A Trigger Language Model-based IR System

ZHANG Jun-lin SUN Le QU Wei-min SUN Yu-fang

Open System & Chinese Information Processing Center
Institute of Software, The Chinese Academy of Sciences
P.O.BOX 8718,Beijing 100080
junlin01@iscas.cn

Abstract

Language model based IR system proposed in recent 5 years has introduced the language model approach in the speech recognition area into the IR community and improves the performance of the IR system effectively. However, the assumption that all the indexed words are irrelative behind the method is not the truth. Though statistical MT approach alleviates the situation by taking the synonymy factor into account, it never helps to judge the different meanings of the same word in varied context. In this paper we propose the trigger language model based IR system to resolve the problem. Firstly we compute the mutual information of the words from training corpus and then design the algorithm to get the triggered words of the query in order to fix down the topic of query more clearly. We introduce the relative parameters into the document language model to form the trigger language model based IR system. Experiments show that the performance of trigger language model based IR system has been improved greatly. The precision of trigger language model increased 12% and recall increased nearly 10.8% compared with Ponte language model method.

1 Introduction

Using language models for information retrieval has been studied extensively recently(Jin et al 2002 Lafferty and Zhai 2001 Srikanth and Srihari 2002 Lavrenko and Croft 2001 Liu and Croft 2002). The basic idea is to compute the conditional probability $P(Q|D)$, i.e. the probability of generating a query Q given the observation of a document D. Several different methods have been applied to compute this conditional probability. In most approaches, the computation is conceptually decomposed into two distinct steps: (1) Estimating a document language model; (2) Computing the query likelihood using the estimated document model based on some query model. For example, Ponte and Croft emphasized the first step, and used several heuristics to smooth the Maximum Likelihood of the document language model, and assumed that the query is generated under a multivariate Bernoulli model (Ponte and Croft 1998). The BBN method (Miller et al 1999) emphasized the second step and used a two-state hidden Markov model as the basis for generating queries, which, in effect, is to smooth the MLE with linear interpolation, a strategy also adopted in Hiemstra and Kraaij (Hiemstra and Kraaij 1999). In Zhai and Lafferty (Zhai and Lafferty 2001), it has been found that the retrieval performance is affected by both the estimation accuracy of document language models and the appropriate modeling of the query, and a two stage smoothing method was suggested to explicitly address these two distinct steps.

It's not hard to see that the unigram language model IR method contains the following assumption: Each word appearing in the document set and query has nothing to do with any other word. Obviously this assumption is not true in reality. Though statistical MT approach (Berger and Lafferty 1999) alleviates the situation by taking the synonymy factor into account, it never helps to judge the different meanings of the same word in varied context. In this paper we propose the trigger language model based IR system to resolve the problem. Though the basic idea of using the triggered words to improve the performance of language model was proposed by Raymond almost 10 years ago

(Raymond et al 1993), Our method adopts a different approach for other objectivity in the IR field. Firstly we compute the mutual information of the words from training corpus and then design the algorithm to get the triggered words of the query in order to fix down the topic of query more clearly. We introduce the relative parameters into the document language model to form the trigger language model based IR system. Experiments show that the performance of trigger language model based IR system has been improved greatly.

In what follows, Section 2 describes trigger language model based IR system in detail. Section 3 is our evaluation about the model. Finally, Section 4 summarizes the work in this paper.

2 Trigger Language Model based IR System

2.1 Inter-relationship of Indexing Words

In order to find out the inter-relationship of words in some specific context, we consider the co-occurring times of different words within fixed sized text window of the document. When the co-occurring time is large enough, we think that relationship is meaningful. Mutual Information is a common tool to be applied under this situation. So we compute the mutual information as following:

$$\omega(w_a, w_b) = \frac{\frac{N(w_a, w_b, L_w)}{N_w \cdot (L_w - 1)}}{\left(\frac{N(w_a)}{N_w}\right) \cdot \left(\frac{N(w_b)}{N_w}\right)} = \frac{N(w_a, w_b, L_w) \cdot N_w}{(L_w - 1) \cdot N(w_a) \cdot N(w_b)} \quad (1)$$

where N_w denotes the size of the vocabulary, $N(w_a, w_b, L_w)$ is the co-occurring times of word w_a and w_b within L_w sized window in training set. $N(w_a)$ is the count of the word w_a appearing in the training set and $N(w_b)$ is the count of word w_b appearing in the training set.

