Biterm language models for document retrieval — **Source link** ↗

Munirathnam Srikanth, Rohini K. Srihari

Institutions: University at Buffalo

Related papers:

- A language modeling approach to information retrieval

- A general language model for information retrieval

- Dependence language model for information retrieval

- A Markov random field model for term dependencies

- Information retrieval as statistical translation

Biterm Language Models for Document Retrieval

Munirathnam Srikanth
State University of New York at Buffalo
Buffalo, NY, 14228
srikanth@cedar.buffalo.edu

Rohini Srihari
State University of New York at Buffalo
Buffalo, NY, 14228
rohini@cedar.buffalo.edu

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information search and retrieval—*Retrieval models*; H.3.3 [**Information storage and retrieval**]: Systems and software—*Performance evaluation*

General Terms

Algorithms, Experimentation, Theory

Keywords

Language models, Information retrieval, N-Grams, Biterms

1 Introduction

Statistical Language Models(LM) have been used in many natural language processing tasks including speech recognition and machine translation [5, 2]. Recently language models have been explored as a framework for information retrieval [9, 4, 7, 1, 6]. The basic idea is to view each document to have its own language model and model querying as a generative process. Documents are ranked based on the probability of their language model generating the given query. Since documents are fixed entities in information retrieval, language models for documents suffer from sparse data problem. Smoothed unigram models have been used to demonstrate better performance of language models against vector space or probabilistic retrieval models for document retrieval.

Song and Croft [10] proposed a general language model that combined bigram language models with Good-Turing estimate and corpus-based smoothing of unigram probabilities. Improved performance was observed with combined bigram language models. The language models explored for information retrieval mimic those used for speech recognition. Specifically, in the bigram model a document d represented as word sequence $w_1, w_2, \cdots, w_n$ is modeled as

$$P(w_1, w_2, \cdots, w_n) = P(w_1)P(w_2|w_1) \cdots P(w_n|w_{n-1}) \quad (1)$$

Here the order of terms is important. A combined bigram model is an interpolation model with back-off to unigram

probabilities.

$$P(w_i|w_{i-1}) = \lambda_1 \; P_2(w_i|w_{i-1}) + (1 - \lambda_1) \; P_1(w_i) \quad (2)$$

where λ_1 is bigram weighting factor. In practice, the bigram probability P_2 is approximated to the ratio of the occurrence counts of (w_{i-1}, w_i) to that of w_{i-1} in document d.

$$P_2(w_i|w_{i-1}, d) \approx \frac{C(w_{i-1}, w_i|d)}{C(w_{i-1}|d)} \quad (3)$$

2 Language models designed for IR

Unlike language modeling for speech recognition, the language models for information retrieval need only to record co-occurrence of features or words. For a query of *information retrieval*, a back-off bigram model will give more weight to document containing *information retrieval* than a document containing *retrieval of information*. To this end, *Biterm Language Models* are introduced. Biterm language models are similar to bigram language models except that the constraint of order in terms is relaxed. A document containing *information retrieval* and a document containing *retrieval of information* will be assigned the same probability of generating the query using biterm language models.

To distinguish unordered word-pairs from order word-pairs or *bigrams* in statistical language modeling terminology, we refer to the former as *biterms*. Unordered word-pairs have been explored as document features for document retrieval [8] in vector space models as well as text categorization [3] applications.

The biterm probabilities can be approximated using the frequency of occurrence of terms. Three approximation methods are suggested here. In the first case, biterm probability of $\{w_{i-1}, w_i\}$ is viewed as an average of bigram probabilities (P_{bg}) of the ordered pairs – (w_{i-1}, w_i) and (w_i, w_{i-1}).

$$P_{BT1}(w_{i-1}, w_i|d) \approx \frac{1}{2}[P_{bg}(w_{i-1}, w_i|d) + P_{bg}(w_i, w_{i-1}|d)] \quad (4)$$

The second approximation is based on computing biterm probabilities from the term frequency in documents. It is similar to the bigram probability in (3). The biterm probability of term pair $\{w_{i-1}, w_i\}$ is computed as the ratio of the number of occurrences of the term pair $\{w_{i-1}, w_i\}$ to that of the occurrence count of the term w_{i-1} in document d.

$$P_{BT2}(w_i|w_{i-1}, d) \approx \frac{C(w_{i-1}, w_i|d) + C(w_i, w_{i-1}|d)}{2 \times C(w_{i-1}|d)} \quad (5)$$

In P_{BT2} the denominator depends on the occurrence count of w_{i-1}. To make the approximation truly independent of the term order, it can be set to the ratio of the frequency

of the term-pair to the minimum of the frequencies of terms w_{i-1} and w_i.

$$P_{BT3}(w_i|w_{i-1}, d) \approx \frac{C(w_{i-1}, w_i|d) + C(w_i, w_{i-1}|d)}{2 \times min\{C(w_{i-1}|d), C(w_i|d)\}} (6)$$

P_{BT2} and P_{BT3} are *ad-hoc* approximations for biterm probabilities.

