%

Table 4.4: MAP, R-prec and *P*@10 in Hindi (50 TDN queries)

	Without stopword removal			With stopword removal		
Retrieval Model	MAP	R-prec	P@10	MAP	R-prec	P@10
BM25	0.5044	0.2739	0.4787	0.5185	0.2745	0.483
TF-IDF	0.4979	0.2711	0.483	0.5095	0.2718	0.4787
In_expC2	0.4168	0.2451	0.4319	0.425	0.2472	0.4468
In_expB2	0.4508	0.2564	0.4511	0.4636	0.2593	0.4617
InL2	0.4454	0.2511	0.4468	0.445	0.2513	0.466
Hiem_LM	0.2627	0.1898	0.3447	0.2881	0.1962	0.366
Mean	0.4296	0.2479	0.4393	0.4416	0.25	0.4503
% Change				+2.79%	+0.84%	+2.5%

smallest in size, and b) the relevant documents are retrieved at later ranks. We observe that the stopword removal provides a slight improvement in MAP score in different Indian languages.

Table 4.5: MAP, R-prec and P@10 without and with stopword removal in Sanskrit (50 TDN queries)

	Without stopword removal			With stopword removal		
Retrieval model	MAP	R-prec	P@10	MAP	R-prec	P@10
BM25	0.405	0.3807	0.218	0.4209	0.4016	0.224
tf_idf	0.4023	0.376	0.212	0.421	0.3931	0.224
In_expC2	0.4077	*0.3988*	0.222	0.4205	*0.4087*	0.23
In_expB2	*0.4091*	0.3892	*0.226*	*0.4232*	0.4037	*0.234*
InL2	0.3913	0.3697	0.212	0.403	0.3756	0.214
Hiem_LM	0.3581	0.3505	0.174	0.3877	0.3772	0.196
Mean	0.3956	0.3775	0.2107	0.4127	0.3933	0.2203
% Change				+4.33%	+4.19%	+4.58%

To get more insights, we also perform a query-by-query analysis. Here, we show it for the BM25 retrieval model in Marathi and Bengali, In_expB2 model for Gujarati and Sanskrit and language model for Hindi, all being the best-performing models for the respective languages. On closer observation, in Marathi, stopword removal improves the average precision scores for 29 topics while reducing for seven. The average precision score of each query for Marathi is shown in Figure 4.1. For example, in Topic #12, सियाचिन सभोवतीच्या सैन्याच्या स्थानाविषयी मनमोहन सिंह आणि परवेझ मुशर्रफ ह्यांच्यामधील चर्चा (Manmohan Singh, Pervez Musharraf discuss troop position around Siachen), stopword removal improves average precision score compared to when it is not done by 4.22% (highest). A similar observation is found in Topic #25 मोनिका बेदी आणि बनावटी पारपत्र खटला (Monica Bedi and fake passport lawsuit) (improvement of 1.27%). Similarly, the changes in average precision score due to stopword removal at per-query-level for Bengali, Gujarati, Hindi and Sanskrit are shown in Figures 4.2, 4.3, 4.4 and 4.5 respectively. As shown earlier, in Bengali, Gujarati, Hindi and Sanskrit, stopword removal improves the average precision score in 41, 36, 39 and 36 topics while reducing the scores in 9, 10, 6 and 6 topics, respectively. In summary, stopword removal improves AP scores for most of the queries, though reducing the score in a small number of cases. This fact was not that pronounced in Tables 4.1–4.5, but it is certainly an important revelation. For Bengali, Gujarati and Sanskrit, in particular, the improvement is noticeable ($>= 0.05$) for a number of queries.

