Applying Language Model into IR Task

Zhang Junlin, Sun Le, Zhang Yongchen, Sun Yufang
Chinese Information Processing Center,
Institute of Software,Chinese Academy of Sciences,
P.O.Box 8717,Beijing,100080,P.R.China
E_mail:junlin01@iscas.cn

Abstract

The language modeling approach to information retrieval is a new framework that has been proposed and developed within the past five years. In NTCIR4, we focus our experiments on evaluating the effectiveness of the LM based IR method. In C-C run, we observe that the average precision of two-stage smoothing language modeling IR increase about 10.09% compared with VSM method when "DESC" field is used as query while decrease nearly 7.81% when query is "TITLE" field. This proved that the two-stage language modeling IR method could increase the performance in longer query effectively. In E-C run, the average precision of two-stage smoothing LM IR decreases about 38.09% compared with VSM method when "DESC" field is used as query while decreases nearly 32.60% when query is "TITLE" field. We think the two-stage smoothing LM IR method is more sensitive to the noise words introduced by wrongly translated query than SVM method.

Keywords: Information Retrieval, Language Model, Two-stage smoothing, CLIR

1 Introduction

The language modeling approach to information retrieval (IR) is a new framework that has been proposed and developed within the past five years, although its roots in the IR literature go back more than twenty years. Research carried out at a number of sites has confirmed that the language modeling approach is a theoretically attractive and potentially very effective probabilistic framework for building IR systems.[1,2,3,4,5,6,7]

This paper mainly describes the methods and procedures we took in participating in NTCIR4 CLIR track. We focus our experiments on evaluating the effectiveness of the language model based IR methods.

The paper is organized as follows: Section 2 is the brief introduction of the language model based IR system. Section 3 describes the language model methods and relative procedures we took in Monolingual IR subtasks and Bilingual CLIR subtasks in NTCIR4. In Section 4 we analyze the experiment results. Section 5 concludes this paper.

2 Language Model Based IR System

Recent advances in Information Retrieval are based on using Statistical Language Models (SLM) for representing documents and evaluating their relevance to user queries. Language Model (LM) has been explored in many natural language tasks including machine translation and speech recognition. In LM approach to document retrieval, each document D is viewed to have its own language model, M_D. Given a query Q, documents are ranked based on

the probability, $P(Q|M_D)$, of their language model generating the query. While the LM approach to information retrieval has been motivated from different perspectives, most experiments have used smoothed unigram language models that assume term independence for estimating document language models.

In most approaches, the computation of language model based IR method 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 [1] 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. The BBN method [2] 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. In Zhai and Lafferty [8][9][10], 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 proposed to explicitly address these two distinct steps.

There are many advantages for language model IR paradigm compared with traditional IR models. Firstly, being able to estimate retrieval parameters is a major advantage of using language models for information retrieval. Another advantage of using language models is that we can expect to achieve better retrieval performance through the more accurate estimation of a language model or through the use of a more reasonable language model. Thus, we will have more guidance on how to improve a retrieval model than in a traditional model. Finally, language models are also useful for modeling the sub-topic structure of a document and the redundancy between documents.

3 Our Work at NTCIR4

We participated in 2 subtasks in CLIR track of NTCIR4. Data listed in Table 1 shows the average precision of each subtask we participated in.

In NTCIR4, We aim at evaluating the effectiveness of the Language Model based IR method. So we design several experiments to compare the Language Model IR system with traditional VSM method.

Run Types		Average Precision		
		D	T	DN
C-C	VSM	0.1774	0.1944	0.1774
	Language Model	0.1953	0.1792	/
E-C	VSM	0.0021	0.0273	0.0021
	Language Model	0.0013	0.0184	/

Table1.Average precision of all subtasks of ISCAS-------Relax

3.1Monolingual IR Subtask (C-C Run)

Since word boundaries are not marked in Chinese written text, word segmentation is necessary to break Chinese sentences into indexing terms, which can be words, single characters, two characters, and so on. All the subtasks which are relevant with Chinese document collection are word based index. Our segmentation algorithm is called bi-direction maximal match algorithm. It scans the Chinese sentence two times by looking up the maximal match term in a general purpose dictionary: The first time is from left to right and the second time reverse the scan order from right to left. This way we can identify and avoid some type of segmentation ambiguity.

