

Experiments in N -gram based indexing and retrieval in Marathi

Ashish Almeida, Pushpak Bhattacharyya
IIT Bombay
ashishfa@gmail.com, pushpakbh@gmail.com

Abstract

For European languages, n -gram has proved to be the cost effective alternative to morphological processing during indexing task and it has been studied and analyzed extensively using CLEF data. We adapted this work for our experiments on n -grams in Marathi language. Our experiments indicate that 4-gram produces the best results among n -grams of different lengths. Also we find that n -gram based retrieval provides improvements over mere word based retrieval for Marathi which is a morphologically rich language. We obtain the MAP (Mean Average Precision) score of 35.79% for n -gram based indexing against baseline MAP score of 23.94%.

1. Introduction

Lexical analysis of documents such as morphological processing, stop-word elimination *etc.* is studied by Ashish Almeida and Pushpak Bhattacharyya [7] in detail for Marathi. The major drawback of such systems is that it depend on the accuracy of morphology processing. In contrast, the n -gram based system is easy to develop and since it is language neutral, there is no need of lexical resources. In this report, we discuss experiments using n -grams to improve performance of information retrieval system and compare it with the baseline. The setting is Marathi monolingual IR.

2. Related work

In the paper "Character n -gram tokenization for European language text retrieval" by Paul McNamee and James Mayfield [1], use of overlapping character n -grams based indexing in ten languages is discussed. Paul McNamee's thesis [2] "Textual representations for corpus-based bilingual retrieval" elaborately discusses retrieval models.

Paul McNamee discusses his experiments using the IR system called HAIRCUT [1]. He reports that out of n -grams 4-grams produce the best results for most of the European languages. His report is based on experiments conducted on the test-collections of Cross-Language Evaluation Forum (CLEF) [URL: <http://www.clef-campaign.org/>]. He analyses the retrieval accuracy in eight European languages using n -gram based indexing. He also deliberates over advantages of n -gram over other indexing methods. His thesis explains how the performance of the IR system is affected when large words are broken into small parts and how the word spanning n -gram captures associations in the text.

In the paper "s-grams: defining generalized n -grams for information retrieval", Anni Jarvelin, Antti Jarvelin and Kalervo Jarvelin [8] discuss s -gram which is generic version of n -gram. The paper describes how s -grams help in cross lingual and monolingual information retrieval. It formalises the notion of n -gram and s -gram with definitions and describes various similarity measures to compare them.

3. Dataset

For our experiment, we used dataset of FIRE 2008 [URL: <http://www.isical.ac.in/~fire/>] for Marathi. The corpus contains 99,275 documents from news domain. These news articles, dated

from April 2004 to September 2007, are extracted from websites of two widely read newspapers, Maharashtra Times [URL: <http://maharashtratimes.indiatimes.com>] and Sakal [URL: <http://www.esakal.com>]. The activity of corpus creation was carried out under purview of “Cross Language Information Access” consortia project coordinated by IIT Bombay [URL: <http://www.clia.iitb.ac.in>].

Document format: It is adapted from the TREC [URL: <http://trec.nist.gov>] document format. Each document is stored physically in a separate file and it has 2 fields, DOCNO and TEXT. DOCNO is a unique identifier assigned to each document. TEXT field contains entire news article in plain text. The example is shown below.

```
<DOC>
<DOCNO>MaharashtraC06E811B.txt</DOCNO>
<TEXT>
घटक चाचणीत सरकार अनुत्तीर्ण
सरकारने आज शंभर दिवस पूर्ण केले आहेत. ठराविक काळानंतर होणार्या घटक चाचणी परीक्षांत आपली तयारी
आजमावण्याची संधी विद्यार्थ्यांना मिळते. सरकारच्या कामगिरीचे मूल्यमापनही त्याच धर्तीवर करावे, या हेतूने हे
प्रगतिपुस्तक मांडले आहे.
....
</TEXT>
</DOC>
```

Queries: Out of 100 queries, relevance judgement was available for a set of 50 queries (query number 26 to 75) from FIRE 2008 workshop. These 50 queries were used as the training set. The test set consists of query numbers 76 to 125. The queries were chosen such that they cover all major events from Indian media for the period 2004-2007.

