Choosing the Right Bigrams for Information Retrieval

Maojin Jiang, Eric Jensen, Steve Beitzel, and Shlomo Argamon

Information Retrieval Laboratory & Laboratory of Linguistic Cognition,
Computer Science Department
Illinois Institute of Technology,
10 W. 31st Street, Chicago, IL 60616
jianmao@ir.iit.edu, ej@ir.iit.edu, steve@ir.iit.edu, argamon@iit.edu

Abstract

After more than 30 years of research in information retrieval, the dominant paradigm remains the "bag-of-words", in which query terms are considered independent of their coocurrences with each other. Although there has been some work on incorporating phrases or other syntactic information into IR, such attempts have given modest and inconsistent improvements, at best. This paper is a first step at investigating more deeply the question of using bigrams for information retrieval. Our results indicate that only certain kinds of bigrams are likely to aid retrieval. We used linear regression methods on data from TREC 6, 7, and 8 to identify which bigrams are able to help retrieval at all. Our characterization was then tested through retrieval experiments using our information retrieval engine, AIRE, which implements many standard ranking functions and retrieval utilities.

1 Introduction

For the most part, information retrieval (IR) has traditionally viewed queries and documents as bags of words [SYW75], treating each term as independent of all other terms. Despite the obvious shortcomings of this approach, in that a great deal of language's meaning is carried in the co-occurrence and order of words, the "bag-of-words" (BOW) approach is still dominant, after more than 30 years of IR research. There are several reasons for this. One is that, with sufficiently sophisticated term-weighting, BOW works surprisingly well. Furthermore, users appreciate the 'convenience' of just throwing a set of relevant keywords to create a query; without a convincing improvement in results, users are unlikely to spend time thinking about the structure of the query.

And third, work at incorporating phrases or other syntactic information into IR systems has given modest and inconsistent improvement, at best.

This paper is a first step at investigating more deeply the question of using *phrases* in IR. Use of phrases typically means using term *bigrams* (sequential pairs) in addition to *unigrams* (individual terms). Previous work on using query bigrams for retrieval has given modest improvements, and often highly inconsistent performance (as we describe in more detail below). Our hypothesis is that only certain kinds of bigrams are likely to aid retrieval; by identifying the class of 'good bigrams', we hope to improve retrieval effectiveness by using only the good bigrams in the retrieval process. Surprisingly, there have apparently not been any studies on this question previously. We first compared 'super-optimal' retrieval effectiveness with and without bigrams for the Ad-Hoc Retrieval tracks of TREC 6, 7, and 8, and studied the ranking of the documents produced by our learner in order to identify where we expect bigrams to be at all helpful. A coarse characterization is that Classifier-Thing bigrams ought to be helpful in most cases (though there are other useful bigrams as well). We test this characterization through retrieval experiments using our information retrieval engine, AIRE, which implements many standard ranking functions and retrieval utilities [CBJSGF00].

2 Prior Work

The many attempts at integrating phrases, whether they be linguistic or statistical, into existing retrieval frameworks have shown at best disappointingly small improvements over the simple bag of words model. This is in contrast with many natural language processing tasks in which context around a word has been shown to significantly improve effectiveness (speech recognition, part-of-speech tagging, etc.). However, in addition to the intuitive motivation that phrases should aid retrieval effectiveness, interactive work with users manually selecting phrases to expand their queries has suggested that the addition of certain phrases can significantly improve average precision [SK98]. There are two typical explanations for the failure of phrases to improve effectiveness of ad-hoc information retrieval. First, incorporating phrase frequencies with word frequencies when ranking documents is a challenge due to their differing distributions. Second, it is difficult to determine which query phrases will improve performance and which will degrade it. In addition to these practical issues, incorporating phrases into existing IR models is often difficult to formally justify as the words composing them are naturally correlated with each other, violating the independence assumptions on which many models are based.

Most IR ranking functions, including the relatively new language modeling approaches, can be shown to be similar in that they rank documents via a linear combination of term (word or phrase) weights [HV00]. Much work has been devoted to combining phrases with terms inside of the same model

[Strz99]. In order to compensate for the vastly greater rarity of phrases than terms across the collection (typically resulting in much larger term weights), these approaches often discount the weights of phrases for all queries by a heuristically-set constant factor. Recent work has shown that dynamically computing phrase weights based on query length can be as effective as static weights empirically tuned and tested on the same query set [Chow01]. Studies incorporating context into language models for information retrieval typically interpolate the conditional probability of each query term given the previous one with its unigram probability, performing no phrase selection. Although intuitive, this gives no significant improvement over baseline unigram models and often yields unintuitive optimal interpolation parameters, with bigrams having only minimal weight [JJBA04] [SC99] [MLS99] [Hiem01].