We use the corpus provided by IR task of NTCIR2 (NTCIR 2002) as the training set to compute the mutual information of words. This corpus contains nearly 100 thousands news articles encoding in BIG5 charset. We think the mutual information which is larger than 25 is meaningful. Considering the stop words in document or query are useless to represent the content, we remove 200 highest frequent words from the document before computation. Table 1 shows some examples with higher mutual information.

测试 (test)	字母表(alphabet):1895 铺轨(rail):1353 限界(delimitation):758 风洞实验(windtunnel):473 测试仪(meter):421 试纸(test paper):403
飞弹 (missile)	空对空飞(antiaircraft):1063 研制(develop):708 长射程(long-range):472 反坦克(anti-tank):354
贿赂 (bribe)	偷税(tax dodging):3462 营私舞弊(jobbery):2603 联邦调查(FBI):1041 投票人(voter):730 湛江(zhanjiang):478 犹他州(Utah):427
毁灭性 (truculency)	检查人员(scrutator):710 长程飞弹(long-range missile):497 恐怖主义(terrorism):457 生化(biochemistry):390 均势(equipoise):327 瘟疫(plague):334 巴格达(Bagdad):325

Table 1. Examples of Mutual Information

2.2 Algorithm of Triggered Words by Query

Generally speaking, a word always represents many different meanings and its exact meaning adopted in specific topic can be determined by the co-occurring words in its

context. Different meaning of a word often lead to the different vocabulary set of related word.

In order to find out the exact meaning of the words contained by the query in IR system, we design the algorithm to compute the triggered vocabularies of query. It is just these triggered words that show the exact meaning of the words in query in some specific context and help fix down the topic of query more clearly. The basic idea behind the algorithm is as following: By computing the mutual information, we can derive the relative words of a query word. All these words mean the semantically related vocabularies of the query word under different contexts. We propose that if the intersection of the derived related words of different words in query is not null, the words in the intersection is useful to judge the exact meaning of the words in query. At the same time, the more times an intersection word appears in related vocabulary set of different query word, the higher the weight of this word to fix down the topic of the query is. So we design the following algorithm to compute the triggered vocabulary set of query:

Algorithm 1: Triggered vocabularies by query
Input: Vocabulary set I of query word and its co-occurring words after removing the stop words in the query.

$$I = \{< q_1, S_1 >, < q_2, S_2 >, ... < q_i, S_i > ..... < q_n, S_n >\}$$

Output: Triggered vocabulary set T.

Setp 1. Initialize the set $T = \phi$.

Setp 2. for(i =2;i<=n;i++)

{

for(j=1;j<= C_n^i ;j++)

{

2.1get the different combination $L_j = \{<q_{j,1}, S_{j,1}>, <q_{j,2}, S_{j,2}>, ...<q_{j,i}, S_{j,i}>\}$ which contains i elements from set I;

2.2 if any vocabulary set $S_{j,k}(1 < k <= i)$ in L_j contains no element, then we turn to 2.4 , otherwise we turn to 2.3;

2.3 Compute the intersection $T_{i,j}$ of all vocabulary set $S_{j,k}(1 < k <= i)$ in L_j . Here $T_{i,j} = \{< w_1, \alpha_1 >, < w_2, \alpha_2 >, ...< w_m, \alpha_i >\}$, where $\alpha_w = \frac{\log i}{2}$, ($1 =< w <= i$). α_w is the word weight decided by the length of L_j;

2.4 $T = T \cup T_{i,j}$, adopting the higher word weight α_w during the merging process;

}

}

Step 3. Output the triggered vocabulary set T;

2.3 Similarity Computation of Query and Document

We use the similar strategy with Ponte language model method (Ponte and Croft 1998) to compute the similarity between the query and the document. That is, we firstly construct the simple language model according to the statistical information of vocabulary and then compute the generative probability of the query. The difference is that the trigger language model method takes the context information of a word into account. So we compute the triggered words set of query q according to algorithm 1.This way we get the triggered vocabulary set $T_q = \{< w_1, \alpha_1 >, < w_2, \alpha_2 >, ... < w_m, \alpha_m >\}$. This set contains the words triggered by query and it is these triggered words that determine the exact meaning of the vocabularies in query among the several optional choices. This helps fix down the topic of query more clearly. Introducing the triggered words factor into the document language model, we can form the trigger language model based information retrieval system.

