

# Incorporating Query Term Dependencies in 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

## 1 Introduction

Recent advances in Information Retrieval are based on using Statistical Language Models (SLM) for representing documents and evaluating their relevance to user queries [6, 3, 4]. Language Modeling (LM) has been explored in many natural language tasks including machine translation and speech recognition [1]. 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 [3, 4], most experiments have used smoothed unigram language models that assume term independence for estimating document language models.

N-gram, specifically, bigram language models that capture context provided by the previous word(s) perform better than unigram models [7]. Biterm language models [8] that ignore the word order constraint in bigram language models have been shown to perform better than bigram models. However, word order constraint cannot always be relaxed since a *blind venetian* is not a *venetian blind*. Term dependencies can be measured using their co-occurrence statistics. Nallapati and Allan [5] represent term dependencies in a sentence using a maximum spanning tree and generate a sentence tree language model for the story link detection task in TDT. Syntactic parse of user queries can provide clues for when the word order constraint can be relaxed. Syn-

tactic structures have been utilized for improving N-gram models [2] targeted towards speech recognition. However, language modeling requirements of speech recognition differ from that of information retrieval [8]. We propose two novel methods for query likelihood retrieval that incorporate term dependencies in estimating the document language models.

## 2 Maximum Bigram Language Models

In query likelihood model, the bigram approximation of probability of relevance of a document,  $D$ , to a query,  $Q$ , is given by

$$P(Q|M_D) \approx \prod_i P(q_i|q_{i-1}, M_D). \quad (1)$$

The order in which the terms appear in a query is not necessarily the order in which they will appear in a relevant document. Since queries are not necessarily well-formed sentences and it is difficult to enumerate all possible dependencies between that occurrence of query terms in documents, we use the following approximation to estimate query term probabilities:

$$P(q_i|q_1, q_2, \dots, q_{i-1}, M_D) = \max_{j=i, j=1, \dots, n} P(q_i|q_j, M_D) \quad (2)$$

where the  $q_j$  corresponding to the maximum probability value can be viewed as the best trigger term for predicting  $q_i$ . In other words, assuming complete dependency between query terms, the dependency between  $q_i$  and  $q_j$  is more pronounced in the document model. We refer to this approximation as *maximum bigram probability* (MBG) of query term  $q_i$  in document  $D$ .

## 3 Incorporating Query Concepts in Language Models

Some phrases in the query that represent specific concepts do appear in that order in relevant documents. For such queries, the complete dependence assumption of maximum bigram language models should be relaxed. For the query *nuclear power plants*, relevant documents have exact match for this phrase or concept and documents that talk about *power plants* are not necessarily relevant to the topic. The query can be viewed as a concept sequence  $\{c_1, c_2, \dots, c_k\}$  where each  $c_j$  is a sequence of terms  $\{q_{i_1}^j, q_{i_2}^j, \dots, q_{i_k}^j\}$ . Then, the query likelihood probability is given by

$$P(Q|M_D) = \prod_j P(c_j|c_1, c_2, \dots, c_{j-1}, M_D) \quad (3)$$

$$\approx \prod_j P(c_j|M_D) \quad (4)$$

where (4) corresponds to a unigram model approximation referred to as *concept unigram language model*. Concepts can be identified using a syntactic parser. In our experiments we used a part-of-speech tagger and a shallow parser to parse the queries and group terms. The concept unigram probabilities are estimated using a smoothed bigram approximation given by

$$P(c_j|M_D) = P(q_1^j) \prod_{i=2}^{i_k} P(q_i^j|q_{i-1}^j, M_D) \quad (5)$$

## 4 Experiments and Results

We implemented different retrieval systems and performed experiments on the Wall Street Journal (WSJ) subset of TREC4 test collection (data size 250MB of 74,520 documents) and the TREC4 test collection (2GB of 567,529 documents). TREC4 topic queries were used in the evaluation. Table 1 summarizes average precision and R-precision values for different retrieval methods on the WSJ dataset. The numbers in brackets indicate percentage change over SMLE and BG(10) retrieval methods, respectively.