Previous approaches for selecting phrases have been either statistically or syntactically motivated. Mitra selected only those phrases that appear at least 25 times in the corpus for inclusion in the vector-space model and saw no significant improvement [MBSC97]. Turpin further examined these results, trying many permutations of topics with varying length and phrase selection techniques, and also could find no significant improvement from statistical phrases [TM99]. Attempts to incorporate syntactic phrases date to the beginning of information retrieval itself [Salt68]. In a recent study, Voorhees analyzes the consistent failure to produce improvement of many attempts to integrate NLP techniques with the statistical methods widely used by document retrieval systems [Voor99]. She observed that these studies often produce inconsistent results, e.g. [Faga87]; quite often there is improvement for some topics and a reduction in effectiveness for others. Arampatzis and colleagues proposed a framework for information retrieval that incorporates linguistically selected phrases [AWKB98] [AWKB00]. They found that exploiting co-occurrence of noun-phrases that contain query terms could improve recall, but precision dropped significantly. Zhai, et. al, selected noun phrases where any ambiguity is resolved through the statistical addition of structure and found improvements in precision for some topics, but damaged performance on others [ZTME96]. Narita and Ogawa also examined the use of noun phrases for ad-hoc retrieval and saw no significant improvement in overall average precision [NO00]. Lewis and Croft clustered syntactic phrases to group redundant phrases in an attempt to mediate the phrase sparseness problem, but found only small improvements in ad-hoc retrieval [LC90]. Kraaij and Pohlmann experimented with both statistical and syntactic phrases and found that neither significantly improved effectiveness, and often performed equivalently [KP98].

Text categorization often employs machine learning algorithms using features mined from unstructured text. Similarly to most IR ranking algorithms, many learning algorithms employ a linear combination of weighted feature values. Although integration of phrase features has been slightly more effective in categorization than ad-hoc retrieval, improvements are still disappointingly low. Koster and Seutter compared several combinations of head-modifier phrases with a word-only baseline and found that providing phrase features

to the learning algorithm did not improve effectiveness in text categorization [KS03].

3 Methodology

In order to differentiate between different bigrams for their possible usefulness in information retrieval, we constructed an experiment to compare unigram and bigram retrieval under "more optimal than optimal" conditions, with the additional goal of doing so in a system-independent fashion. We constructed a 'pseudo-ranking function' for each query by performing linear regression from a set of query-dependent parameters of each document (described below) to the values 0 or 1, depending on whether or not the document was prejudged to be relevant to that document (we used the TREC 6, 7, and 8 queries and relevance judgements). The ranking thus produced can then be evaluated on the document set (which was used to compute the regression function) for precision. Comparison of the resultant document rankings between using just unigrams or using unigrams and bigrams thus gives an optimistic measure of the potential contribution of bigrams to the retrieval process.

Our methodology is as follows. Given a document d and query q, we compute a set of unigram parameters $u_i = f(q_i, d)$, one for each word in the query, as well as a set of bigram parameters $b_i = f(q_i - 1, q_i, d)$, one for each bigram in the query. Given this representation, and assuming a particular query, each document in the collection is then represented by either a unigram vector, $U_d = [u_1 \dots u_n]$, or a bigram vector, $B_d = [u_1, \dots, u_n, b_2, \dots, b_n]$. Note that we only considered bigrams appearing 25 times or more in the collection, as done in prior statistical phrasing approaches [MBSC97]. We take all relevant documents for the query along with 3 times as many top-ranked non-relevant documents to produce a document set C and then compute the following two weight vectors minimizing the sum-of-squares difference between weight-vector dot-products and the binary relevance judgements R_d:

$$w_u = argmin_w \sum_{d \in C} (w^T U_d - R_d)^2 \tag{1}$$

$$w_b = argmin_w \sum_{d \in C} (w^T B_d - R_d)^2 \tag{2}$$

Note that the vectors have different dimensions. Each such weight vector can then be used to rank all the documents in the collection by their 'estimated relevance' to the query; a higher dot-product indicates more likely relevance. This use of linear regression is meant to simulate, on a coarse scale, the action of a 'typical' IR system, whose ranking functions can be formulated as nearly a linear function of parameters of query word occurrences in target documents (see [HV00]).