The similarity of query and document can be computed as following:

$$P(Q \mid M_d) = \prod_{i=1}^{l(Q)} (\sum_{j=1}^{l(d)} C_j \cdot p(q_i \mid d_j) + \frac{tf(q_i)}{cs}) \quad (2)$$

$$p(q_i \mid d_j) = \begin{cases} 1 & q_i = d_j \\ \alpha_j & (q_i \neq d_j) \wedge (d_j \in T_q = \{<w_1, \alpha_1>, <w_2, \alpha_2>, ...<w_m, \alpha_m>\}) \\ 0 & other \end{cases} \quad (3)$$

(1) $Q = \{q_1, q_2, ...q_i, ...q_{l(Q)}\}$ denotes query and $l(Q)$ is the length of the query;

(2) M_d denotes the trigger language model of document d;

(3) $d = \{d_1, d_2, ...d_j, ...d_{l(d)}\}$ denotes a document in document set and $l(d)$ is the length of the document;

(4) $C_j = \frac{f(d_j)}{l(d)}$ is the weight parameter of words d_j in a document. Here $f(d_j)$ means the account of the words d_j appearing in the document.

(5) $p(q_i \mid d_j)$ denotes the probability of q_i being triggered by the document word d_j .When 2 words are same, the probability equals 1. If they are different and the word d_j belongs to the triggered vocabulary set of query, the probability equals the according parameter in the T_q ,otherwise the probability is 0。

(6) $\frac{tf(q_i)}{cs}$ is used for data smoothing; here $tf(q_i)$ denotes times of query word q_i appearing in document set and cs denotes the total length of documents which contains the word q_i .

3 Experiment Results

3.1 Corpus

The corpus we used to evaluate the performance of our proposed trigger language model IR system is the document set offered by the traditional Chinese Document set of NTCIR3 for the IR task. The corpus consists of 381681 news articles from Hong Kong and Taiwan with varied topics. After the word segmentation, the document set contains 150700953 words. Among them,127519 different words are the entries of the vocabulary. The average length of each document is 394.

The 50 queries offered by NTCIR3 IR task are contained in a XML file and each query consists of following elements: Topic Number(NUM),Topic Title(TITLE),Topic question(DESC),Topic Narrative(NARR) and Topic Concepts(CONC). In order to make it easer to compare the performance of the different IR methods, we adopt the Topic Question field as the query and regard the top 1000 retrieval documents as the standard result of the experiment.

3.2 Analysis of Experiment Results

We design 3 relative experiments to evaluate the trigger language model IR method: vector space model, Ponte language model based method and the trigger language model approach. Precision and recall are two main evaluation parameters. As for the trigger model IR method, the optimal size of the text window is 20 content words and the mutual information over 25 is regarded as the meaningful information. Experiment results can be seen in table 2.

The data of column $\%\Delta 1$ in table 2 shows the performance improvement of Ponte language model compared with vector space model. The data tells us that the precision of language model based method increased 10% and recall increased nearly 13.7%. The data of column $\%\Delta 2$ in table 2 shows the performance improvement of trigger language model compared with Ponte language model method. From the data we can see that the precision of trigger language model increased 12% and recall increased nearly 10.8%. We can draw the conclusion that the trigger language model has improved the performance greatly. The performance comparison can be showed more clearly in figure 1.