Relevance judgements: For query number 76 to query number 125, the pool was created by combining the runs submitted to FIRE 2010 workshop and judged by human evaluators. Since the same set of queries is created for different Indian languages, it is possible that given a query, corpus in a particular language might not have any relevant document. In Marathi, for 11 queries we did not have any relevant document.

4. Why *N*-grams?

According to Heap's law [4], dictionary size increases with increase in number of the documents while indexing. But, if *n*-grams are used, the size is restricted by the number of alphabets rather than number of words. Recently it was also found that large size *n*-gram helps improve the accuracy as it tends to become unique [3]. Since *n*-grams split words at the character level, they are resilient to simple spelling mistakes or spelling variations. The disadvantage of *n*-grams is that *n*-gram based indexes occupy large space where *n* is large. But with the advent of cheap secondary storage devices, this is not a major concern.

In our experiment reported in [7], using morphological analyser for normalising the words while creating index, we came across some of its limitations. It could not process words which are not present in its lexicon. Also it could not process any unknown suffix. This poses a great problem in the information retrieval task such as this, as it contains news domain corpus which often includes new words, phrases and non-Marathi words. Also errors introduced by the morphological analyser reduce the performance of the system. In contrast to morphological processing, the *n*-gram approach does not need any language specific resource.

5. Implementation

In our basic *n*-gram experiment, we use fixed length *n*-grams in indexing and retrieval. The text article is separated into sentences. Apart from sentences, entities such as title, heading, address, byline are also treated as sentences. Each sentence is prefixed and suffixed by ' _ '(underscore). It is

also used to represent a space character.

Thus the sentence "या फुटबॉलपटुंना प्रसिद्धीची गरज आहे." (These football players deserves fame) becomes “_या_फुटबॉलपटुंना_प्रसिद्धीची_गरज_आहे_”. This produces 4-grams such as _या_, या_फ, ा_फु, _फुट, फुटब, ुटबॉ, ... , गरज_, रज_आ, ज_आह, _आहे, आहे_. Note that the length of word is measured in terms of Unicode characters which represent it. Thus the word प्रसिद्धी has nine characters (प+्+र+स+ि+द+्+ध+ी). The successive n -grams span word boundaries.

Experiment

We used the open source information retrieval system named Terrier [5] which is developed in University of Glasgow. It implements indexing and various retrieval models in a modular fashion. Thus, it is easy to customize and is well suited for academic research.

Though terrier 2 is Unicode compatible it is not able to handle Marathi in Unicode. We changed few lines of code to make it work on Marathi. The changes include using Unicode compatible functions and disabling the default actions such as changing case in the class hierarchy. For our experiments, we extend Terrier 2 implementation to index documents with n -grams and also convert the query to n -grams. We created the indexes for $n=2$ to 8. Table 1 indicates the size of all index data structures where n varies from 2 to 8. We submitted three runs for FIRE 2010 workshop, which are indicated with * mark in the following tables.

N in n -gram	Index data structures (MB)
2	87
3	179
4	685
5	627
6	1124
7	1872
8	2834

Table 1: Size of index data structures for different n -grams

Evaluation: MAP (mean average precision) is used as the measure of performance. It is a single value measure based on the harmonic mean of precision and recall for a given set of queries.

Retrieval algorithms: For every index, we tested the system using four different retrieval models. These models are available in terrier 2 as shown in Table 2, along with default parameters they take.

Retrieval model	parameters
TF-IDF	-
BM25	b=0.75
DFR_BM25	c=1
In_expC2	c=1

Table 2: Retrieval models used in the experiments

Experiments

1. Baseline: In this experiment, we used the terrier system modified to handle Marathi in Unicode. We did not use any pre-processing such as stop-word removal or stemming. *i.e.*, we indexed all FIRE 2010 documents with word as indexing unit. The results are mentioned in Table 3.