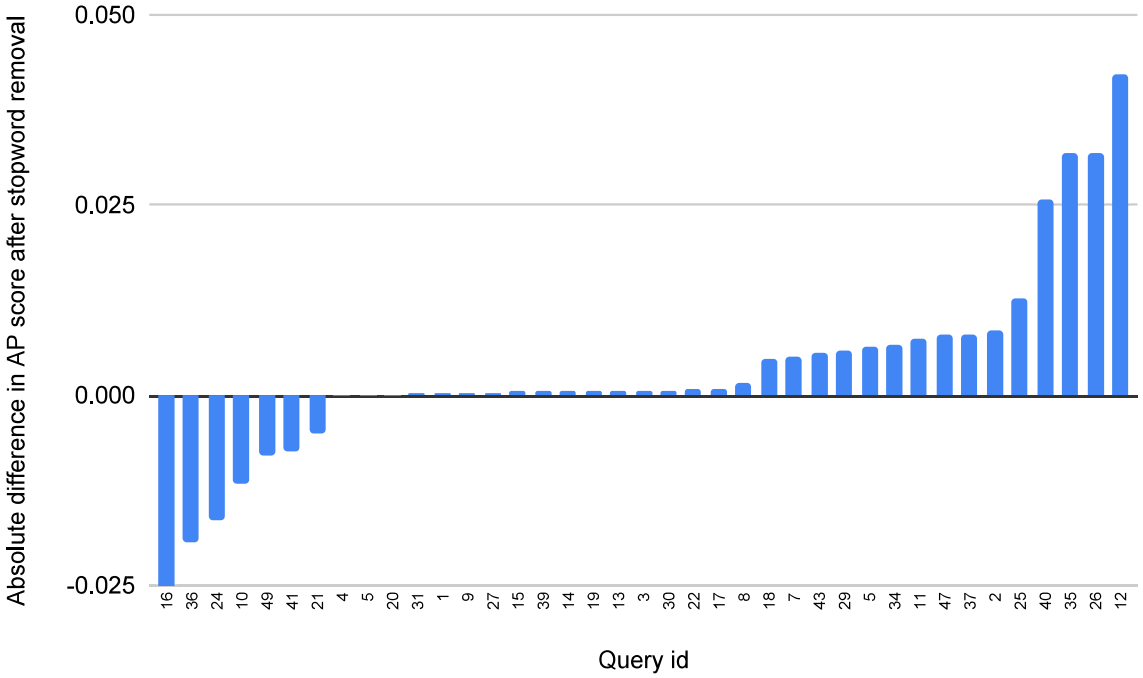

Figure 4.1: A query-by-query evaluation in Marathi by BM25 model

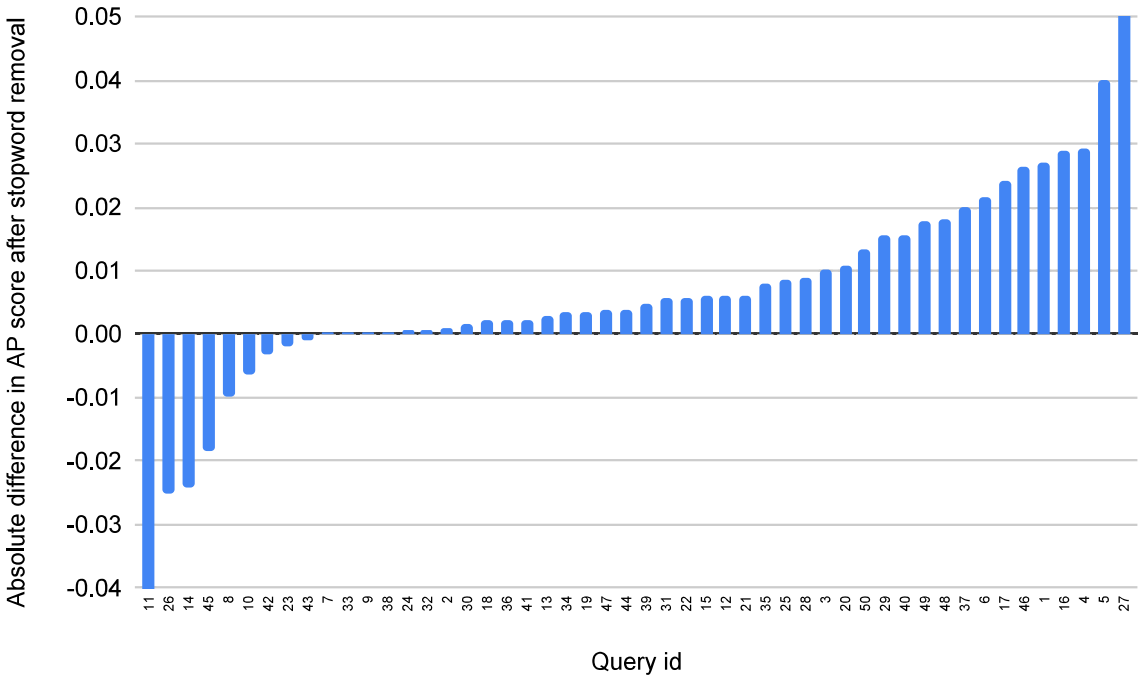

Figure 4.2: A query-by-query evaluation in Bengali by BM25 model

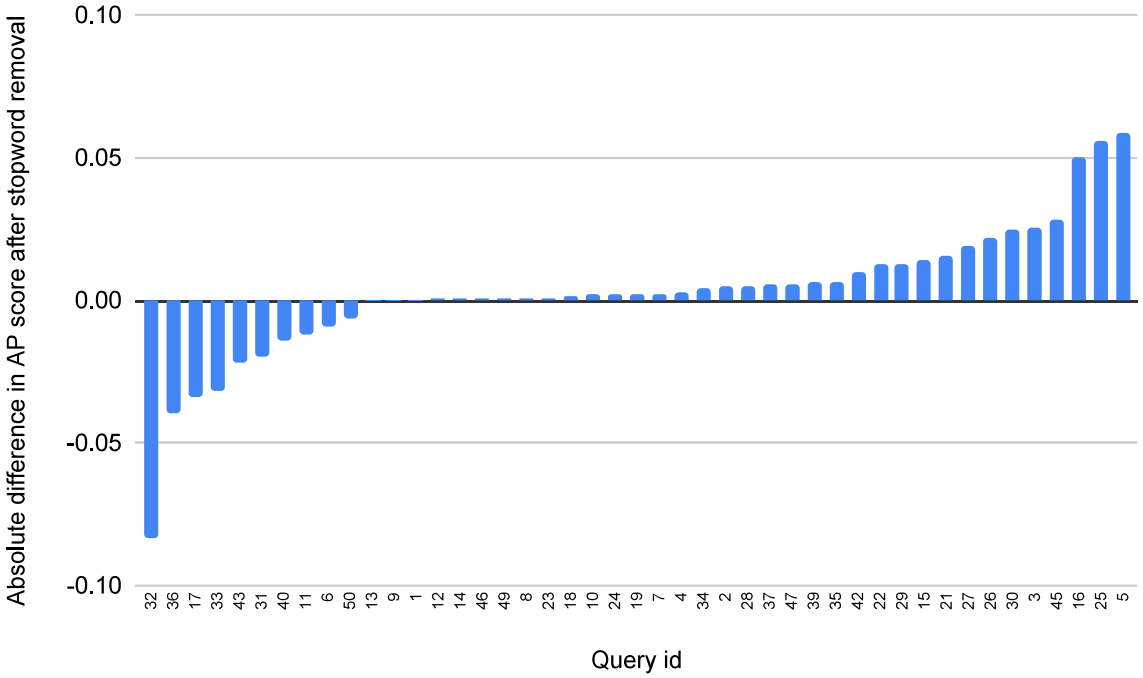

Figure 4.3: A query-by-query evaluation in Gujarati by In_expB2 model

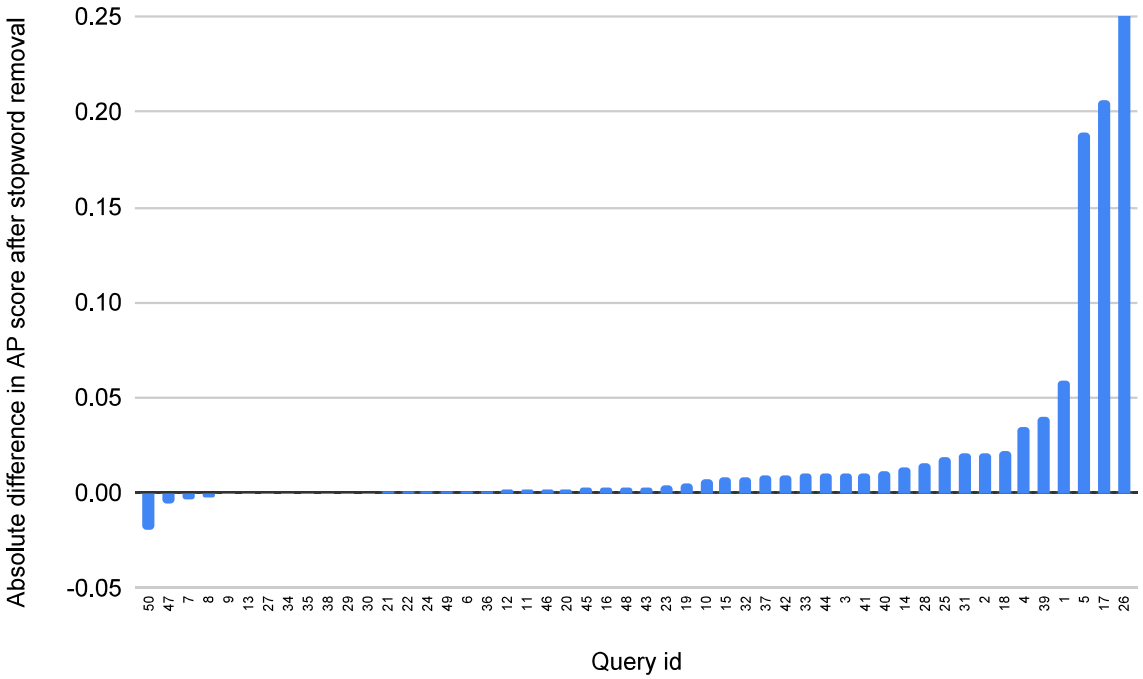

Figure 4.4: A query-by-query evaluation in Hindi by Hiem_LM model

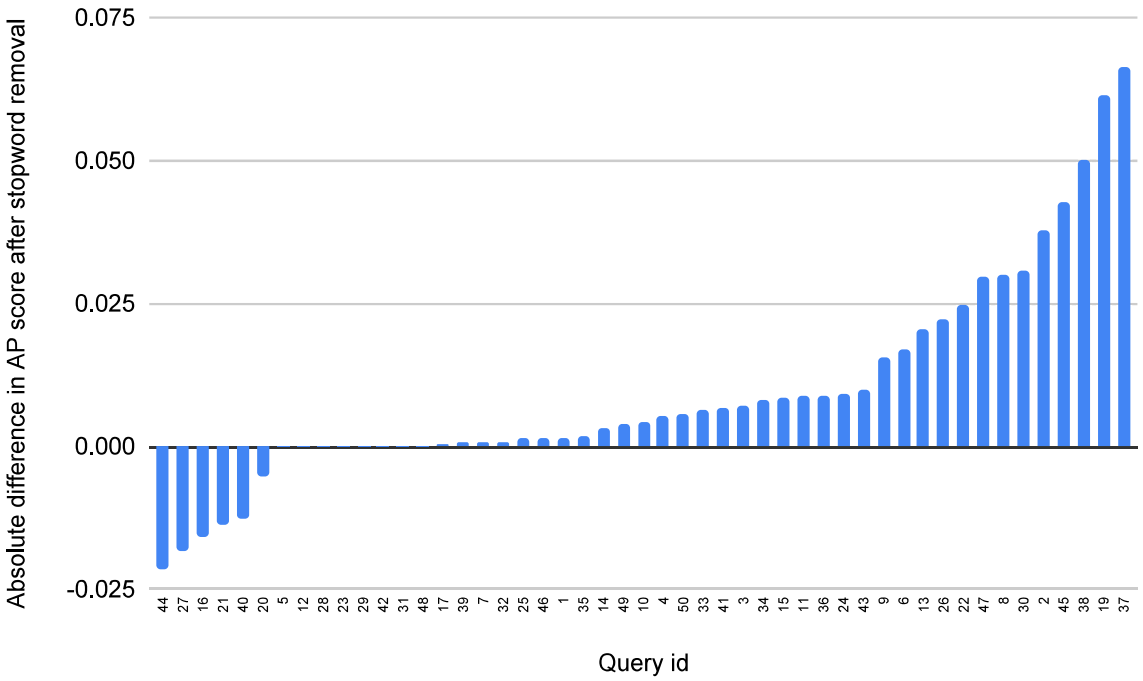

Figure 4.5: A query-by-query evaluation in Sanskrit by In_expB2 model

4.4.2 Interaction between stopwords and document length

The second set of experiments is conducted to see the effect of stopword removal on retrieval effectiveness with changes in document length. In a long document, diverse terms occur, and each term