In our experiment in NTCIR4, we performed several different C-C runs based on either VSM or Language Model method to

compute the similarity of the query and documents. VSM is employed as a baseline to evaluate the language model method. In VSM, the term of vector is word. If T={ t_j } is a term set, then query vector v_j of topic j can be express $V_j=(v_{j1},v_{j2},\ldots v_{jn})$,in which v_{jk} denotes the weight of t_k in v_j.The vector $D_i=(d_{i1},d_{i2},\ldots,d_{in})$ denotes a document ,d_{ik} denotes the weight of t_k in d_i.The similarity between v_j and d_i is calculated by following formula

$$s_j = \sum_{k=1}^{n} d_{ik} * v_{jk} \Big/ \sqrt{\sum d_{ik}^2 + \sum v_{jk}^2}$$

The language model IR method is our main concern in NTCIR4. The query likelihood retrieval method is based on the original language modeling approach proposed by Ponte and Croft in 1998[1]. It involves a two-step scoring procedure. First, estimate a document language model for each document, and, second, compute the query likelihood using the estimated document language model directly. In our experiments, we try to evaluate the language model using the two-stage smoothing approach proposed by Zhai in literature [9]. That is, the first stage uses Dirichlet prior smoothing method to improve the estimate of a document language model; this method normalizes documents of different lengths appropriately with a prior sample size parameter. The second stage is intended to bring in a query background language model to explicitly accommodate the generation of common words in queries. The retrieval scoring formula is as following :

$$P(Q \mid D,S) = \prod_{i=1}^{n} ((1-\lambda)\frac{c(q_i,D)+\mu P(q_i \mid S)}{\mid D \mid +\mu} + \lambda P(q_i \mid S))$$

In our experiments, the parameter λ is set to be 0.3 and the optimal value of parameter μ is 542.

An important advantage of the two-stage language models is that they explicitly capture the different influences of the query and document collection on smoothing. It is known that the optimal setting of retrieval parameters generally depends on both the document collection and the query; decoupling the influence of the query from that of documents makes it easier to estimate smoothing parameters independently according to different documents and different queries.

3.2 Query Translation of CLIR

The main concern of subtasks in the Bilingual CLIR is query translation. The easiest way to find translations is to look up each query term in a bilingual dictionary. However, We can't neglect problems brought by this method such as coverage, spelling norms. Applying MT in CLIR is also a straightforward approach. Another option to using translation dictionaries is using a parallel or comparable corpus, that is, the same or similar text written in different languages.

Our aim is the evaluation of the language model IR method in NTCIR4, So we didn't do much work on the query translation in E-C subtask of NTCIR4. We directly use lexical approach to translate the English query into Chinese. Then we search the relevant documents in the Chinese document collection with our Chinese monolingual IR system.

4 Analysis of the Experiment Results

IR Model	C-C Run ID	E-C Run ID
Language Model+ 2-stage smoothing	ISCAS-C-C-T-02 ISCAS-C-C-D-04	ISCAS-E-C-T-02 ISCAS-E-C-D-04
VSM	ISCAS-C-C-T-01 ISCAS-C-C-D-03	ISCAS-E-C-T-01 ISCAS-E-C-D-03

Table 2. Different runs and their IR model

In order to evaluate the effectiveness of the language model IR method, we design several experiments to compare LM based method with VSM method. Table 2 shows the relationship

between the various Run ID and the IR model they adopted.

From the experimental results, we observe the following rules:

(1) In C-C run, we observe that the average precision of two-stage smoothing language modeling IR increase about 10.09% compared with VSM method when "DESC" field is used as query(fig 1) while decrease nearly 7.81% when query is "TITLE" field.(fig 2) Generally speaking, the title fields are the concise and short queries which contain only several keywords compared with the "Desc" field. So we can draw the conclusion from the experiments that two-stage language modeling method can be a better choice if the query is relatively longer. This shows that the two-stage language modeling IR method could reduce the noise in longer query containing common words effectively.