Retrival model	MAP
TF-IDF	0.2384
BM25	0.2390
DFR_BM25	0.2394
In_expC2	0.2371

Table 3: Mean Average Precision for word based indexing

2. Using basic n -grams: In this experiment, index is built using n -gram as explained above. N -grams are generated from a stream of characters from which all punctuation marks were removed. Four different models are used to retrieve the ranked documents list. The results are summarized in table 4. MAP values given are based on title + description + narration (TDN).

n -gram	Model	MAP (TDN)
2	BM25b0.75	0.1581
	DFR_BM25c1.0	0.1572
	TF_IDF	0.1530
	In_expC2c1.0	0.1548
3	BM25b0.75	0.3356
	DFR_BM25c1.0	0.3356
	TF_IDF	0.3284
	In_expC2c1.0	0.3304
4	BM25b0.75	0.3543
	DFR_BM25c1.0	0.3563
	TF_IDF	0.3508
	In_expC2c1.0	0.3547
5	BM25b0.75	0.3370
	DFR_BM25c1.0	0.3385
	TF_IDF	0.3354
	In_expC2c1.0	0.3412
6	BM25b0.75	0.2987
	DFR_BM25c1.0	0.2994
	TF_IDF	0.2982
	In_expC2c1.0	0.3094
7	BM25b0.75	0.2606
	DFR_BM25c1.0	0.2604
	TF_IDF	0.2605
	In_expC2c1.0	0.2636
8	BM25b0.75	0.2239
	DFR_BM25c1.0	0.2250
	TF_IDF	0.2245
	In_expC2c1.0	0.2370

Table 4: MAP scores for different value n of n -grams

3. Using n -grams and small words: While generating n -grams, small words do not get represented by single individual n -gram (when n is larger than the word's length). Thus, retrieval and indexing does not take care of small words which are shorter than n . Those words will not be searched as individuals. In this experiment, while generating the n -grams, if we get a word with length less than n , then we also include the word as it is. This process is used at the retrieval end as well. Also, we broke the n -gram sequences at the potential boundaries such as sentence end markers, braces, quotes and comas, since these markers break the association between two consecutive words. The results of this experiment are shown in table 5. Figure 1 shows the effect of length of n -grams on mean average precision score for DFR_BM25 model. It can be observed from the graph that MAP value is highest when n is 4 and it decrease on both sides of it, *i.e.*, the 4-gram based index not only improves the recall but also provides adequate amount of unique tokens to have good precision.

n -gram	Model	MAP (TD)	MAP (TDN)
2	BM25b0.75	0.1563	0.1585
	DFR_BM25c1.0	0.1548	0.1588
	TF_IDF	0.1593	0.1199
	In_expC2c1.0	0.1849	0.1639
3	BM25b0.75	0.3260	0.3345
	DFR_BM25c1.0	0.3278	0.3352
	TF_IDF	0.3236	0.3019
	In_expC2c1.0	0.3329	0.3283
4	BM25b0.75	0.3331	0.3569
	DFR_BM25c1.0	0.3355	0.3579
	TF_IDF	0.3342	0.3526
	In_expC2c1.0	0.3381*	0.3558*
5	BM25b0.75	0.3125	0.3393
	DFR_BM25c1.0	0.3136	0.3405
	TF_IDF	0.3132	0.3391
	In_expC2c1.0	0.3216	0.3475
6	BM25b0.75	0.2747	0.2989
	DFR_BM25c1.0	0.2753	0.2994
	TF_IDF	0.2756	0.2994
	In_expC2c1.0	0.2881	0.3085
7	BM25b0.75	0.2292	0.2615
	DFR_BM25c1.0	0.2306	0.2615
	TF_IDF	0.2299	0.2624
	In_expC2c1.0	0.2436	0.2362
8	BM25b0.75	0.1933	0.2236
	DFR_BM25c1.0	0.1958	0.2236
	TF_IDF	0.1952	0.2248
	In_expC2c1.0	0.2111	0.2658

