Using Term Location Information to Enhance Probabilistic Information Retrieval

Baiyan Liu, Xiangdong An, Jimmy Xiangji Huang
Information Retrieval and Knowledge Management Research Lab
School of Information Technology
York University, Toronto, ON M3J 1P3, Canada
{baiyan, xan, jhuang}@yorku.ca

ABSTRACT

Nouns are more important than other parts of speech in information retrieval and are more often found near the beginning or the end of sentences. In this paper, we investigate the effects of rewarding terms based on their location in sentences on information retrieval. Particularly, we propose a novel Term Location (TEL) retrieval model based on BM25 to enhance probabilistic information retrieval, where a kernel-based method is used to capture term placement patterns. Experiments on five TREC datasets of varied size and content indicate the proposed model significantly outperforms the optimized BM25 and DirichletLM in MAP over all datasets with all kernel functions, and excels the optimized BM25 and DirichletLM over most of the datasets in P@5 and P@20 with different kernel functions.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

Keywords

Term location, probabilistic information retrieval, noun

1. INTRODUCTION

English has 5 basic sentence patterns [1, 10]: (1) Subject + Verb (e.g., Joe swims); (2) Subject + Verb + Object (e.g., Joe plays the guitar); (3) Subject + Verb + Complement (e.g., Joe becomes a doctor); (4) Subject + Verb + Indirect Object + Direct Object (e.g., I give her a gift); and (5) Subject + Verb + Object + Complement (e.g., We elect him president). Most English simple sentences follow these 5 patterns and exceptions are rare (compound and complex sentences can be split into simple sentences) [10], where nouns and noun-phrases are mostly located in the beginning or the end of sentences. On the other hand, past research has indicated that nouns and noun-phrases are more information-bearing than the other parts of speech in information retrieval (IR) [6, 3, 8, 12]. Therefore, integrating term location information into IR may help improve retrieval performances.

Two illustrative experiments are conducted to verify the hypothesis. The first illustrative experiment on WT2g dataset (247,491 web documents, TREC'99 Web track) indicates nouns do concentrate on both ends of sentences as shown in Table 1, where $AvgDis$ is the average of the normalized distances of a set T of nouns from the middle of their sentences as defined by Eq. 1. In Eq. 1, $|T|$ means the cardinality of set T (In this paper, except as explicitly noted, |x| means absolute value of x).

$$AvgDis = \frac{\sum_{t \in T} avg(|Mid(t) - Pos(t)|/Mid(t))}{|T|} \quad (1)$$

where $Mid(t)$ is the middle position of the sentence that noun t is in, $Pos(t)$ is the position of t, and T is the set of nouns. Since a term may appear in a document more than once, average function $avg(.)$ is used. $AvgDis$ has a range of $[0, 1]$, and is closer to 0 if all nouns are gathered in the middle of sentences.

Table 1 shows that $AvgDis > 0.5$ on both halves of the sentences, which means that the nouns are nearer to the beginning or the end of sentences than to the middle of sentences.

Table 1: Noun placement in sentences

End	AvgDis	# Nouns	Avg. sent. len.	# Sentences
Left	0.5901	24,918,926	10.5619	14,360,676
Right	0.613	26,286,542		

Table 2: Relevant noun placement in sentences

Term	Score	Term	Score
louisa	43	head	-24
gradgrind	27	ladi	-17
tom	23	hand	-16
bounderbi	22	countri	-16
slackbridg	15	time	-16

The second illustrative experiment on *Hard Times* by Charles Dickens [2] illustrates that relevant terms are more likely located in the beginning or the end of than in the middle of sentences as shown in Table 2, where $Score > 0$ if a term is more often found near the beginning or the end of sentences, and $Score < 0$ otherwise. To obtain $Score$ of a term t in a document D, sentences in D are each partitioned into three parts, $\{p_1\ p_2\ p_3\}$, where $|p_1| = |p_3|$ and $|p_2| = |p_1| + |p_3|$, and a score to t for its each occurrence in D is assigned by Eq. 2:

$$Score(t) = \begin{cases} 1 & \text{if t } \in p_1 \cup p_3 \\ -1 & \text{if t } \in p_2 \end{cases} \quad (2)$$

Then $Score$ of t for all of its occurrences in D is given by Eq. 3:

$$Score(t, D) = \sum_{t_i \in D} Score(t_i) \quad (3)$$

where t_i is the ith occurrence of t in D.

Table 2 shows that the highest scoring terms "louisa", "gradgrind", "bounderbi", "tom", and "slackbridg" turn out to be the

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767827.

main or minor characters in the book, whereas the lowest scoring terms are not particularly related with the novel.

The results from the two illustrative experiments above indicate the hypothesis deserves a deeper investigation. The main contributions of this paper are as follows:

- We extend BM25 naturally with term location information to enhance probabilistic information retrieveal;
- In order to reward terms that are more likely to be nouns, we propose a novel kernel-based Term Location (TEL) retrieval model to capture term placement patterns;
- Experiments on five TREC datasets of varied size and content indicate the proposed model is highly promising.

