Exploiting Proximity Feature in Bigram Language Model for Information Retrieval

Seung-Hoon Na*, Jungi Kim*, In-Su Kang# and Jong-Hyeok Lee*
Pohang University of Science and Technology, South Korea *
Kyungsung University, South Korea #
{nsh1979,yangpa,jhlee}@postech.ac.kr*, dbaisk@ks.ac.kr#

ABSTRACT

Language modeling approaches have been effectively dealing with the dependency among query terms based on N-gram such as bi-gram or trigram models. However, bigram language models suffer from adjacency-sparseness problem which means that dependent terms are not always adjacent in documents, but can be far from each other, sometimes with distance of a few sentences in a document. To resolve the adjacency-sparseness problem, this paper proposes a new type of bigram language model by explicitly incorporating the proximity feature between two adjacent terms in a query. Experimental results on three test collections show that the proposed bigram language model significantly improves previous bigram model as well as Tao's approach, the state-of-art method for proximity-based method.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval Models

General Terms: Algorithms, Experimentation

Keywords: Language models, proximity, bigram model, term dependency

1. INTRODUCTION

The locality among query terms is a useful feature to boost the retrieval effectiveness since the semantics of some queries can be identified based on multi-term level rather than single-term level. Regarding this, language modeling approaches have been flexibly extended to bi-gram or N-gram language model in order to exploit the sequential dependency among query terms [2]. Researchers have observed that bi-gram model is helpful to improve the retrieval effectiveness [2].

However, the paraphrasing phenomenon makes a serious problem when estimating bi-gram language models. Due to the paraphrasing, when two adjacent terms in a query are given, these terms in highly-relevant documents can be represented by different manners such as reversing their positions, or inserting other terms between two terms, thus resulting in adjacency-sparseness problem. Thus, the estimated bi-gram model becomes biased, causing the unfair preference of only to documents where query terms are adjacently-appeared but not to documents where query terms are locally-appeared.

To deal with this, this paper proposes a new type of language model to explicitly incorporate the proximity feature. Note that the previous bigram language model exploited the probability to generate next query term for a given current query term in the stream of a given document. In this paper, we define a general type of bigram language model by the average of the generative probabilities of next query term from the set of local contexts of the current query term. To reasonably estimate the new bigram language model, we define the local context of a term by minimum-length passages between current query term and next query term. The proposed approach has some advantages compared to Tao's approach [3], the state-of-the-art proximity-driven approach which explicitly uses the proximity feature but is independently designed to retrieval models. 1) While Tao's approach is largely heuristic, our approach is model-driven approach without losing the elegancy of language modeling approaches. 2) While Tao's approach should calculate distances among all possible pairs in a query, our approach calculate distances among only adjacent query terms in a query, thus providing more efficient manner. 3) Our approach shows better performance than Tao's approach in standard TREC test collections.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. EXPLOITING PROXIMITY FEATURE IN BIGRAM LANGUAGE MODEL

2.1 Background

First, let us re-visit the background of bigram language model. Suppose that query Q is given by the stream of $q_1 \cdots q_n$. Then, the query likelihood of query Q from document D is calculated as follows [2]:

$$P(Q|D) = P(q_1|\ _D) \prod_{i=2}^{Q} P(q_i|q_{i-1},\ _D) \quad (1)$$

where $P(q_i|\ _D)$ and $P(q_i|q_{i-1},\ _D)$ is unigram model and bigram language model of document D, respectively.