potentially occurs with high frequency. Any retrieval model prefers a long document over a short one because the probability of a query term occurring in a long document is higher. To overcome this retrieval bias towards longer documents, modern IR systems use document length normalization [130]. For a heuristic study to see the effect of stopword removal on short vs long documents, we divide each corpus into two parts. We consider short documents of smaller file sizes over larger file sizes based on suitable thresholds, ensuring an almost equal number of documents in each part. In each set, we evaluate the effect of stopwords on retrieval effectiveness as shown in Tables 4.6–4.10. In all the languages, most retrieval models give a comparable MAP score in short and long documents. However, the language model gives poor MAP scores in all languages.

Table 4.6: MAP scores of the effect of stopword in short vs. long docs in Marathi

	Short document length		Long document length	
Retrieval Model	None	Stopword removal	None	Stopword removal
BM25	0.1314	0.1321	0.2276	0.2296
TF-IDF	0.1298	0.1293	0.2249	0.2257
In_expC2	0.1109	0.1121	0.2224	0.2248
In_expB2	0.1158	0.1178	0.2254	0.2278
InL2	0.1201	0.1229	0.2232	0.225
Hiem_LM	0.1094	0.1119	0.1913	0.199
Mean	0.1196	0.121	0.2191	0.222
% Change		+1.21%		+1.3%

Table 4.7: MAP scores of the effect of stopword in short vs. long docs in Bengali

	Short document length		Long document length	
Retrieval Model	None	Stopword removal	None	Stopword removal
BM25	0.0843	0.0843	0.199	0.2033
TF-IDF	0.0834	0.0841	0.1978	0.2034
In_expC2	0.0686	0.0691	0.1739	0.1753
In_expB2	0.075	0.0745	0.1848	0.1871
InL2	0.0758	0.0764	0.1892	0.1896
Hiem_LM	0.0683	0.0706	0.1516	0.1583
Mean	0.0759	0.0765	0.1827	0.1862
% Change		+.79%		+2.07%

Table 4.8: MAP scores of the effect of stopword in short vs. long docs in Gujarati

	Short document length		Long document length	
Retrieval Model	None	Stopword removal	None	Stopword removal
BM25	0.1538	0.1537	0.1837	0.1921
TF-IDF	0.148	0.1521	0.179	0.1923
In_expC2	0.1366	0.1369	0.166	0.1736
In_expB2	0.1462	0.1463	0.1769	0.1849
InL2	0.1424	0.1434	0.1845	0.1909
Hiem_LM	0.1186	0.1277	0.1202	0.1285
Mean	0.1409	0.1433	0.1683	0.1767
% Change		+1.71%		+5%

4.5 Discussion

In the first set of experiments (shown in Tables 4.1–4.5), we observe that the non-corpus-based stopword removal improves retrieval effectiveness and gives comparable MAP scores

