A First Investigation on Mongolian Information Retrieval

Guanglai Gao, Wei Jin, Fei Long, Hongxu Hou
School of Computer Science
Inner Mongolia University
Hohhot China 010021
Email：{csggl, csjinwei, csfeilong, cshhx}@imu.edu.cn

Abstract

In this paper we present an attempt to build a test collection for Mongolian IR as well as some preliminary tests about the key issues in Mongolian Information Retrieval: using a stoplist and using word stemming. Our preliminary tests will show that while these basic operations on Mongolian can bring slight improvements in retrieval effectiveness, many problems remain.

The results using stemming and stoplist show that the stemming can potentially lead to some gain in retrieval effectiveness; The stoplist slightly improve retrieval effectiveness, but it can reduce the index significantly.

Keywords: Mongolian IR, Language Model, Test Collection, Mongolian Information Processing

1、 Introduction

Traditional Mongolian (for short “Mongolian”) is the main language of Inner Mongolia Autonomous Region in China. It has a long history and is a widely used language [1]. Despite the fact that Mongolian is a language spoken by millions of people, no IR study has been carried out on it utill now.

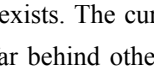
To our knowledge, no test collection for Mongolian IR exists. The current state of Mongolian processing is far behind other languages. This is the very reason that motivated our study: to construct a test collection in Mongolian and to promote research activities in Mongolian IR.

Our study has two main purposes. First, we aim to construct a test collection following the TREC methodology. Such a test collection can benefit other researchers. Second, we aim to perform some preliminary tests on the test collection to see if the traditional IR methods are effective in Mongolian IR. In particular, we will test several basic options for IR: the utilization of a stoplist, and word stemming.

In the following sections, we will first describe the Mongolian language and the problems for Mongolian IR. Then we describe our effort to construct a test collection. Several IR methods will be tested on the test collection.

2、The Mongolian Language and Mongolian IR

Mongolian is a phonemic script language. It has 35 letters. A Mongolian word is formed by several letters. In the written language, a word is written top-down by letters joined cursively together. Figure 1 shows a Mongolian sentence. We can see that words are separated by spaces.

Generally, a Mongolian word can be divided into two parts: stem and affix. Mongolian language is an agglutinative language. It does not have pre-affix and mid-affix, but only has suffix. More than one suffix can be added one after another [2]. For example, the word  (bolbasuragul – maturate) is formed by the stem  and three suffixes. Schematically, we can illustrate the formaytion of the word as follows:

 (bol) – ripe (verb)+ (basun) →

(bolbasun) – mature(vi.), practice + (ra) → (bolbasura) [1] - mature(adj.), civilized + (gul) → (bolbasuragul) - maturate.

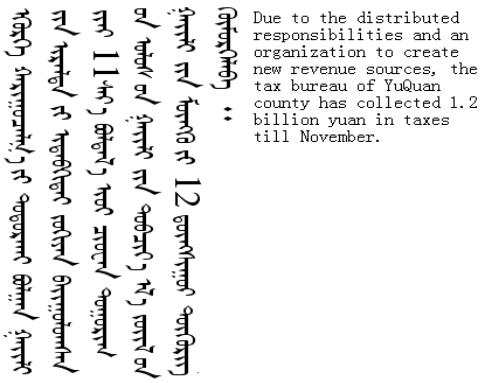


Figure 1. An example of Mongolian sentence

This example shows the necessity to perform word stemming. However, we can also see the potential problem. If we perform a slight stemming, then some other suffixes will remain, making it difficult to unify words with similar meanings. On the other hand, if the stemming process is too aggressive, then we will unify words that have quite different meanings, even though they may share the same stem. For example, if we remove all the suffixes from the above word, we will transform the word meaning maturate to a stem meaning ripe. This creates the danger to retrieve documents on a different topic.

In some cases, suffixes are written separated by Mongolian space. Most of these suffixes are often added on Noun. For example, (bagsi-in, meaning teacher's) = (bagsi, meaning "teacher") + (in a possessive mark).

3、Construction of a Mongolian Test Collection

3.1 Document collection

The construction of document collection is very important for information retrieval, but it is also a time consuming work. By now, there is no functional IR system available because the computational processing of Mongolian started very late. It is currently difficult to collect Mongolian documents because of the limited sources where digital Mongolian texts are available.