2.2 Bigram Language Model using Proximity Feature

The proposed bigram language model indicates the average of probabilities to generate term q_i from the set of local contexts of q_{i-1}. Generally, an arbitrary pseudo passage sample containing q_{i-1} can be used as a local context of q_{i-1}. Let $LC(q_{i-1}, D)$ be the set of such local contexts of q_{i-1} in document D. Then, the proposed model redefines the estimation of $P(q_i|q_{i-1},\ _D)$ as follows:

$$P(q_i|q_{i-1}, D) = \sum_{lc \in LC(q_{i-1}, D)} P(q_i|lc) P(lc|LC(q_{i-1}, D)) \quad (2)$$

where lc indicates a member of $LC(q_{i-1}, D)$. $P(lc|LC(q_{i-1}, D))$ indicates the prior probability of local context lc, which is assumed to be uniformly distributed - $1/|LC(q_{i-1}, D)|$. $P(q_i|lc)$ is defined as follows:

$$P(q_i|lc) = \begin{cases} c(q_i; lc)/len(lc) & \text{if } len(lc) \leq W \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where W is a new parameter which is regarded as the maximally allowed window size where two terms q_{i-1} and q_i proximally appear in the local context lc. When $len(lc)$ is larger than W, $P(q_i|lc)$ simply indicates the generative probability of q_i from local context lc which is obtained from MLE(maximum likelihood estimation) - $c(q_i;lc)/len(lc)$.

$LC(q_{i-1},D)$ plays the most important role to estimate $P(q_i|q_{i-1},\ _D)$. To define $LC(q_{i-1},D)$, this paper assumes that minimum cover of q_i for each occurrence of q_{i-1} is an element of $LC(q_{i-1},D)$. Suppose that $q_{i-1}^{(k)}$ is k-th occurrence of q_{i-1} in D. We denote the minimum cover of q_i for $q_{i-1}^{(k)}$ by $span(q_{i-1}^{(k)},q_i)$, which is defined by the minimum-length passage among candidate passages which contain both $q_{i-1}(k)$ and q_i. For example, suppose that D is given by "1:t_1, 2:t_2, 3:t_3, 4:t_1, 5:t_2, 6:t_4, 7:t_5, 8:t_2, 9:t_4, 10:t_1" and Q is given by "t_1 t_2". Let us denote a passage by $[p,q]$ where p and q are the start position and the end position of passage, respectively. For the first term occurrence of $t_1^{(1)}$ - 1:t_1, the minimum cover - $span(t_1^{(1)},t_2)$ is [1-2], and for $t_1^{(2)}$ - 4:t_1, $span(t_1^{(2)},t_2)$ is [4-5], for $t_1^{(3)}$ - 10:t_1, $span(t_1^{(3)},t_2)$ is [8-10]. Thus, $LC(t_1,D)$ is {[1-2], [4-5], [8-10]}. Note that $|LC(q_{i-1},D)|$ is $c(q_{i-1};D)$ from this definition. Thus, Eq. (2) is rewritten by

$$P(q_i|q_{i-1},D) = \sum_{span(q_{i-1}^{(k)},q_i)\in LC(q_{i-1},D)} \frac{P(q_i|span(q_{i-1}^{(k)},q_i))}{c(q_{i-1},D)} \quad (4)$$

2.3 Applying Dirichlet-Prior Smoothing

Since it is well-known that Dirichlet-prior smoothing is better than Jelinek-Mercer smoothing (for short keyword queries), we focus on how the proposed bigram-model can be applied to Dirichlet-prior smoothing. Dirichlet-prior smoothing for unigram language model is formulated as follows [4]:

$$P(q_i|\ _D) = \frac{len(D)P(q_i|\hat{}_D) + \mu P(q_i|\ _C)}{len(D)+\mu} \quad (5)$$

where μ is a smoothing parameter, $P(q_i|\hat{}_D)$ is MLE for unigram language model of D - $c(q_i;D)/len(D)$, and $P(q_i|\ _C)$ is the background collection language model. Note that $len(D)$ plays an important role for obtaining smoothed model $P(q_i|\ _D)$ as additional evidences. Unlike unigram language model, the evidence sample for bigram language model is not unique, i.e. it consists of several minimum covers. In this paper, we assume that the length of the evidence sample for smoothing of the bigram language model is simply $c(q_{i-1};D)W$. Thus, the bigram model version of Eq. (4) is formulated by

$$P(q_i|q_{i-1},D) = \frac{c(q_{i-1}W\cdot P(q_i|q_{i-1},\hat{}_D) + \mu P(q_i|\ _C)}{c(q_{i-1},D)W+\mu} \quad (6)$$

Unigram language model is used as back-off for bigram language model, i.e. when $c(q_{i-1};D)$ is 0 or $P(q_i|q_{i-1},\hat{}_D)$ is 0, we use Eq. (5) of unigram language model instead of Eq. (6).