**Table 1: Experimental results: WSJ data set**

Method	AveP(%changes)	RPres(%changes)
SMLE	0.2282(-,-)	0.2395(-,-)
BG(10)	0.2447(7.23,-)	0.2475(3.34,-)
BT(10)	0.2506(9.82,2.41)	0.2449(2.25,-1.05)
MBG(10)	0.2406(5.43,-1.68)	0.2564(7.06,3.60)
CULM(10)	0.2512(10.08,2.66)	0.2543(6.18,2.75)

SMLE is a smoothed unigram language model using a Dirichlet prior [9] with parameter  $\mu$  set to 1000. Higher order models are smoothed using the Dirichlet smoothed unigram language models. BG(10) is a smoothed bigram language model which interpolates empirical bigram probability with smoothed unigram probability with the weighting parameter for bigram probabilities set at 10%. BT(10) is a smoothed biterm language model which uses the min-Adhoc approximation [8] given by

$$P_{BT}(w_i|w_{i-1}, D) \approx \frac{C(w_{i-1}, w_i|D) + C(w_i, w_{i-1}|D)}{2 * \min\{C(w_{i-1}|D), C(w_i|D)\}} \quad (6)$$

where  $C(w_l, w_k|D)$  is the occurrence count of the ordered word pair  $(w_l, w_k)$  in document  $D$ , and  $C(w_l|D)$  is the occurrence count of word  $w_l$  in  $D$ . MBG(10) is a maximum bigram language model which interpolates the empirical maximum bigram probability a term with its corpus smoothed unigram probability. Here the bigram probabilities have a weighting of 10% in the interpolation. CULM(10) corresponds to the concept unigram language with concept probabilities estimated using smoothed term bigram probabilities with bigrams weighted at 10%.

In Table 1, BT(10), MBG(10) and CULM(10) methods show significant improvements (9.82%, 5.43% and 10.08%, respectively) over baseline smoothed unigram language models. While the average precision performance is comparable for BG(10) and MBG(10), the R-precision values, which is biased towards the relevance decision made by the retrieval method on each document, show significant improvement of MBG(10) over BG(10). For larger TREC4 test collection (ref. Table 2), BT(10) and MBG(10) do not perform

**Table 2: Experimental results: TREC4 data set**

Method	AveP(%changes)	RPres(%changes)
SMLE	0.1876(-,-)	0.2475(-,-)
BG(10)	0.1901(1.33,-)	0.2465(-0.40,-)
BT(10)	0.1893(0.91,-0.42)	0.2442(-1.33,-0.93)
MBG(10)	0.1747(-6.87,-8.10)	0.2278(-7.96,-7.59)
CULM(10)	0.2020(7.68,6.26)	0.2589(4.61,5.03)

as well as bigram language model. However, the concept unigram language model consistently performs better than other models for both test collections. This indicates that improved retrieval performance is achievable by incorporating query term dependence in the estimation of document language models.

## 5 Conclusion

We have proposed two language models for information retrieval that incorporate term dependencies based on statistical (Maximum Bigram Language Models) and syntactic (Concept Language Models) information from user query. While better R-precision values are obtained for Maximum Bigram Language Models on the WSJ data set, the performance of Concept Unigram Language Models is consistently better than other models on both datasets. In future, we plan to explore higher order concept language models that exploit the context in which concepts appear in relevant documents.

## REFERENCES

- [1] E. Charniak. *Statistical Language Learning*. The MIT Press, Cambridge, Massachusetts, 1993.
- [2] C. Chelba and F. Jelinek. Exploiting Syntactic Structures for Language Modeling. In *Proceedings of the COLING-ACL Meeting*, pages 225–231, Montreal, Canada, 1998.
- [3] 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.
- [4] V. Lavrenko and W. B. Croft. Relevance-based Language Models. In *Proceedings of SIGIR*, pages 120–127. ACM, New York, 2001.
- [5] R. Nallapatti and J. Allan. Capturing Term Dependencies using a Sentence Tree based Language Model. In *Proceedings of CIKM'02*, pages 383–390, 2002.
- [6] 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.
- [7] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of SIGIR'99*, pages 279–280, 1999.
- [8] M. Srikanth and R. Srihari. Biterm Language Models for Document Retrieval. In *Proceedings of SIGIR*, pages 425–426. ACM, New York, 2002.
- [9] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'01*, pages 334–342, 2001.