For generality, we used three different methods of computing the parameters. First was to use raw count of the number of each n-gram occurring in the given document (**cnt**). The second method was to use a Dirichlet-smoothed unigram language model (with the μ parameter set to 3000 as recommended in [ZL01]) for single words, and a maximum likelihood bigram model for phrases to get a 'probability' of the term given the document (as used in language modeling), which we term here *prob*. And third was a logarithmically scaled inverse probability (to avoid $log(0)$), termed *log*, computed as $log(1 - prob)$. Retrieval effectiveness was measured using two standard techniques: average precision which averages precision at 11 points from 0% to 100% recall, and R-precision which is the retrieval precision for the top r documents, where r is the total number of relevant documents for the query. The potential improvement of using bigrams together with unigrams was measured by relative precision improvement, I_prec , defined as $I_{prec} = (prec_b - prec_u)/prec_u$. A query was considered bigram-good if I_{prec} for the query was non-negative for all three parameter types under both effectiveness metrics, and positive for at least one parameter type under each effectiveness metric; a query is bigram-bad if I_{prec} was negative for at least one type under each metric and not positive for any; otherwise, a query is considered bigram-neutral.

4 Results

Table 1 shows the bigram-good queries from TRECs 6, 7, and 8, with their I_{prec} values for different parameters, while Tables 2 and 3 show the bigram-bad and bigram-neutral queries respectively. An examination of the queries in Table 1 reveals that many of the queries contain a bigram whose first word is a Classifier for the second, which is a Thing (and the head of a nominal phrase). These are functional roles within a nominal group structure, as analyzed in Systemic Functional Linguistics [Hall94]; a Classifier is a nominal modifier which effectively narrows the domain of reference to a subcategory of the category indicated by the nominal head. For example, "airport" in the phrase "airport security" (contrast with "national security"). Not all modifiers are classifiers, however; for example, "tight" in "tight security" is an Epithet, describing an attribute of the Thing, as shown by the fact that it can be intensified ("very tight security") and can be used to modify a variety of subtypes (i.e., we can have "tight airport security" as well as "tight national security"). The few queries without Classifier-Thing bigrams are all special cases, for which individual explanations can be easily found why two of the words would appear in the same irrelevant document, but only together in a relevant document. For example "Iran" and "Iraq", since the are in the same region, may tend to be mentioned in the same articles about the Middle East, even ones not about relations between the countries; however, when they are mentioned in succession, the likelihood of relevance to cooperative relations between the countries is much higher.

Table 1. Bigram-good queries from TREC 6, 7, and 8. Columns give Iprec for different measures: $P(cnt)$ for average precision on **cnt**, $R(cnt)$ for R-precision on **cnt**, and so forth. Maximum I_{prec} is boldfaced for each query, as are "Classifier-Thing" bigrams.

