Preliminary Qualitative Analysis of Segmented vs Bigram Indexing in Chinese

Mun-Kew Leong, Hong Zhou
Institute of Systems Science
Singapore

Abstract

This paper investigates merging multiple methods of indexing for Chinese IR. Identical queries, differently segmented, are used to retrieve individual lists of documents which are then merged before evaluation. Two simple merge methods are discussed. Results on Chinese TREC queries 1 to 28 show improvement over either one of the indexing schemes by themselves. In addition, we examine the difference in the documents returned by each indexing method, i.e., do different indexing schemes retrieve different documents, or the same documents ranked differently, or something else. While we contrasted bigram based indexing with segmented based indexing, the same methods would apply between any two forms of indexing.

1 Introduction

Different queries, different search engines and different indexing schemes give different result sets for any given topic. We expect these results to differ, but it has not been investigated just *how* they differ. Specifically, we might ask the following questions:

- are there trends identifiable when holding one (or two) of query, engine, and indexing method constant, and varying the remainder?
- if there are trends, or correlations, do they identify with any intuitive (cognitive) classes?
- how consistent are these differences? For example, chinese trigram indexing consistently works well when searching for (Chinese) names, but on the average (based on the Chinese TREC queries) perform at about 60% of bigram indexing.

Lastly, we need also be concerned with what kinds of methodology and tools are needed and which are available for doing investigations into this area.

This TREC experiment was aimed at addressing some of those questions above. We focused on short queries using only the Chinese "Desc" field, and compared the result sets obtained by keeping query and engine the same, and varying the indexing method. Specifically, we looked at bigram vs. segmented indexing for the Chinese ad-hoc task. We also investigated various methods for merging the differing result sets to see their effects, and to try and account for differences, if any.

In preliminary work, using the TREC-5 queries, we obtained modest to significant improvement in recall and precision when we merged the results of individual bigram and segmented indexing runs. In both normalized and unnormalized fusion experiments, the merged results were never significantly worse than either individually over a range of query lengths, types of n-gram indexing, and two types of segmentation. In some cases, we had significant improvements, as will be described later.

2 Outline of Procedure

The experiment had the following basic outline.

1. create two indices, I_1, I_2, corresponding to two different indexing methods (e.g., bigrams and segmented (or "word" based) approches).
2. for each query, q_i, in a given set of queries,
   (a) run q_i against I_1 and I_2 to give lists α_i and β_i, each of n ranked documents.
   (b) extract the following: $\alpha_i \cup \beta_i, \alpha_i \backslash \beta_i, \beta_i \backslash \alpha_i, \alpha_i \cap \beta_i$. These are each sets of results showing some relationship between the two indexing methods for that query.
   (c) take top n documents in each list, and run against relevance judgements for that query
3. merge the results for all the queries, and run against all the relevance judgements to get comparative precison and recall figures to compare against baselines, etc., and to get an overall idea of relative performance of the two indexing methods under investigation.
4. analyse data for the two steps above

In our case specifically, the engine was the same for both indexing schemes, and the queries generated from each topic were the same, except insofar as the tokenization for bigram or segmented querying was necessary.

2.1 Variables

While simple, the methodology above allows us to investigate quite a few different variables, including the following:

- types of indexing:
  - segmentation (compound words, phrases)
    - greedy, short algorithms
    - dictionaries, types of dictionaries
  - n-grams (bigrams, unigrams, trigrams)
- merging returned records:
  - using raw scores
  - using normalized scores
    - within index
    - w.r.t. query terms (no., size)

A slight variation in the outline would also allow us to investigate differences in queries generated from the same topics. However, this would require that we could reliably qualitatively differentiate the differences in queries in a non-statistical manner.

2.2 Data Available

There is also a wealth of data to be mined from the statistics obtainable above. For example:

- Data:
  - S = segmented document result list
    B = bigrams document result list
    $n = 100$, $\cup$ = simple merge based on raw scores
  - $\mid S \cup B \mid = 140 \Rightarrow \mid S \cap B \mid = 60$ (i.e., 60 common documents between the lists)
  - Consider $T = S \cup B \mid 100$ (the list restricted to the top 100 docs)
  - $\mid T_S \mid = \mid T \backslash B \mid = 25$, $\mid T_B \mid = \mid T \backslash S \mid = 75$
- Relevance judgements on:
  - S, B, $S \cup B$, T, $T \backslash S$, $T \backslash B$, $S \backslash T$, $B \backslash T$, etc.
- Series of n = 100, 200, 300, ...
  - avoid floor and ceiling effects

2.3 Data Analysis

Given the data above, what are the interesting things to look for?

- What can the data tell us?
  - do different indices imply different documents retrieved?
  - do different indices imply different *relevant* documents retrieved?
  - how do the different indices relate with respect to precision vs. recall?
  - at what point do they begin to differ?
  - can we get both (high precision and high recall) with judicious fusion?
  - consider also the distribution of relevant documents
- Look to the document content for "why"
  - intuitive classes of documents?
  - consistency?
  - trends?

3 Results from TREC-5 Queries

The following results (Figure 1) were obtained from Chinese queries 1 TO 28 using bigrams in one case, and a dictionary based segmentation method in the other. The dictionary used was the one compiled by the University of Berkeley, and a greedy (longest match) algorithm was employed. The merging algorithm was to use the maximum score for intersecting documents and the original weight returned otherwise. No corrections were done for the different IDF values in the two indices.