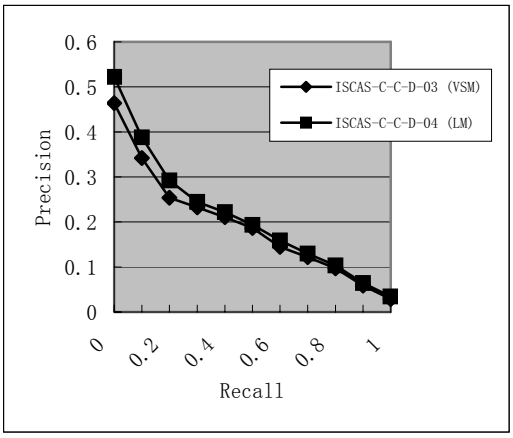

Fig 1. Precision-Recall of C-C-D Run

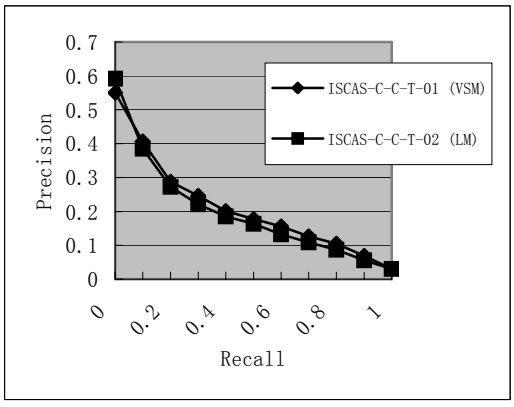

Fig 2. Precision-Recall of C-C-T Run

(2)In E-C subtasks (figure 3), the performance of the language model method is always worse than the VSM, no matter which fields are used as the query (Title or Desc). The average precision of two-stage smoothing LM IR decreases about 38.09% compared with VSM method when "DESC" field is used as query while decreases nearly 32.60% when query is "TITLE" field. This result is out of our expectation because we thought the result should be similar with the C-C run. We explain this as following: We only directly use lexical approach to translate the English query into Chinese, so too much noise of translation is introduced into the translated query. Even though the 2-stage smoothing approach can deal with the noise in query, it is only effective for the common words in query. The wrongly translated query contains many content words that are beyond the ability of the two-stage smoothing approach. While it seems VSM didn't suffer so much from the worse translation. We think the two-stage smoothing LM IR method is more sensitive to the noise words introduced by wrongly translated query than SVM method.

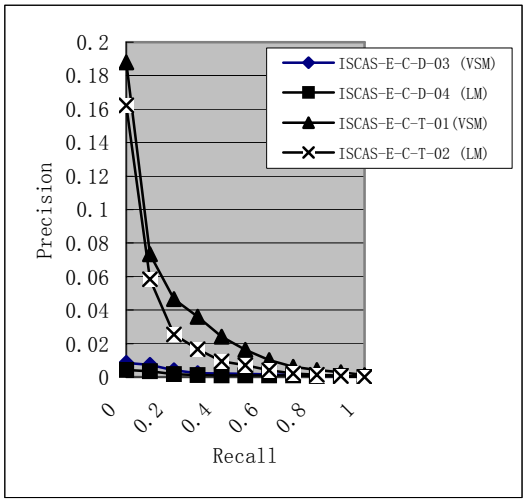

Fig 3. Precision-Recall of E-C Subtask

(3) In E-C subtasks, we found that the performance get worse whenever we use the "Desc" field as the query compared with the "Title" field. We thought it's because the "Desc"

field is longer compared with “title” field and we just use the simple query translation approach by directly looking up the lexicon. This simple query translation method will bring much noise into the translated query. So much more irrelevant documents are searched out.

5 Conclusion

This paper mainly describes the methods and procedures we took in participating in NTCIR4 CLIR track. We focus our experiments on evaluating the effectiveness of the language model based IR method. The analysis of the language model based IR method compared with VSM method is also presented in the paper. In C-C run, we observe that the average precision of two-stage smoothing LM IR increase about 10.09% compared with VSM method when “DESC” field is used as query while decreased nearly 7.81% when query is only “TITLE” field. This indicates that the two-stage language modeling method can achieve better performance in longer query effectively. In E-C subtasks, the average precision of two-stage smoothing decrease about 38.09% compared with VSM method when “DESC” field is used as query while decreased nearly 32.60% when query is “TITLE” field. We think the LM method is more sensitive to the noise words introduced by wrongly translated query than VSM method.

Acknowledgments

This work is supported by the National Science Fund of China under contact 60203007 and Beijing New Star Plan of Technology & Science (NO.H020820790130). We also thank the anonymous reviewers for kind suggestions.

Reference

[1] J.Ponte and W.B.Croft , A Language Modeling Approach to Information Retrieval. In Proceedings of the 1998 ACM SIGIR Conference on Research and Development in Information Retrieval. Pages 275-281,1998.

[2]D.H.Miller, T.Leek and R.Schwartz. A hidden Markov model information retrieval system. In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval. Pages 214-221,1999.

[3]A.Berger and J.Lafferty. Information retrieval as statistical translation. In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval. Pages 222-229,1999.

[4]T.Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval. Pages 50-57,1999

[5]S.Deerwester,S.T.Dummais etc. Indexing by latent semantic analysis. Journal of the Society for Information Science,41(6):381-407,1990

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

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

[8]C Zhai and J Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of the 2001 ACM SIGIR Conference on Research and Development in Information Retrieval. 2001

[9] C Zhai and J Lafferty. Two-stage language model for information retrieval. In Proceedings of the 2002 ACM SIGIR Conference on Research and Development in Information Retrieval. pp.49-56. 2002

[10]C.Zhai. Risk Minimization and Language Modeling in Text Retrieval. Ph.D. dissertation, Carnegie Mello University.