The sparse data problem in representing documents using language models is handled by smoothing biterm probabilities using unigram probabilities. The unigram probability $P(w_i|d)$ of a term in a document is in turn smoothed using its corpus probability $P(w_i|C)$. Thus the biterm and unigram probabilities are computed by

$$
\begin{aligned}
P(w_i|w_{i-1}, d) &= \alpha_1 P_{BT}(w_i|w_{i-1}, d) + \\
&\quad (1 - \alpha_1) P(w_i|d) \qquad (7) \\
P(w_i|d) &= \alpha_2 P_1(w_i|d) + (1 - \alpha_2) P(w_i|C) \quad (8)
\end{aligned}
$$

where α_1 and α_2 are constants.

The weights for biterm and document models can be learned from a training corpus [7]. However, we conducted our experiments using fixed model weights.

3 Experiments and Results

We implemented different retrieval systems and performed experiments on the Wall Street Journal (WSJ) subset of TREC4 test collection. The WSJ collection has 74,520 documents with around 250MB of data. TREC4 topic queries were used in the evaluation. SMLE(40) is a unigram language model that combines a document model with corpus model through linear interpolation of term probabilities with the weighting factor of 40% set to the document model. BG(40+10) is a bigram language model which combines bigram document model and smoothed unigram language model. The weighting parameter between document and corpus models in the unigram model is set to 40% and the weighting parameter for bigram document model set to 10%.

Biterm retrieval systems were implemented with different approximations for biterm probabilities. The interpolation parameters were set at 40% for document model over corpus model and 10% for biterm models over unigram models. BT1(40+10) uses average of bigram probabilities as given by (4). BT2(40+10) is the biterm language model using the *ad-hoc* probability given by (5) and BT3(40+10) corresponds to the biterm probability in (6). For comparison, the language model suggested by Ponte and Croft [9] was implemented (referred as PCLM).

Table 1 shows the comparison of non-interpolated average precision between different retrieval systems on the WSJ data set.

Method	AveP	%Change	%Change	%Change
SMLE(40)	0.2023	-	-	-
PCLM	0.2070	2.32	-	-
BG(40+10)	0.2229	10.18	7.68	-
BT1(40+10)	0.2176	7.56	5.12	-2.38
BT2(40+10)	0.2247	11.07	8.55	0.81
BT3(40+10)	0.2272	12.31	9.76	1.93

Table 1: Experimental results on the WSJ data set

Based on our experiments on a WSJ data set, the performance of different biterm approximations is around the

bigram language models. With fixed model weights, the *ad-hoc* approximations of biterm probabilities in biterm language model perform better than bigram language models. The reduction in average precision for BT1 in comparison to BG can be attributed to the averaging of probabilities reducing the effect of term pairs. The *ad-hoc* approximation P_{BT2} for biterms improves slightly over bigram language model since it ignores the order of term-occurrence while computing the frequency of term-pair occurrence.

If w_{i-1} occurs more often than w_i in the term pair $\{w_{i-1}, w_i\}$ in a document, the effect of the co-occurrence of term-pair to the biterm probability is reduced. Choosing the minimum of the term occurrence count for the denominator as in BT3 has improved performance than bigram language model. All approximations of biterm probabilities perform better than Ponte and Croft language model (PCLM) and smoothed unigram language model (SMLE(40)).

4 Conclusion

We present a novel method for generating language models for documents for information retrieval tasks. It is our belief that the assumptions for deriving language models for information retrieval are different from those for speech recognition or machine translation. Biterm language models capture term co-occurrence better than bigram models. Different approximations for biterm probabilities have been shown to provide better average precision than bigram language model. We have used *ad-hoc* approximations for biterm probabilities. We plan to explore better models to represent biterms. While constant weighting parameters were used in our experiments, the performance can be improved by learning and optimizing the combination weights.

REFERENCES

[1] A. Berger and J. D. Lafferty. Information retrieval as statistical translation. In *Proceedings of SIGIR'99*, pages 222–229, 1999.

[2] P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[3] M. F. Caropreso, S. Matwin, and F. Sebastiani. Statistical phrases in automated text categorization. Technical Report IEI-B4-07-2000, Istituto di Elaborazione dell'Informazione, Pisa, IT, 2000.

[4] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *European Conference on Digital Libraries*, pages 569–584, 1998.

[5] F. Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts, 1998.

[6] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR'01*, pages 111–119, 2001.

[7] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of SIGIR'99*, pages 214–221, 1999.

[8] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *Proceedings of RIAO-97, 5th International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 200–214, Montreal, CA, 1997.

[9] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR'98*, pages 275–281. ACM, New York, 1998.

[10] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of SIGIR'99*, pages 279–280, 1999.