Query	P(cnt)	P(log)	P(prob)	R(cnt)	R(log)	R(prob)	Query terms
301	0.24%	**0.75%**	0.05%	2.21%	**6.60%**	0.00%	International **Organized Crime**
304	0.03%	3.60%	**3.77%**	0.00%	28.57%	**33.31%**	**Endangered Species** (Mammals)
306	**0.92%**	0.76%	0.01%	2.90%	**5.73%**	0.00%	African **Civilian Deaths**
313	1.85%	9.40%	**10.36%**	5.13%	19.49%	**22.50%**	**Magnetic Levitation**-Maglev
314	**1.65%**	0.05%	0.00%	**16.65%**	0.00%	0.00%	**Marine Vegetation**
315	**4.37%**	1.88%	0.00%	**19.05%**	4.02%	0.00%	Unexplained **Highway Accidents**
317	**11.96%**	5.66%	5.66%	**12.51%**	0.00%	0.00%	**Unsolicited Faxes**
321	**5.05%**	1.54%	0.12%	**22.86%**	4.62%	0.00%	Women in Parliaments
326	**5.72%**	2.58%	2.41%	**6.46%**	0.00%	0.00%	**Ferry Sinkings**
328	**7.69%**	0.00%	0.00%	**14.28%**	0.00%	0.00%	Pope Beatifications
330	2.84%	**7.85%**	4.78%	2.49%	**7.32%**	2.32%	Iran-Iraq Cooperation
331	**2.20%**	1.65%	1.83%	4.44%	**6.38%**	**6.38%**	**World Bank** Criticism
332	0.15%	0.41%	**0.82%**	0.00%	**2.08%**	**2.08%**	**Income Tax Evasion**
336	0.00%	**3.33%**	**3.33%**	0.00%	**9.09%**	**9.09%**	**Black Bear** Attacks
339	**9.87%**	0.00%	0.00%	**25.00%**	0.00%	0.00%	Alzheimer's **Drug Treatment**
341	0.49%	**4.87%**	0.37%	0.00%	**4.65%**	0.00%	**Airport Security**
343	0.34%	**0.50%**	0.22%	**2.99%**	0.00%	0.00%	**Police Deaths**
346	1.60%	**1.66%**	0.72%	**2.32%**	2.26%	0.00%	**Educational Standards**
350	0.27%	**7.36%**	1.46%	7.15%	**13.34%**	0.00%	Health and **Computer Terminals**
352	0.00%	**1.19%**	0.00%	0.00%	**9.38%**	0.00%	**British Chunnel** impact
357	**6.64%**	3.88%	2.70%	**34.34%**	10.06%	2.32%	**territorial waters** dispute
358	**44.29%**	33.41%	26.68%	**41.17%**	32.35%	21.63%	**blood-alcohol** fatalities
359	0.64%	**2.29%**	1.36%	**4.54%**	0.00%	**4.54%**	**mutual fund** predictors
360	0.70%	**0.98%**	0.26%	**2.21%**	**2.21%**	0.00%	**drug legalization** benefits
365	**2.27%**	0.33%	0.33%	**2.94%**	0.00%	0.00%	**El Nino**
369	0.00%	**5.34%**	**5.34%**	0.00%	**16.68%**	**16.68%**	**anorexia nervosa** bulimia
372	0.70%	**2.75%**	2.18%	2.78%	0.00%	**2.86%**	**Native American** casino
376	5.88%	**6.43%**	2.27%	**12.90%**	12.50%	0.00%	**World Court**
377	**1.38%**	0.24%	0.25%	**14.82%**	3.70%	3.70%	**cigar smoking**
384	**15.19%**	7.93%	5.59%	**14.29%**	6.99%	4.56%	**space station** moon
385	**1.27%**	0.77%	0.56%	**2.38%**	0.00%	0.00%	**hybrid fuel** cars
386	0.58%	**2.76%**	0.00%	0.00%	**6.66%**	0.00%	teaching **disabled children**
396	**3.32%**	1.73%	1.34%	**4.88%**	2.45%	2.45%	**sick building** syndrome
398	**1.49%**	0.96%	0.02%	**5.27%**	2.38%	0.00%	dismantling **Europe's arsenal**
404	**0.82%**	0.07%	0.07%	**2.10%**	0.00%	0.00%	Ireland, **peace talks**
406	**9.73%**	8.36%	8.36%	**20.01%**	**20.01%**	**20.01%**	**Parkinson's disease**
408	**16.42%**	10.07%	10.23%	**41.95%**	25.00%	25.00%	**tropical storms**
412	**5.99%**	1.83%	1.83%	**12.52%**	2.29%	2.29%	**airport security**
413	0.00%	**1.94%**	**1.94%**	0.00%	**2.13%**	**2.13%**	**steel production**
415	**10.67%**	1.53%	1.55%	**25.63%**	0.00%	0.00%	drugs, **Golden Triangle**
423	**0.88%**	0.58%	0.58%	0.00%	**6.25%**	**6.25%**	Milosevic, Mirjana Markovic
430	**4.17%**	8.33%	8.33%	**20.00%**	**20.00%**	**20.00%**	**killer bee** attacks
438	0.00%	**0.58%**	**0.58%**	0.00%	**2.05%**	**2.05%**	tourism, increase
440	**16.06%**	8.06%	8.06%	**13.32%**	9.67%	9.67%	**child labor**
441	**22.03%**	6.35%	6.35%	**45.45%**	9.09%	9.09%	**Lyme disease**
447	**11.00%**	0.00%	0.00%	**7.14%**	0.00%	0.00%	**Stirling engine**
450	**0.72%**	0.31%	0.32%	**2.44%**	0.00%	0.00%	King Hussein, peace