Table 5: MAP scores for experiment which includes small words

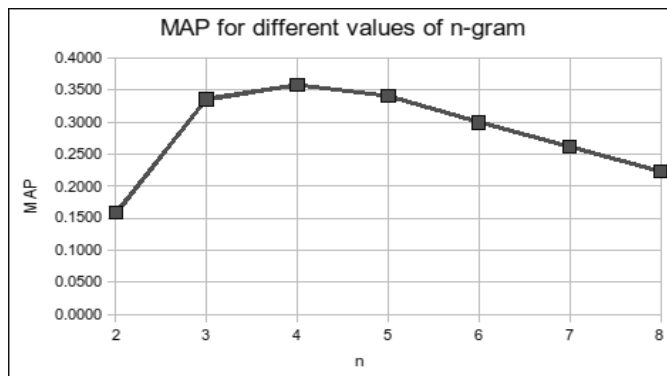


Figure 5: Effect of length of n in n -gram on MAP (TDN)

4. Using combination of two n -grams: In this setup, we used the combination of two different length n -grams. We verified whether the benefits of combination will be offset by problems incurred due to combining those. We used different combinations of n -grams as shown in table 6. N -gram1 was chosen to be longer than the optimal case, *i.e.*, greater than 4, and n -gram2 was chosen such that $n \leq 4$. It is found that, there is no significant improvement compared to 4-grams in any combination.

n -gram1	n -gram2	MAP
5	2	0.3499
5	3	0.3586
5	4	0.3519
6	2	0.3185
6	3	0.3387
6	4	0.3366
7	2	0.2847
7	3	0.3196
7	4	0.3220

Table 6: MAP scores for combination of 2 different n -grams for DFR_BM25.

6. Conclusion

It was found that among different length n -grams, 4-grams produces best results. It is because it strikes a balance between improving the precision by providing sufficient length for unique tokens and improving recall by breaking the words into enough small size chunks. Also we found that the performance of DFR_BM25 model is the best among all four models we have selected. From the results, it is also concluded that n -grams improve the accuracy considerably without using the language resource such as morphological analyser. In the fourth experiment, the combination of n -gram did not improve the MAP score. For future work, we propose to carry out a detailed investigation and analysis of experimental observations. The n -gram based technique can be applied to other Indian languages as well. We propose to carry out these experiments too.

Future work

In Unicode block for Devnagari script, there are some meta-characters which are not part of Devnagari script but they are required for the word formation. Its effect on formation of n -gram needs further investigation. Also we would like to work on skip-grams and its effect on Marathi.

References

1. Paul McNamee and James Mayfield, *Character N-gram Tokenization for European Language Text Retrieval*. Information Retrieval, 7:73-97, 2004.
2. Paul McNamee, *Textual Representations for Corpus-Based Bilingual Retrieval*, PhD Thesis, University of Maryland Baltimore County, December 2008.
3. Paul McNamee, Charles Nicholas, and James Mayfield, *Don't Have a Stemmer? Be Un+Concern+ed*. of the 31st Annual International Conference on Research and Development in Information Retrieval (SIGIR-08), Singapore, pp. 813-814, July 2008.
4. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, 1999.
5. Terrier information retrieval system. <http://ir.dcs.gla.ac.uk/terrier/>
6. Cross Lingual Information Access (CLIA), www.clia.iitb.ac.in
7. Ashish Almeida and Pushpak Bhattacharyya, *Using Morphology to Improve Marathi Monolingual Information Retrieval*, FIRE 2008. Kolkata, India
8. Anni Jarvelin, Antti Jarvelin, Kalervo Jarvelin, *s-grams: Defining generalized n-grams for information retrieval*, Elsevier, 2006