Indexing	Segmented	Bigram	Merged
Recall/2182	1959	2119	2124
Precision @0.1	0.5125	0.5760	0.5724
Precision @0.5	0.3504	0.3864	0.3881
Avg Precision	0.3328	0.3683	0.3686
Exact R-Precision	0.3602	0.3998	0.3975

Figure 1: Various Indexing approaches, ad-hoc task using TREC-5 Data

Switching to a normalized merging algorithm (Figure 2), where scores were normalized with respect to the average score for a query returned by the respective indexing method, we get the following improvement:

Indexing	Segmented	Bigram	Merged
Recall/2182	1959	2119	2125
Precision @0.1	0.5125	0.5760	0.5776
Precision @0.5	0.3504	0.3864	0.3960
Avg Precision	0.3328	0.3683	0.3802
Exact R-Precision	0.3602	0.3998	0.4101

Figure 2: Improved merging algorithm, ad-hoc task using TREC-5 Data

The improvement is approximately 3% and is consistent across the various measures.

3.1 Examples of Specific Queries

We can also look at the statistics obtained from each of the queries. This gives us a better picture of which query was helped, hindered, or unchanged by the merging.

3.1.1 Topic 1

For example, consider just the top 50 records from topic 1. As we can see from the table below (Figure 3), looking at the fine grained level of just the top 50 records can give us some interesting insights.

Indexing	Sgmt	Bigm	Union	Inter	$S\backslash B$	$B\backslash S$
Cardinlty	50	50	85	15	35	35
Relvnt/13	3	9	9	3	0	6
Avg Prec	0.069	0.295	0.262	0.173	0	0.094
Prec 10	0.200	0.300	0.300	0.200	0	0.100
Prec 20	0.150	0.150	0.150	0.150	0	0.200
Prec R	0.154	0.231	0.231	0.231	0	0.077

Figure 3: Topic 1, top 50 documents, various data

As we can see, every relevant document found by the segmented method was also found by the bigram method. If we follow this query for the next 50 records, and then the next, etc., we will find the point (If ever) that the segmented method might find a record not found by the bigram method.

3.1.2 Topic 11

Similar to above, but for topic 11.

Indexing	Sgmt	Bigm	Union	Inter	$S\backslash B$	$B\backslash S$
Cardinlty	50	50	77	23	27	27
Relvnt/186	21	10	24	7	14	3
Avg Prec	0.042	0.011	0.031	0.012	0.038	0.006
Prec 10	0.200	0.200	0.200	0.300	0.500	0.200
Prec 20	0.300	0.200	0.200	0.350	0.500	0.150
Prec R	0.113	0.054	0.129	0.038	0.075	0.016

Figure 4: Topic 11, top 50 documents, various data

In this case, we see that the segmented approach works better, but there are 3 documents retrieved (in the top 50) by the bigram approach that are not retrieved by the segmented.

Again, further investigation with the next 50 results, etc., will give more information.

4 Results from TREC-6 Queries

Given the positive results from the TREC-5 topics, we decided to investigate the relative contributions of the two indexing methods in TREC-6, e.g., what kind of overlap was there in the documents retrieved on the same query by the different methods, where records retrieved by only one method ranked, and where *relevant* records retrieved by only one method ranked. We avoided using pseudo-relevance feedback even though that has significant improvement in earlier experiments (including the TREC-5 experiments above) to minimize the variables involved. The merging method used was to merge based on the raw score, removing the lower scoring duplicate documents.

The following table summarizes the results obtained in the submitted runs for TREC-6:

Indexing	Segmented	Bigram	Merged
Recall/2958	2619	2802	2723
Precision @0.1	0.7686	0.8103	0.7950
Precision @0.5	0.4964	0.6037	0.5095
Avg Precision	0.4709	0.5646	0.4903
Exact R-Precision	0.4689	0.5515	0.4941

Figure 5: Various Indexing approaches, ad-hoc task using TREC-6 Data

The results here go opposite from what occured with the TREC-5 topics. In all the measures above, the merged results are worse than the bigram results. However there is also consistency: the segmented approach was worst in all cases, the bigram was best in all cases, and the merged approach was between them. The merged numbers are closer to the segmented numbers, which may indicate that the merging algorithm favoured the segmented approach.

In looking at the results, this turned out to be the case. Changing the merging algorithm to the average based one used for TREC-5 queries in the previous section brought the merged numbers much closer to the bigram numbers, but bigrams still turn out to perform better.

There is some possibility that a ceiling effect may be in operation. Looking at overall statistics, the bigram approach had 18 $\geq$ median and 4 best at the 100 relevant retrieved level, 24 $\geq$ median and 16 best at the 1000 relevant retrieved level, and 19 $\geq$ median and 1 best with respect to average precision.

Note also that the above results were obtained without the use of pseudo-relevance feedback. When used on the TREC-5 queries, this feedback generally improved precision and recall by about 10%.

Unfortunately, due to lack of time, finer grain analysis (i.e., at the 50 document level, etc.) demonstrated for TREC-5 queries in the previous section was not done on the TREC-6 results. This would greatly help towards understanding how merging changes the documents retrieved and their ranking with the number of documents considered, and would avoid any ceiling or floor effects.

5 Conclusions

There are two sets of conclusions to draw here. With respect to the results obtained from the TREC-6 Chinese track, we found the following:

- TREC-6 results indicate bigrams still perform best
  - documents returned by other indexing methods are mostly subsets of bigram documents
  - segmentation is better only in isolated cases
- so far, no clear cognitive or linguistic classes are identifiable
- segmentation rankings and bigram rankings have different shapes, implying that simple merging algorithms may not work well.

With respect to the methodology proposed, however, it is obvious that such methods generalize between any two indexing schemes. We believe that the data gathered and the possible analyses to be done can shed a lot of light on the qualitative differences between different forms of indexing.

We hope to continue the analysis started on the TREC-6 data, and also intend to apply similar analysis between various n-grams to further understand the relationship between n-gram and segmented indexing. There is some anecdotal evidence that suggests that trigrams would give better precision, and this, among other hypotheses, will be tested in future work.