An examination of Table 2 allows us to refine this hypothesis somewhat. Even though each of the bigram-bad queries contains a Classifier-Thing bigram, such bigrams are not central to the meaning of the query. To understand this, consider first the query "Legionnaire's disease". It is highly unlikely that any document in the collection contains the word "Legionnaire" while not being about this disease. Similarly for "Schengen" or "obesity". In the case of "encryption equipment export", we find that "encryption export" is also an excellent query. So to be more precise, we believe that queries that contain Classifier-Thing bigrams that contrast with other Classifiers for the same Thing in the corpus will be useful bigrams for retrieval. We validated this method by examining retrieval performance with only words, all phrases, and only our list of (hand-chosen) Classifer-Thing phrases when using a common, highly effective retrieval strategy, Robertson's probabilistic model BM25 [RWBGP95]. Phrases were weighted such that their scores counted for only .25 of terms' scores. When using the list, phrases not appearing in our list of relevant phrases did not contribute any weight to a document. We used a hand-tailored set of conflation classes for stemming [XC98] and the 342-word stop list from Cornell's SMART system [Corn04]. A summary of these results is given in Table 3.

Table 2. Bigram-bad queries from TREC 6, 7, and 8. Columns give I_{prec} for different measures: P(cnt) for average precision on cnt, R(cnt) for R-precision on cnt, and so forth. Lowest I_{prec} is boldfaced for each query, as are "Classifier-Thing" bigrams.

Query	$P(cnt)$	$P(log)$	$P(prob)$	$R(cnt)$	$R(log)$	$R(prob)$	Query terms
373	-1.16%	**-2.05%**	0.00%	0.00%	**-4.00%**	0.00%	**encryption equipment** export
380	**-3.55%**	-2.02%	-2.02%	**-16.66%**	0.00%	0.00%	obesity **medical treatment**
410	0.00%	0.00%	**-61.54%**	0.00%	0.00%	**-80.01%**	**Schengen agreement**
421	-1.49%	**-1.50%**	**-1.50%**	**-2.05%**	-1.99%	-1.99%	**industrial waste disposal**
429	**-6.38%**	-2.49%	-2.49%	**-18.18%**	-9.09%	-9.09%	**Legionnaires' disease**

Table 3. Bigram improvement for average precision and R-precision, using BM25 retrieval. Averages and standard deviations are shown for bigram-good, bigram-bad, and all queries in TRECs 6, 7, and 8.

Query Type	**Measure**	**min**	**max**	**avg**	**stderr**
good	avg. prec.	-70%	2827%	**113%**	74%
good	R-prec	-100%	400%	**12%**	11%
bad	avg. prec.	-30%	400%	**74%**	81%
bad	R-prec	-13%	20%	**1.5%**	5.2%
all	avg. prec.	-82%	2827%	**42%**	23%
all	R-prec	-100%	400%	**5.3%**	4.0%

These results clearly confirm that the set of queries singled out by our 'overly optimal' linear regression technique as where bigrams may possible be useful, indeed get higher levels of retrieval improvement when using bigrams. The results for bigram-bad queries, however, are misleading, since the strongly positive results are due to a single query, "industrial waste disposal", while all other bad queries give neutral or negative effect from using bigrams in retrieval. Indeed, this query fits our proposed pattern of "Classifer-Thing" bigrams, containing two of them.

5 Discussion and Future Work

We have proposed a system-independent methodology for determining which bigrams are likely to be useful for retrieval, and have validated the methodology by showing that those queries our method shows to be good candidates for bigram use indeed get higher improvements from using bigrams than other queries. We examined the phrases in the queries that improved and concluded that an important characteristic of "good" bigrams (for retrieval purposes) is that they are "Classifier-Thing" pairs, in which the first word effectively selects for a subclass of the type referred to by the second word. At the same time, there are a few other interesting types of bigrams which are more difficult to characterize directly in this way. Future work will include devising and evaluating methods for automatically determining the good bigrams without the use of relevance judgments, as well as incorporating such selective bigram use in our retrieval system.