3. EXPERIMENTATION

For evaluation, we used standard TREC test collections for ad-hoc retrieval - WT2G and WT10G (WT2G is used for TREC8, and WT10G is used for TREC9 and TREC10). The standard method was applied to extract index terms. We first separated words based on space characters, eliminated stopwords, and then applied Porter's stemming. The title field is utilized as query type of all test collections. The parameter μ for Dirichlet-prior smoothing is differently selected depending on each collection so that it maximizes the retrieval effectiveness of unigram language model. W is fixed to 5. Table 1 shows the retrieval effectiveness on these three test collections. For comparison with previous methods, we append the results of the previous bigram language model and Tao's approach. We used Tao's *MinDist* for the proximity of query terms - (Q,D) due to its better performance, and used Dirichlet-prior smoothing using the same μ as the basic retrieval model. Four methods are abbreviated by Unigram (the baseline), Bigram (the previous bigram model), Tao's Prox (Tao's approach [3]), and ProxBigram (the proposed bigram model) in the table. To check whether or not bigram-based or proximity-based method significantly improves the baseline (Unigram), we performed the Wilcoxon sign ranked test and attached † and ‡ to the performance number of each cell in the table when the test passes at 95% and 99% confidence level, respectively.

Table 1: Retrieval performances (Mean Average Precision) of four different methods. All methods use Dirichlet-prior smoothing as basic retrieval model. Unigram, Bigram and ProxBigram indicate unigram language model, the previous bigram model, and the proposed bigram model, respectively. Tao's Prox indicates the most recent work of proximity-based approach.

Collection	*Unigram*	*Bigram*	*Tao's Prox*	*ProxBigram*
WT2G	0.3101	0.3149	0.3165	**0.3324**‡
TREC9	0.1965	0.2062 †	0.2013‡	**0.2149**‡
TREC10	0.1946	0.1964	0.1965	**0.2000**†

As shown in table 1, the proposed bigram language model significantly improves the unigram language model, and better performs than previous bigram language model and Tao's proximity-based approach. This result consistently shows that the proposed model can effectively resolve the adjacency-sparseness problem.

4. CONCLUSION

This paper proposed a new type of bigram language model to explicitly support the proximity feature, in order to resolve the adjacency-sparseness problem. Experimental results indicate that the proposed model is promising by showing a better performance than Tao's proximity-based method, and previous bigram language model. Furthermore, the proposed model is more efficient than Tao's proximity method, since our model allows us to calculate only within-document distances between only adjacent query terms. In addition, the proposed model is much more efficient than the dependency language model which makes non-trivial burden since full syntactic parsing should be pre-applied to whole documents [1]. In the future, we will explore the proposed bigram model on various different setting of local context sets.

Acknowledgement

This work was supported in part by MKE & IITA through IT Leading R&D Support Project and also in part by the BK 21 Project in 2008.

5. REFERENCES

[1] J. Gao, J.-Y. Nie, G. Wu, and G. Cao. Dependence language model for information retrieval. In *SIGIR '04*, pages 170–177, 2004.

[2] F. Song and W. B. Croft. A general language model for information retrieval. In *CIKM '99*, pages 316–321, 1999.

[3] T. Tao and C. Zhai. An exploration of proximity measures in information retrieval. In *SIGIR '07*, pages 295–302, 2007.

[4] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, pages 334–342, 2001.