11-2005

Accurately Extracting Coherent Relevant Passages Using Hidden Markov Models

Jing JIANG
University of Illinois at Urbana-Champaign, jingjiang@smu.edu.sg

ChengXiang ZHAI
University of Illinois at Urbana-Champaign

Accurately Extracting Coherent Relevant Passages Using Hidden Markov Models

Jing Jiang
Department of Computer Science
University of Illinois at Urbana-Champaign

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

ABSTRACT

In this paper, we present a principled method for accurately extracting coherent relevant passages of variable lengths using HMMs. We show that with appropriate parameter estimation, the HMM method outperforms a number of strong baseline methods on two data sets.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

General Terms

Algorithms

Keywords

Passage Retrieval, Hidden Markov Models

1. INTRODUCTION

It is often desirable for information retrieval (IR) systems to retrieve relevant passages as opposed to whole documents to further filter out irrelevant information. A critical problem in passage retrieval is how to accurately locate the boundaries of *coherent relevant passages*, which we refer to as *passage extraction*. Passage retrieval involves both passage extraction and passage ranking. Accurate passage extraction not only allows an IR system to precisely point to the most relevant parts of a document, but can also potentially improve document ranking and relevance feedback by using short, relevant passages rather than long, noisy whole documents. Despite its importance, passage extraction has not been seriously addressed in existing work. Passage retrieval methods have so far been evaluated for document ranking[3, 4], passage ranking[1], and question answering[6]. In all three tasks, however, ranking is the main component being evaluated. As a result, existing passage retrieval methods are not designed to extract passages that are both query dependent and coherent in content.

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

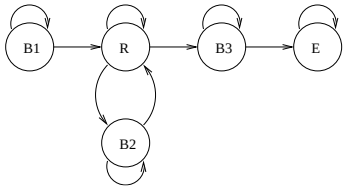

Figure 1: HMM Structure

In this paper, we directly address the passage extraction problem. We focus on studying how to accurately extract the single most relevant passage from a relevant document. We present a principled method for accurately detecting coherent and query-specific relevant passages of variable lengths using hidden Markov models (HMMs). Our method is similar to [5] and [2] in spirit, but different in both the HMM architecture and the parameter estimation methods. Evaluation on two data sets shows that with appropriate parameter estimation, our HMM method outperforms a number of strong baseline methods. We further show that the HMM method can be applied on top of any basic passage extraction method to improve the passage boundaries.

2. HMM-BASED PASSAGE EXTRACTION

In the language modeling approach to information retrieval, a document is treated as a bag of words, where each word is independently drawn from a language model. For passage extraction, we treat a document as a sequence of words, and model this word sequence as being generated from two language models, where the relevant segment is generated from a *relevance language model*, and the non-relevant segments are generated from a *background language model*. These two stochastic processes are connected though another stochastic process, which determines when the language model switches from one to the other. The whole process essentially forms a hidden Markov model.

Figure 1 shows the 5-state HMM structure we constructed for extracting a single relevant passage from a document. Each state is associated with a language model. An arrow in the figure indicates a non-zero transition probability. A document starts from background state B_1, where a non-relevant segment is generated from the background language model. At some point, the document switches to state R, which, together with state B_2, generates the relevant passage. R is associated with the relevance language model, which is related to the specific query. B_2 is used as a smoothing factor to account for the words not captured

by R. Transition probabilities between R and B_2 adjust the degree of smoothing. Finally the document switches from state R to background state B_3, where another non-relevant segment is generated from the background language model. The last state E generates only a special symbol, which is appended to each document to enforce that a document goes through the complete HMM. When all the output probabilities and all the transition probabilities in the HMM are set, we can use Viterbi algorithm to find the state sequence that has most likely generated the word sequence of a document, and hence locate the relevant passage in the document.

The language models are estimated as follows. For the background language model, we use the whole document collection as samples and maximum likelihood estimator to estimate the probability of each word. For the relevance language model, we explored three estimation strategies, all using maximum likelihood estimator. The first is to simply use the original query words. We call this method *HMM-q*. The second strategy incorporates *within-document pseudo feedback*. The idea is to first extract from the document a short passage highly likely to be relevant to the query, and then use words in this starting passage as samples for estimation. Such an estimated language model should presumably attract the text surrounding the starting passage that is similar to the starting passage, and thus extend the passage to the true relevant passage that is coherent in content and has natural topical boundary. We call this method *HMM-wd*. The third strategy incorporates *cross-document pseudo feedback*. The idea is similar to *HMM-wd*, but we use a set of starting passages from different documents that are relevant to the same query to estimate a relevance language model for this query. We call this method *HMM-cd*. Once all output probabilities are set, we can train the transition probabilities. Since passage length is document specific, and transition probabilities affect passage length, we train the transition probabilities for each document.