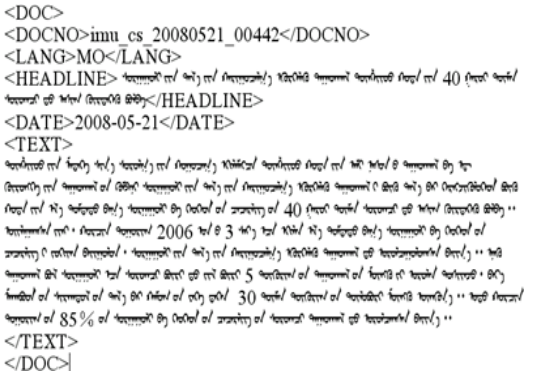


Figure 2. An example of Mongolian document

We collected Mongolian documents from two sources: one part of the document collection is provided by Inner Mongolia Ordos Daily. It contains 4,984 documents of approximately 23MB. The other part of document collection comes from Inner Mongolia Daily. It contains 15,341 documents of approximately 82MB. For our test collection,

Finally, we added tags into these documents according to TREC document's format. Figure 2 gives a sample document formatted in TREC style. We have in total 20,235 documents.

3.2 Topic sets

As in TREC collections which include a series of research topics, we also determined a set of test topics in the TREC nomenclature. Each topic contains title, brief description, detailed description, and related word phrases.

We have determined 11 topics. Figure 3 show a sample topic and Figure 4 shows a translation of it in English (Notice that the translation is not part of the collection. It is provided here to help understanding the query).

[1] The letter (n) will disappear after the next suffix.

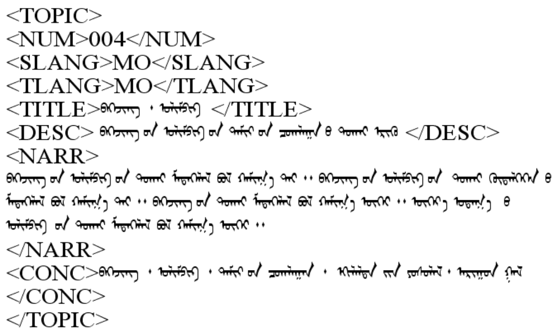


Figure 3. An example of Mongolian search topic

<TOPIC>
<NUM>004</NUM>
<SLANG>MO</SLANG>
<TLANG>MO</TLANG>
<TITLE> Beijing, Olympics </TITLE>
<DESC> find information of the Beijing Olympics </DESC>
<NARR>
The content of the related documents should include the news of Beijing Olympics and the news of Related activities of Beijing Olympics. The document will be regarded as irrelevant if it only introduced the news of Beijing and the news of other Olympics.
</NARR>
<CONC> Beijing, Olympic, game, opening ceremony, cauldron
</CONC>
</TOPIC>

Figure 4. Translation of topic in Figure 3

3.3 Relevance judgments

We have adopted the pooling method to evaluate the relevance of documents. First we use Lemur and Lucene to make the pool. For each system and for each search topics, we use 100 top documents retrieved by each system. The documents in the pool are then shown to human assessors who ultimately decide the relevance of these documents.

The collection is still in development. We are adding new documents and new queries in it. However, the size of the collection already allows us to perform some preliminary experiments on Mongolian IR. In the next section, we will describe our preliminary tests with the collection.

4、 Mongolian IR methods

4.1 Indexing units for Mongolian IR

For indexing, we are faced with two basic problems: word stemming and stop words.

- Word stemming

Term is the basic unit of indexing in traditional IR. A Mongolian sentence is composed of many words separated by space, each word being a string of Mongolian characters. Mongolian words can be divided into two parts: stem and affix, such as (nisugsen) (flied) = (nis)(fly)+ (ugsen) (affix for past tense). As we mentioned, Mongolian only has suffix. In many cases, although the addition of suffix slightly change the meaning of a word, for IR purpose, the difference in meaning is not significant. Therefore, it may be desirable to perform word stemming to remove one or more word forming suffix. Word stemming generally produces two effects: It reduces the size of index, and it may also increase recall.

- Stoplist

Extremely common words often have little value in helping select documents matching a user need. These words are included in a stoplist. Our general strategy for determining a stop-list is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent terms, and hand-filtered them according to their meaning and roles in the language. The selected stop words are mostly prepositions, adverbs, conjunctions, such as (and), (what), (about), (this).