Table 4.9: MAP scores of the effect of stopwords in short vs. long docs in Hindi

	Short document length		Long document length	
Retrieval Model	None	Stopword removal	None	Stopword removal
BM25	0.156	0.1569	0.4407	0.4583
TF-IDF	0.1505	0.1522	0.4456	0.464
In_expC2	0.1373	0.1432	0.4038	0.4151
In_expB2	0.1446	0.1496	0.4256	0.4381
InL2	0.1408	0.1442	0.3992	0.4286
Hiem_LM	0.1051	0.1134	0.2051	0.2378
Mean	0.139	0.1432	0.3866	0.4069
% Change		+3.02%		+5.25%

Table 4.10: MAP scores of the effect of stopwords in short vs. long documents in Sanskrit

	Short document length		Long document length	
Retrieval Model	None	Stopword removal	None	Stopword removal
BM25	0.1971	0.2102	0.2355	0.2536
TF-IDF	0.2002	0.2122	0.2320	0.2514
In_expC2	0.1939	0.2035	0.2337	0.2421
In_expB2	0.1942	0.205	0.2350	0.2485
InL2	0.1959	0.2018	0.2250	0.2371
Hiem_LM	0.1589	0.1738	0.2194	0.234
Mean	0.19	0.2017	0.23	0.2433
% Change		+5.83%		+5.79%

when non-corpus-based stopwords lists are used [37]. On closer observation, we also find that the effect of stopwords varies from one language to another. In Marathi and Bengali, the effect of stopword removal on retrieval effectiveness is relatively low (less than 1%). However, in Gujarati, Hindi and Sanskrit, it is higher (more than 2%). In Marathi, Bengali and Hindi, the BM25 and TF-IDF models give better MAP scores than DFR-based retrieval models. Different retrieval models provide similar effectiveness in Gujarati and Sanskrit. Among the different retrieval models, the effectiveness of the language model is poor in different Indian languages. In summary, we find that the non-corpus-based stopword list improves effectiveness in document retrieval. Still, the effectiveness is relatively low compared to other corpus-based stopword removals in Indian and European languages [37]. A corpus-based stopword list may improve effectiveness in South-Asian languages like other European languages, but this is yet to be ascertained.

In the second set of experiments (shown in Tables 4.6–4.10), we observe that the

effect of stopwords is relatively low in short documents compared to long ones. This can be explained by the fact that a long document has many stopwords, and each such stopword is also likely to have a higher frequency there than in a short one. Hence, removing stopwords causes more improvement in longer documents than their shorter counterparts. In Sanskrit, the effect of stopword removal is quite similar in both short and long documents. This can be attributed to the fact that the size of the Sanskrit collection is small, with slight variations in length. In Bengali, Gujarati and Hindi, the effect of stopword removal is less noticeable in short documents, whereas in long ones, it is noticeable. In Marathi, Bengali and Hindi, the differences in the MAP scores between short and long documents are pretty high. However, corresponding differences in the MAP scores are comparatively low in Gujarati and Sanskrit. The principal reason is that Bengali, Hindi and Marathi collections contain more relevant long documents compared to short ones. However, in Gujarati and Sanskrit, the collection consists of nearly equal number of relevant documents in both long and shorts. We also observe that the different retrieval models prefer long documents at early ranks to short ones, demonstrating that the evaluated retrieval models still preserve bias towards long documents.

4.6 Summary

Stopword removal is an effective pre-processing step in the IR domain. In the aforementioned experiments, we observed that removing stopwords improved MAP scores compared to without stopword removal. In Marathi, Bengali and Hindi, the BM25 and TF-IDF models give better MAP scores than the DFR-based retrieval models. Different retrieval models provide similar effectiveness in Gujarati and Sanskrit. However, the language model is not promising for the Indian languages considered. We also found that the effect of stopword removal on retrieval effectiveness is relatively low in short documents compared with long ones. In Marathi, Bengali and Hindi, the differences in MAP scores between short and long documents are pretty high, whereas, in Gujarati and Sanskrit, they are comparatively low. In all languages, most retrieval models give similar MAP scores in short and long documents. In Sanskrit, the collection size is very small compared to the other four languages. Hence, the difference in length between short and long documents is minimal.