Effective Use of Phrases in Language Modeling to Improve Information Retrieval

Maojin Jiang, Eric Jensen, Steve Beitzel
Information Retrieval Laboratory
Illinois Institute of Technology
{jianmao, ej, steve}@ir.iit.edu

Shlomo Argamon
Linguistic Cognition Group
Illinois Institute of Technology
argamon@iit.edu

Abstract

Traditional information retrieval models treat the query as a bag of words, assuming that the occurrence of each query term is independent of the positions and occurrences of others. Several of these traditional models have been extended to incorporate positional information, most often through the inclusion of phrases. This has shown improvements in effectiveness on large, modern test collections. The language modeling approach to information retrieval is attractive because it provides a well-studied theoretical framework that has been successful in other fields. Incorporating positional information into language models is intuitive and has shown significant improvements in several language-modeling applications. However, attempts to integrate positional information into the language-modeling approach to IR have not shown consistent significant improvements. This paper provides a broader exploration of this problem. We apply the backoff technique to incorporate a bigram phrase language model with the traditional unigram one and compare its performance to an interpolation of a conditional bigram model with the unigram model. While this novel application of backoff does not improve effectiveness, we find that our formula for interpolating a conditional bigram model with the unigram model yields significantly different results from prior work. Namely, it shows an 11% relative improvement in average precision on one query set, while yielding no improvement on the other two.

1. Introduction

Information retrieval traditionally views queries and documents as a bag of words [Salton75], implying that the occurrence of each term is independent from occurrences of all other terms. This is obviously inaccurate and it is easy to find examples of its failures. Financial documents, for example, would likely have many instances of the word "exchange" following the word "stock," whereas agricultural documents may talk more about the "exchange of livestock." Prior attempts at integrating phrases into other information retrieval models have shown improvements in retrieval effectiveness on modern test collections.

The language modeling approach to information retrieval ranks documents' similarity to queries by modeling them as statistical language models and calculating the probability of one given the other [Ponte98]. However, the traditional term independence assumption is typically applied. Attempts to incorporate phrases into language models have not shown consistent, significant improvements on modern test collections. These techniques have all been based on slightly different linear interpolations of unigram word probabilities with bigram phrase probabilities. We propose the backoff strategy, which has been successful in using language models for speech recognition [Katz87], as an attempt to garner the improvements seen from phrases in other retrieval strategies.

2. Prior Work

Prior work in this area consists of on the development of the language modeling approach to information retrieval, the incorporation of phrases into that methodology through interpolation, and the incorporation of phrases into other retrieval strategies.

2.1. Language Modeling in Information Retrieval

The language modeling approach to information retrieval ranks documents based on $p(d|q)$, the probability that a document generates an observed query. Since this is difficult to measure directly, however, Bayes Theorem is often applied and a document-independent constant is dropped (Equation 1).

$$p(d|q_{1...}q_n) \propto p(q_{1...}q_n|d)p(d)$$

Equation 1: Language Model Document Ranking

In practice $p(d)$, the prior probability that a document is relevant to any query, is assumed to be uniform. Also common in most work is that the next step: applying the bag of words assumption (Equation 2) by estimating the probability of the sequence as the product of the probabilities of the individual terms.

$$p(q_{1...}q_n|d) \approx \prod_i p(q_i|d)$$

Equation 2: Unigram Language Model

Since a document is a very sparse language model, the next necessary step is for this estimate to be smoothed to account for query terms unseen in the document. This is typically accomplished by incorporating the probability of the unseen term in the collection as a whole through one of the many smoothing methods available (such as the successful Dirichlet smoothing in Equation 3). Katz' backoff technique has also been compared to smoothing for the case of unseen terms and did not perform as effectively [Zhai01].

$$p_\mu(q_i|d) = \frac{C(q_i;d) + \mu_1 \cdot p_{MLE}(q_i|C)}{|d| + \mu_1}$$

$C(q_i;d)$ is the frequency of query term q_i in document d

$|d|$ is the number of terms in document d

$p_{MLE}(q_i|C)$ is the maximum likelihood estimate of the probability of q_i in the collection

Equation 3: Dirichlet Smoothed Unigram Model

2.2. Incorporating Phrases

There has been much work attempting to use phrases to enhance the effectiveness of information retrieval. Most often this does not show more than a 10% improvement over the simple bag of words approach, with improvements closer to 5% being more common Mitra97 Kraaij98 Turpin99 Narita00 Chowdhu01a. However, phrases consistently provide some improvement in retrieval strategies outside the language modeling paradigm, rarely harming retrieval performance.