4.2 IR using Language model

Language modeling is a quite general formal approach to IR. The goal is to rank documents by P(D|Q) where D is a document and Q is a query. Using Bayes rule, we have:

$$P(D|Q) = P(Q|D)P(D)/P(Q)$$

P(Q) is the same for all documents, and so can be ignored [3]. The prior P(D) is often treated as uniform across all D and so it can also be ignored. But given these simplifications, we return results ranked by simply P(Q|D), the probability of the query Q given D. This probability is estimated using a language model θ_D of the document.

The most common way to do this is using the multinomial unigram language model. Under this

model, we have that:

$$P(Q \mid \theta_D) = \prod_{w_i \in Q} P(w_i \mid \theta_D)$$

5、Experimentations

The Indri system is used for the experimentation [4].

5.1 Experimented methods

Here is an example to our queries.  (The information about Beijing Olympic Games). In this query,  (about) and  (inquire) are stop words. The terms  (Pekinese),  (Olympic's),  (locomotor),  (pageant's) are translated into  (Beijing),  (Olympic),  (athletics) and  (pageant) if word stemming is used.

Smoothing [5] is a method used to overcome both the 'zero probability' and data sparseness problem. Three kinds of tests have been carried out to compare Mongolian IR with/without word stemming and with/without stoplist.

- Dirichlet smoothing

$$P(w_i \mid \theta_D) = \frac{tf_{w_i,D} + \mu P(w_i \mid \theta_C)}{|D| + \mu} \quad (1)$$

Where μ=2500 is the Dirichlet prior, θ_C is a language model of the collection and |D| is the total count of words in document D.

- Jelinek-Mercer smoothing

$$P(w_i \mid \theta_D) = \frac{(1-\lambda) tf_{w_i,D}}{|D|} + \lambda P(w_i \mid \theta_C) \quad (2)$$

Where λ=0.4 is a smoothing parameter.

- Two-stage smoothing

$$P(w_i \mid \theta_D) = \frac{(1-\lambda)(tf_{w_i,D} + \mu P(w_i \mid \theta_C))}{|D| + \mu} + \lambda P(w_i \mid \theta_C) \quad (3)$$

Where λ=0.4 and μ=2500 are the smoothing parameters.

5.2 Experimental results

The experimental results with the three smoothing methods are described in the following tables.

- Smoothing methods

Figure 5 shows a comparison between different smoothing methods for document models. All the methods illustrated in the figure use stemming and stoplist. We can see that different smoothing methods lead to slightly different retrieval effectiveness. However, globally, the effectiveness is comparable. We can conclude that all these smoothing methods that have been successfully applied to other languages also apply to Mongolian IR.

- Using stemming

From Tables 1, we can see that when stemming is used, the effectiveness is improved. And stemming allowed us to greatly reduce the size of the index. The number of unique indexes is reduced from 125,796 (without stemming) to 74,001 (with stemming). As a consequence, the total size of the index is also reduced by 20%.

- Using stoplist

We can observe that using stoplist, we can also obtain slightly effectiveness than without using stoplist.

6、 Concluding remarks and future work

This paper describes our first attempt to Mongolian IR. In this paper, we investigated the following fundamental problems in IR: word stemming and utilization of a stoplist.

In order to test different retrieval methods, we constructed a small test collection. This is the first test collection for this language. Although it is still at its first stage, we have been able to perform some preliminary tests. These tests suggest the following conclusions:

- Word stemming in Mongolian IR is important. It can improve retrieval effectiveness.
- Using stoplist can slightly improve retrieval effectiveness, and it can reduce the index significantly.