References

1. Arampatzis, A. T., T. P. van der Weide, C. H. A. Koster and v. Bommel (1998). "Phrase-based Information Retrieval." *Information Processing and Management* **34** (6): 693-707.
2. Arampatzis, A. T., T. P. van der Weide, C. H. A. Koster and v. Bommel (2000). *An Evaluation of Linguistically-motivated Indexing Schemes.* BCSIRSG '2000.
3. Chowdhury, A., S. Beitzel, E. Jensen, M. Saelee, D. Grossman and O.Frieder (2000). *IIT-TREC-9 - Entity Based Feedback with Fusion.* Ninth Annual Text Retrieval Conference, NIST.
4. Chowdhury, A. (2001). *Adaptive Phrase Weighting.* International Symposium on Information Systems and Engineering (ISE 2001).
5. Cornell Cornell University - SMART - ftp://cs.cornell.edu/pub/smart.
6. Fagan, J. (1987). *Automatic phrase indexing for document retrieval.* 10th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'87), New Orleans, Louisiana.
7. Halliday, M. A. K. (1994). *An Introduction to Functional Grammar.* London, Edward Arnold.
8. Hiemstra, D. and A. d. Vries (2000). Relating the new language models of information retrieval to the traditional retrieval models, Centre for Telematics and Information Technology.
9. Hiemstra, D. (2001). *Using language models for information retrieval, Center for Telematics and Information Technology*: 164.
10. Jiang, M., E. Jensen, S. Beitzel and S. Argamon (2004). *Effective Use of Phrases in Language Modeling to Improve Information Retrieval.* 2004 Symposium on AI & Math Special Session on Intelligent Text Processing, Florida.
11. Koster, C. H. A. and M. Seutter (2003). *Taming Wild Phrases.* ECIR'03.
12. Kraaij, W. and R. Pohlmann (1998). *Comparing the effect of syntactic vs. statistical phrase index strategies for dutch.* 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL'98).
13. Lewis and Croft (1990). *Term Clustering of Syntactic Phrases. 13th ACM Conference on Research and Development in Information Retrieval* (SIGIR'90).
14. Miller, D. R. H., T. Leek and R. M. Schwartz (1999). *A hidden Markov model information retrieval system.* 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99).
15. Mitra, M., C. Buckley, A. Singhal and C. Cardie (1997). *An Analysis of Statistical and Syntactic Phrases.* 5th International Conference Recherche d'Information Assistee par Ordinateur (RIAO'97).
16. Narita, M. and Y. Ogawa (2000). *The use of phrases from query texts in information retrieval. 23rd ACM Conference on Research and Development in Information Retrieval* (SIGIR'00).
17. Robertson, S. E., S. Walker, M. M. Beaulieu, M. Gatford and A. Payne (1995). *Okapi at TREC-4.* 4th Annual Text Retrieval Conference (TREC-4), NIST, Gaithersburg, MD.
18. Salton, G. (1968). *Automatic information organization and retrieval.* New York, McGraw-Hill.
19. Salton, G., C. S. Yang and A. Wong (1975). "A Vector-Space Model for Automatic Indexing." *Communications of the ACM* **18** (11): 613-620.

20. Smeaton, A. F. and F. Kelledy (1998). *User-chosen phrases in interactive query formulation for information retrieval.* 20th BCS-IRSG Colloquium, Springer-Verlag Electronic Workshops in Computing.
21. Song, F. and W. B. Croft (1999). *A general language model for information retrieval.* Eighth International Conference on Information and Knowledge Management (CIKM'99).
22. Strzalkowski, T. (1999). *Natural Language Information Retrieval*, Kluwer Academic Publishers.
23. Turpin, A. and A. Moffat (1999). *Statistical phrases for Vector-Space information retrieval.* 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99).
24. Voorhees, E. (1999). *Natural Language Processing and Information Retrieval.* Information Extraction: Towards Scalable, Adaptable Systems (SCIE'99), New York, Springer.
25. Xu, J. and B. Croft (1998). "Corpus-based Stemming using co-occurrence of word variants." *ACM Transactions on Information Systems* **16** (1): 61-81.
26. Zhai, C., X. Tong, N. Milic-Frayling and D. A. Evans (1997). *Evaluation of syntactic phrase indexing - CLARIT NLP track report.* The Fifth Text Retrieval Conference (TREC-5), NIST Special Publication.
27. Zhai, C. and J. Lafferty (2001). *A Study of Smoothing Methods for Language Models Applied to ad-hoc Information Retrieval.* 24th ACM Conference on Research & Development in Information Retrieval, New Orleans, LA, ACM Press.