Several methodologies for integrating phrases have been tried inside the language modeling framework. Although they are all based on a linear interpolation of bigram and unigram models, they differ slightly in formulation and significantly in their results, albeit on differing collections. Song and Croft used a linear interpolation of unigram and joint bigram probabilities[1] inside the document in combination with a linear interpolation for smoothing unseen unigrams by $p(q_i|C)$, the probability of a term in the corpus as a whole (Equation 4) Song99.

$$p(q_i|d) \approx \omega \cdot p(q_i|d) + (1-\varpi) \cdot p(q_i|C)$$
$$p(q_{i-1}, q_i|d) \approx \lambda \cdot p(q_{i-1}, q_i|d) + (1-\lambda) \cdot p(q_i|d)$$

Equation 4: Song and Croft Bigram Interpolation

In their experimentation, they empirically set $\varpi = .40$ and $\lambda = .90$ and saw a relative improvement in average precision over their smoothed unigram language model of 7% for the TREC 4 queries on the Wall Street Journal collection and less than 1% on the entire TREC 4 collection.

Miller, Leek, and Schwartz used a single three-way linear interpolation of unigram and conditional bigram document probabilities and unigram collection probabilities (Equation 5) Miller99.

$$p(q_{1\ldots}q_n|d) \approx \prod_i (a_0 \cdot p(q_i|C) + a_1 \cdot p(q_i|d) + a_2 \cdot p(q_i|q_{i-1}, d))$$

Equation 5: Miller, Leek, and Schwartz Bigram Interpolation

[1] The use of the joint probability here is unclear, as we would not expect the unigram probability to be a reasonable estimate for the bigram joint probability.

They empirically set $a_0 = .70$, $a_1 = .29$, and $a_2 = .01$ and saw less than 5% relative improvement in average precision over their smoothed unigram language model on the TREC 6 and 7 queries over the 2GB SGML collection from TREC disks 4 & 5. Hiemstra describes a similar strategy for integrating proximity into his hidden Markov model framework [Hiemstra01].

All of these approaches treat phrases as lower-weighted units while counting the terms making up the phrases at a higher weight. This is often explained heuristically by citing the high weights associated with phrases due to their rarity in the collection (their collection probabilities are much smaller than their document probabilities) and the corresponding need to normalize for this. Work outside of the language modeling framework shows that phrase weighting based simply on query lengths can be as effective as static weights empirically tuned and tested on the same 50 TREC queries [Chowdhu01a].

Interactive work with users manually selecting phrases from the lexicon to expand their queries has suggested that the addition of certain phrases can significantly improve average precision [Smeaton98]. The authors suggest that since phrases have very different frequency distributions than individual terms, integrating phrases with the unigram words used in queries may require two separate term-weighting functions and thus a more fusion-centered approach similar to what has been used to combine other forms of multiple query representations [Chowdhu01b].

3. Methodology

We further explore the problem of integrating phrases into language models for information retrieval by applying the backoff technique that has been successful in other applications. As our baseline, we propose an alternative interpolation of conditional bigram and unigram models that seeks to maintain a separation between the bigram interpolation parameter and the unigram smoothing parameter (Equation 6).

$$p(q_i|q_{i-1}, d) \approx \lambda \cdot p(q_i|q_{i-1}, d) + (1 - \lambda) \cdot p(q_i|d)$$

Equation 6: Conditional Bigram Interpolation

Note that this technique counts unigram influence in combination with bigram influence for its final estimation. Our backoff strategy uses unigram term probabilities only when those terms do not form a phrase that appears in the document (Equation 7).

$$p(q_{i-1}, q_i|d) \approx \begin{cases} p_{dml}(q_{i-1}, q_i|d) & \text{if } q_{i-1}, q_i \in d \\ \alpha_d \cdot p(q_{i-1}|d) \cdot p(q_i|d) & \text{otherwise} \end{cases}$$

Equation 7: Backoff from Bigrams to Unigrams

In fact, the amount of the overall probability space dedicated to unigram probabilities is explicitly defined by the document-dependent constant given in Equation 8, which is determined by $p_{dml}(q_{i-1}, q_i|d)$, the discounted maximum likelihood estimate specific to the discounting method used [Zhai01]. In this work, we use the Dirichlet discounting method given by Equation 9.

$$\alpha_d = \frac{1 - \sum_{w_1 w_2 \in D} p_{dml}(w_1 w_2 \mid D)}{\sum_{w_1 w_2 \notin D} p(w_1 \mid D) \cdot p(w_2 \mid D)}$$

Equation 8: Probability space reserved for unseen bigrams in each document

$$p_{dml}(q_{i-1}, q_i|d) = \frac{C(q_{i-1}, q_i; d)}{|d| - 1 + \mu_2}$$

Equation 9: Dirichlet Discounted Maximum Likelihood Joint Bigram Model

In order to practically use this backoff formulation, we must deal with joint bigram probabilities by converting them to conditional probabilities in order to combine them term-by-term as in Equation 10.