	Dirichlet				Jelinek-Mercer				Two-Stage			
	No Stoplist No Stem	Stoplist No Stem	No Stoplist Stemming	Stoplist Stemming	No Stoplist No Stem	Stoplist No Stem	No Stoplist Stemming	Stoplist Stemming	No Stoplist No Stem	Stoplist No Stem	No Stoplist Stemming	Stoplist Stemming
0.0	0.80883	0.80883	0.83798	0.83084	0.83647	0.83647	0.89634	0.89634	0.80883	0.80883	0.84447	0.83798
0.1	0.81734	0.81189	0.82798	0.81957	0.79564	0.79564	0.80944	0.80422	0.81394	0.81038	0.82674	0.81647
0.2	0.80639	0.81167	0.82812	0.82300	0.74695	0.74695	0.76263	0.76263	0.80839	0.81334	0.83544	0.82994
0.3	0.78258	0.78547	0.80401	0.80363	0.71221	0.71221	0.73107	0.72839	0.79579	0.79143	0.79447	0.78843
0.4	0.69485	0.67688	0.73370	0.71963	0.67736	0.67736	0.71772	0.71593	0.68524	0.67067	0.73458	0.72597
0.5	0.61808	0.63787	0.66369	0.66584	0.60741	0.61489	0.65075	0.64300	0.59042	0.61291	0.64518	0.64376
0.6	0.53299	0.55743	0.58749	0.58731	0.52406	0.55099	0.54008	0.5642	0.50921	0.53320	0.56366	0.57028
0.7	0.39715	0.42766	0.47659	0.49094	0.42203	0.44711	0.45876	0.50105	0.35875	0.40444	0.43904	0.45464
0.8	0.27344	0.31301	0.26074	0.30571	0.25226	0.29133	0.27289	0.32652	0.26658	0.31192	0.2619	0.30918
0.9	0.11021	0.11768	0.10944	0.12929	0.10627	0.11604	0.10103	0.12024	0.11914	0.13279	0.11679	0.14784
1.0	0.03243	0.03284	0.03177	0.03253	0.03995	0.04078	0.03853	0.04030	0.03241	0.03288	0.03175	0.03258
Avg	0.53403	0.54375	0.56014	0.56439	0.52005	0.52998	0.54357	0.55480	0.52624	0.53843	0.55400	0.55973

Table1. Language Model smoothing

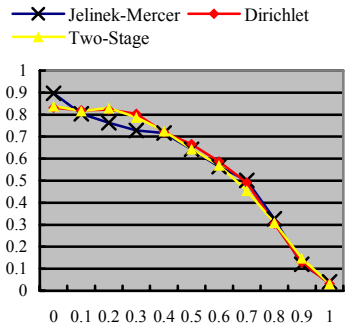


Figure 5. Recall – precision graphs for smoothing methods

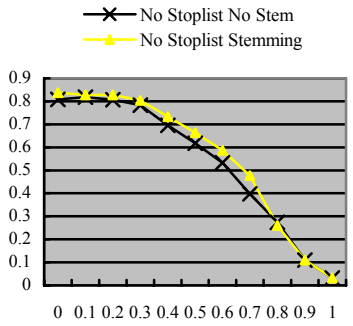


Figure 6. Recall – precision graphs for using stemming graphs

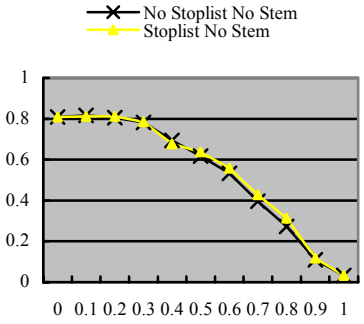


Figure 7. Recall – precision for using stoplist

The present study is still at its preliminary stage. We are still trying to construct a larger test collection for Mongolian IR. It is hopeful that other researchers can benefit from this collection in the future.

Among the future interesting investigations, we would like to investigate Mongolian morphology in more depth. In our current experimentations we only removed the last suffix. The stemming process should continue to remove other possible suffixes. Mongolian has seven vowels, and some of them have the same morphology but have different pronunciation. A possible solution is to perform query expansion by replacing letters in a query word that are confoundable with other letters.

Acknowledgement

We thank professor Jian-Yun Nie of Université de Montréal for helpful comments on this work. This research is funded by State Key Basic Research and Development Program of China (973 Program) (2007CB316503).

References

[1] Quejingzhabu. Mongolian Codes. Inner Mongolia University Publishing House. 2000.

[2] Qingge'ertai. Mongolian Syntax. Inner Mongolia people publishing house. 1991.

[3] Song, F., & Croft, B. A general language model for information retrieval. In Proceedings of the 1999 ACM SIGIR conference on research and development in information retrieval, 279-280, 1999.

[4] Donald Metzler, Trevor Strohman, Howard Turtle and W. Bruce Croft. Indri at TREC 2004: Terabyte Track

[5] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst., 22(2):179–214. 2004.