3. EXPERIMENTS

We implemented a number of baseline methods for comparison. *BL-simple* returns the passage between the first and the last query words in the document. *BL-win* uses a fixed-size sliding window to search for the passage with the most query words. *BL-cos* and *BL-pivoted* are similar to *BL-win*, but use a cosine measure and a pivoted cosine measure to find the most relevant passage. Experiments were carried out on two data sets: a synthetic data set created from TREC DOE abstracts, and a subset of TREC 2004 HARD track data. We use precision, recall and F1 for evaluation. Let T be the true relevant passage, E be the extracted passage, and O be the overlap between T and E. Precision is the length of O (in number of words) divided by the length of E, recall is the length of O divided by the length of T, and F1 is the harmonic mean of precision and recall.

Table 1 shows the experiment results. Stars indicate the best performance figures among all methods on the same data set. For the window-based methods, we set the window size to the average relevant passage length, which is the best the system can do. *HMM-wd* and *HMM-cd* use passages extracted by *HMM-q* for feedback. We see from Table 1 that *HMM-cd* performed the best among all methods if we use F1 as the measure. This shows that with good parameter estimation from feedback, the HMM-based method outperforms all baseline methods. *HMM-q* achieved high precision but low recall. However, when pseudo feedback was used, as in *HMM-wd* and *HMM-cd*, recall increased substantially compared with *HMM-q*, while precision did not decrease much. This improvement agrees with our hypothesis that if we use short, accurate passages as pseudo feedback to estimate the relevance language model, the HMM method can automatically extend the passages to the natural topical boundaries.

Collection		**Precision**	**Recall**	**F1**
DOE	*BL-simple*	0.869	0.591	0.632
	BL-win	0.779	0.777	0.730
	BL-cos	0.764	0.763	0.717
	BL-pivoted	0.749	0.745	0.701
	HMM-q	0.940	0.500	0.561
	HMM-wd	0.932	0.630	0.659
	HMM-cd	0.941*	0.858*	0.862*
HARD04	*BL-simple*	0.670	0.909	0.666
	BL-win	0.668	0.759	0.621
	BL-cos	0.671	0.781	0.628
	BL-pivoted	0.672	0.783	0.629
	HMM-q	0.709*	0.726	0.585
	HMM-wd	0.686	0.877	0.656
	HMM-cd	0.671	0.969*	0.706*

Table 1: HMM Methods vs. Baseline Methods.

4. CONCLUSIONS AND FUTURE WORK

Passage extraction is an essential component of passage retrieval. We proposed an HMM-based method to extract coherent and query-specific relevant passages of variable lengths. We studied three methods for parameter estimation. Experimental results show that with appropriate parameter estimation, our HMM method outperformed a number of baseline methods. Because there is no restriction on the method used for extracting starting passages for feedback, the HMM method can be applied on top of any simple method to improve the passage boundaries, and thus provides a framework for incorporating feedback for passage extraction. Our future work will focus on how to detect multiple relevant passages in a single document and how to exploit the extracted passages to improve document ranking.

5. REFERENCES

[1] J. Allan. Hard track overview in trec 2003: High accuracy retrieval from documents. In *Proceedings of TREC 2003*, 2003.

[2] L. Denoyer and H. Zaragoza. HMM-based passage models for document classification and ranking. In *23rd BCS European Annual Colloquium on Information Retrieval*, 2001.

[3] M. Kaszkiel and J. Zobel. Passage retrieval revisited. In *Proceedings of SIGIR'97*, 1997.

[4] X. Liu and W. B. Croft. Passage retrieval based on language models. In *Proceedings of CIKM'02*, 2002.

[5] E. Mittendorf and P. Schäuble. Document and passage retrieval based on hidden Markov models. In *Proceedings of SIGIR'94*, 1994.

[6] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of SIGIR'03*, 2003.