$$p(q_{1\ldots}q_n|d) \approx p(q_1 \mid d) \cdot \prod_{i=2}^{n} \frac{p(q_{i-1}q_i|d)}{p(q_{i-1} \mid d)}$$

Equation 10: Converting Joint Bigram Probabilities to Conditionals for Combination

Note that using interpolation parallels previous methods of incorporating phrases (inside and outside of the language modeling framework) by regarding phrase matches as an added boost to their component term matches, whereas backoff stresses phrase matches by excluding influence of their component terms. As such the method of phrase selection used with backoff may have a more significant impact than seen in prior studies [Mitra97] [Kraaij98]. In this work, we simply use all sequentially appearing pairs of terms (statistical phrases), as a basic bigram model would be defined. Linear interpolation should perform well when phrases consist of terms that carry their own meaning, such as "academic journal", whereas backoff should perform well when phrases consist of terms that should effectively be stopped if they are inside the phrase such as "New York".

4. Experimentation

We use the TREC 6, 7, and 8 short title (1-4 terms) queries' against the TREC disks 4 and 5 2GB collection of SGML (primarily news) data with our lab's search engine, AIRE, to examine average precision when including the bigram distribution through each of our methods. The only retrieval utilities we use are a 342-word stop list from Cornell's SMART search engine, and a hand-built list of conflation classes in favor of a stemmer. For each model, we empirically find a semi-optimal set of smoothing parameters, publishing these tuning experiments to illustrate stability of smoothing and combination parameters across varying models. We first tune our unigram smoothing

parameter for each set of queries. We then incorporate bigram probabilities through linear interpolation as in Equation 6, and also through backoff as in Equation 7.

4.1. Smoothed Unigram Language Model

As it has been shown to perform well [Zhai01], our initial baseline is a unigram model with Dirichlet smoothing using the corpus unigram probabilities (Equation 3). The optimal parameter value for each set of 50 queries is shown in bold.

Table 1: Avg. Precision for Varied Unigram Smoothing Parameter

μ_1	TREC-6 301-350	TREC-7 351-400	TREC-8 401-450
100	15.13	9.57	14.52
200	**15.50**	9.84	14.81
300	15.48	10.00	15.19
350	15.46	10.04	**15.25**
500	15.38	10.30	15.06
800	15.18	10.41	14.94
1000	14.86	10.45	14.85
1500	14.86	**10.49**	14.58
2000	14.95	10.44	14.47
3000	14.18	10.45	14.18

4.2. Interpolated Bigram and Unigram

As a phrase baseline, we add document phrase probabilities to this through linear interpolation as per Equation 6, using the maximum likelihood estimate for bigram probabilities and Dirichlet smoothing (Equation 3) for unigram probabilities. In order to examine the interplay between smoothing parameter and interpolation parameter, we present results for both the optimal smoothing parameter and a common one (3000) for each query set. As expected, the optimal interpolation parameter does not seem to change when modifying the smoothing parameter. Note that as λ approaches zero, interpolation reduces to the simple unigram model, yielding equivalent performance. Neither TREC 6 nor TREC 8 show any improvement over the unigram model, while TREC 7 shows an 11% relative improvement over the optimal unigram model when phrases weighted quite low. Also, the range of interpolation parameters that provide any improvement at all (between 0.0001 and 0.01) is quite small, but it is similar to the optimal parameter found by Miller, et. al in their interpolation model.

Table 2: Avg. Precision for Varied Bigram Interpolation Parameter

λ	TREC-6 301-350 $\mu_1 = 200$	TREC-6 301-350 $\mu_1 = 3000$	TREC-7 351-400 $\mu_1 = 1500$	TREC-7 351-400 $\mu_1 = 3000$	TREC-8 401-450 $\mu_1 = 350$	TREC-8 401-450 $\mu_1 = 3000$
0.00005	**15.5**	14.05	10.57	10.51	**15.25**	**15.25**
0.0001	**15.5**	14.07	10.63	10.53	**15.25**	**15.25**
0.001	15.43	14.24	11.05	11.31	**15.25**	14.63
0.005	15.40	**14.73**	**11.66**	**11.48**	15.15	14.79
0.01	15.38	14.69	11.59	11.32	15.09	14.81
0.1	15.02	14.14	10.52	10.53	14.16	13.71
0.4	14.47	13.85	10.11	10.12	13.75	13.44
0.6	14.36	13.83	10.03	9.95	13.74	13.44