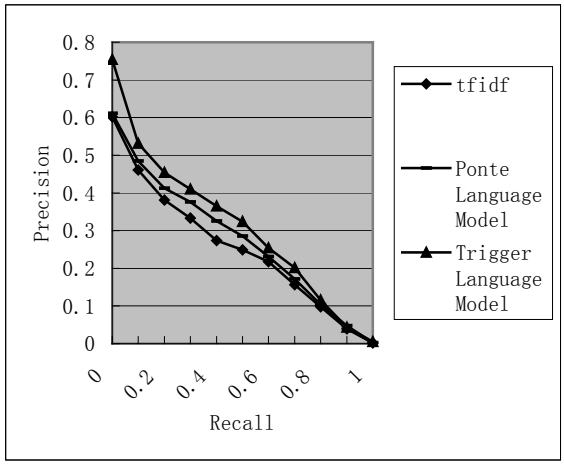

Figure 1. Precision-Recall of 3 methods

	Tfidf	Lm(ponte)	Trigger lm	%Δ1	%Δ2
Relevant:	3284	3284	3284	-----	-----
Rel.ret:	1843	2096	2322	13.7	10.8
Precision:					
0.00	0.6016	0.6109	0.7537	+2	+23
0.10	0.4607	0.4844	0.5314	+5	+10
0.20	0.3812	0.4123	0.4541	+8	+10
0.30	0.3336	0.3757	0.4094	+12	+9
0.40	0.2738	0.3255	0.3648	+18	+12
0.50	0.2495	0.2854	0.3237	+14	+13
0.60	0.2179	0.2313	0.2538	+6	+9
0.70	0.1566	0.1716	0.2011	+9	+17
0.80	0.0978	0.1041	0.1153	+6	+10
0.90	0.0389	0.0474	0.0435	+21	-8
1.00	0.0019	0.0025	0.0055	+31	+120
Avg:	0.2377	0.2610	0.2933	+10	+12

Table 2. Experiment results

4 Conclusion

Language model based IR system proposed in recent 5 years has introduced the language model approach in the speech recognition area into the IR community and improves the performance of the IR system effectively. However, the assumption that all the indexed words are irrelative behind the method is not the truth. Though statistical MT approach alleviates the situation by taking the synonymy factor into account, it never helps to judge the different meanings of the same word in varied context. In this paper we propose the trigger language model based IR system to resolve the problem. . Firstly we compute the mutual information of the words from training corpus and then design the algorithm to get the triggered words of the query in order to fix down the topic of query more clearly. We introduce the relative parameters into the document language model to form the trigger language model based IR system. Experiments show that the performance of trigger language model based IR system has been improved greatly.

Acknowledgement

This work is supported by Beijing New Star Plan of Technology & Science(NO.H020820790130) and the National Science Fund of China under contact 60203007.

References

Berger A. and Lafferty J. (1999). Information retrieval as statistical translation. In *Proceedings of SIGIR '99*. pp. 222-229.

Jin R., Hauptmann A.G. and Zhai C.(2002) Title Language Model for Information Retrieval. In Proceedings of the 2002 ACM SIGIR Conference on Research and Development in Information Retrieval.

Hiemstra D. and Kraaij W. (1999), Twenty-One at TREC-7: ad-hoc and cross-language track, In *Proceedings of the seventh Text Retrieval Conference TREC-7*, NIST Special Publication 500-242, pages 227-238, 1999.

Lafferty J. and Zhai. C. (2001) Document language models,query models and risk minimization for information retrieval. In Proceedings of the 24th ACM SIGIR Conference,pp.111-119.

Lavrenko,V., and Croft,W.B.(2001) .Relevance based language models. In Proceedings of the 24th ACM SIGIR Conference.pp.120-127.

Liu,X. and Croft,W.B.(2002).Passage Retrieval Based on Language Models. In Proceedings of the 11th International Conference on Information and Knowledge Management. Pp.375-382

Miller D., Leek T. and Schwartz R. M. (1999). A hidden Markov model information retrieval system. Proceedings of SIGIR'1999, pp. 214-222. .

NTCIR Workshop (research.nii.ac.jp/ntcir/index-en.html)

Ponte J. and Croft W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of SIGIR' 1998*, pp. 275-281.

Raymond Lau, Roni Rosenfeld and Salim Roukos(1993) Trigger-based Language Models: A Maximum Entropy Apporach. *Proceedings ICASSP '93*, Minneapolis, MN, pp. II-45 - II-48.

Srikanth M. and Srihari. R(2002). Biterm Language Models for Document Retrieval. In Proceedings of the 2002 ACM SIGIR Conference on Research and Development in Information Retrieval.

Zhai C. and Lafferty J.(2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceeding of SIGIR'01*, 2001, pp. 334-342.