4.3. Backoff from Bigram to Unigram

Finally, we examine our backoff strategy. Again, we offer results for both the optimal unigram smoothing parameter and a fixed one. Unfortunately, we see no improvement for any query set. In TREC 7 and TREC 8, performance is impaired by approximately 10%. In terms of the discount parameter, performance is much more stable with a wide range of values than that of interpolation, indicating that we likely did not simply miss the optimal parameter. The large magnitude of the optimal discount parameter, devoting more of the probability space to the unigram model we back off to, may indicate that using phrases without incorporating their component unigram probabilities is unwise. We hypothesize that the reason performance does not approach that of the unigram model even when the parameter is drastically large is due to the additional document length factor that is introduced by the document-dependent backoff constant. As with interpolation, the optimal discount parameter does not seem to change significantly when modifying the smoothing parameter.

Table 3: Avg. Precision for Varied Bigram Backoff Discount Parameter

μ_2	TREC-6 301-350 $\mu_1 = 200$	TREC-6 301-350 $\mu_1 = 500$	TREC-7 351-400 $\mu_1 = 1500$	TREC-7 351-400 $\mu_1 = 500$	TREC-8 401-450 $\mu_1 = 350$	TREC-8 401-450 $\mu_1 = 500$
1000	14.89	14.56	9.16	9.25	12.79	13.39
3000	14.20	14.65	9.26	9.66	13.68	12.79
5000	**15.05**	14.67	**9.68**	**9.69**	**13.71**	13.71
10000	14.77	**14.82**	9.65	9.37	13.59	13.78
50000	14.24	14.02	9.70	9.47	13.79	13.88
90000	13.71	13.94	9.61	9.37	13.65	13.90
100000	13.67	13.67	9.56	9.31	13.59	**14.00**
110000	13.62	13.74	9.49	9.26	13.52	13.95
150000	13.24	13.23	9.52	9.05	13.20	

5. Conclusion and Future Work

We have explored using the backoff technique to incorporate phrases into language models for information retrieval. This technique has proved unsuccessful when using all sequentially appearing pairs of terms as phrases. We have also shown that phrases can help significantly when interpolating them as conditional bigram probabilities with the unigram model. Future work will examine intelligently selecting phrases when applying the backoff strategy and performing an in-depth analysis of why phrases help so much on one set of topics but do not show any improvement on others.

Salton75 G. Salton, C.S. Yang, and A. Wong,. A Vector-Space Model for Automatic Indexing. In *Communications of the ACM* 18(11):613-620, 1975.

Ponte98 J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *21st ACM Conference on Research and Development in Information Retrieval (SIGIR'98)* 275-281, 1998.

Katz87 S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35:400-401, 1987.

Zhai01 C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *24th ACM Conference on Research and Development in Information Retrieval (SIGIR'01)* 334-342, 2001.

Mitra97 M. Mitra, C. Buckley, A. Singhal, C. Cardie. An Analysis of Statistical and Syntactic Phrases. In *5th International Conference Recherche d'Information Assistee par Ordinateur (RIAO'97)*, 1997.

Kraaij98 W. Kraaij and R. Pohlmann. Comparing the effect of syntactic vs. statistical phrase index strategies for dutch. In C. Nikolaou and C. Stephanidis (Eds.) *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL'98)* 605-617.

Turpin99 A. Turpin and A. Moffat. Statistical phrases for Vector-Space information retrieval. In *22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99)* 309-310, 1999.

Narita00 M. Narita and Y. Ogawa. The use of phrases from query texts in information retrieval. In *23rd ACM Conference on Research and Development in Information Retrieval (SIGIR'00)* 318-320, 2000.

Chowdhu01a A. Chowdhury. Adaptive Phrase Weighting. In *International Symposium on Information Systems and Engineering (ISE 2001)*, 2001.

Song99 F. Song and W. B. Croft. A general language model for information retrieval. In *Eighth International Conference on Information and Knowledge Management (CIKM'99)*, 1999.

Miller99 D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden Markov model information retrieval system. In *22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99)* 214-221, 1999.

Hiemstra01 D. Hiemstra. Using language models for information retrieval. Center for Telematics and Information Technoloy, 2001. VIII, 164 p. : ill. ; 25 cm. - (CTIT Ph.D thesis series, ISSN 1381-3617 ; no. 01-32), 2001.

Smeaton98 A. F. Smeaton and F. Kelledy. User-chosen phrases in interactive query formulation for information retrieval. In *20th BCS-IRSG Colloquium, Springer-Verlag Electronic Workshops in Computing*, 1998.

Chowdhu01b A. Chowdhury, S. Beitzel, E. Jensen. Analysis of Combining Multiple Query Representations in a Single Engine. In *2002 IEEE International Conference on Information Technology - Coding and Computing (ITCC)*, 2002.