2. RELATED WORK

Jing and Croft [6] proposed PhraseFinder to automatically construct collection-dependent association thesauri, and then used the association thesauri to assist query expansion and found that nouns and noun-phrases were most effective in improving IR. Liu et al. [8] classified noun-phrases into four types – proper names, dictionary phrases, simple phrases, and complex phrases – and ranked documents based on phrase similarity. Zheng et al [17] used noun-phrases and semantic relationships to represent documents in order to assist document clustering, where noun-phrases were extracted with the assistance of WordNet. Yang et al [14] used a parse tree to transform the sentences in legal agreements into subject-verb-object (SVO) representations, which are then used in conjunction with cue terms to identify the provisions provided by the sentences. However, they found that provision extraction using the SVO representation resulted in high precision but low recall, which could be due to the specificity of SVO sentence patterns and the difficulty in parsing complex sentences. Hung et al. [4] used syntactic pattern matching to extract syntactically complete sentences that express event-based commonsense knowledge from web documents, and then semantic role labeling was used to tag the semantic roles of the terms in the sentences, such as the subject and the object. Ibekwe-SanJuan et al. [5] built finite state automaton with syntactic patterns and synonyms from WordNet, which was used to tag the sentences in scientific documents according to its category. It is difficult to find syntactic patterns that are effective in all the documents of a single corpus, and rule-based part-of-speech taggers are less effective in unseen text [7]. Terms were rewarded based on their locations in a document in [15, 16].

3. OUR APPROACH

3.1 Term Location

We assume that the most important terms in the documents are near the beginning or the end of the sentences. We determine the importance (relevancy) of a term t by examining its distance from the middle of its sentence in document D as defined by Eq. 4:

$$q(t, D) = |Mid(t, D) - Pos(t, D)| \tag{4}$$

$$\text{where } Mid(t, D) = (SL(t, D) - 1)/2,$$

$SL(t, D)$ is the length of the sentence in D that contains t, and $Pos(t, D)$ is the location of t in the sentence. We use the average distance of t from the middle of its sentences in D if t appears more than once in D, which is defined as $r(t, D)$ by Eq. 5:

$$r(t, D) = (\sum_{t_i \in D} q(t_i, D))/tf(t, D) \tag{5}$$

where t_i is the ith occurrence of t in D and $tf(t, D)$ is the term frequency of t in D. We define $m(t, D)$ to be the average length of the sentence(s) that contain t in D as Eq. 6:

$$m(t, D) = \frac{\sum_{t_i \in D} SL(t_i, D)}{\beta * tf(t, D)} + \gamma \tag{6}$$

where parameter β has a larger effect for longer sentences since it is proportional to the lengths of the sentences, and parameter γ has a proportionally smaller effect for longer sentences since it is the same for all sentences. These two parameters are used to balance term weights in particularly short or long sentences.

3.2 Kernel Functions

In order to measure the distances of the terms from the middle of their sentences, we fit a kernel function over each sentence. We adjust the weight of each term based on its average distance to the middle of its sentences. In this paper, we explore the following kernel functions for our location based reward function $RN(t, D)$ used in Eq. 9:

$$\text{Gaussian - } Kernel(r, m) = 1 - e^{\frac{r^2}{-2m^2}}$$

$$\text{Triangle - } Kernel(r, m) = \frac{r}{m}$$

$$\text{Cosine - } Kernel(r, m) = 1 - \frac{1 + \cos\frac{r\pi}{m}}{2}$$

$$\text{Circle - } Kernel(r, m) = 1 - \sqrt{1 - \left(\frac{r}{m}\right)^2}$$

$$\text{Quartic - } Kernel(r, m) = 1 - \left(1 - \left(\frac{r}{m}\right)^2\right)^2$$

$$\text{Epanechnikov - } Kernel(r, m) = \left(\frac{r}{m}\right)^2$$

$$\text{Triweight - } Kernel(r, m) = 1 - \left(1 - \left(\frac{r}{m}\right)^2\right)^3$$

Among them, Gaussian kernel is widely used in statistics and machine learning such as Support Vector Machines, Triangle kernel, Circle Kernel, and Cosine Kernel are applied to estimate the proximity-based density distribution for the positional language model [9]. Since the kernel functions are not maturely applied in IR, we also explore Quartic kernel, Epanechnikov kernel and Triweight kernel in this work. In these kernel functions, r and m are defined by Eqs. 5 and 6, respectively. With these kernel functions, the number of terms that are given maximum reward decreases as $m(t, D)$ increases.

3.3 Integration into BM25

In this experiment, we use the BM25 weighting model as our base weighting model. BM25 is defined as follows:

$$Score(t, D) = TF(t, D) * IDF(t) \tag{7}$$

$$\text{where } TF(t, D) = \frac{(k_3 + 1) * tf(t, D) * qtf(t)}{(k_3 + qtf(t)) * K},$$

$$IDF(t) = log_2 \frac{N - n(t) + 0.5}{n(t) + 0.5},$$

$$K = k_1 * \left(1 - b + \frac{b * |D|}{AvgDL}\right) + tf(t, D),$$

k_1, k_3, and b are tuning parameters for BM25, $qtf(t)$ is the frequency of t in the query, $|D|$ is the number of terms in D, $AvgDL$ is the average length of the documents in the collection, N is the number of documents in the collection, and $n(t)$ is the number of documents in the collection that contain t. We modify $TF(t, D)$ to account for the reward given to the terms based on their locations in the sentences. We define the Term Location score (TL) as follows:

$$TL(t, D) = \frac{(k_3 + 1) * RN(t, D) * tf(t, D) * qtf(t)}{(k_3 + qtf(t)) * K_{TL}} \tag{8}$$

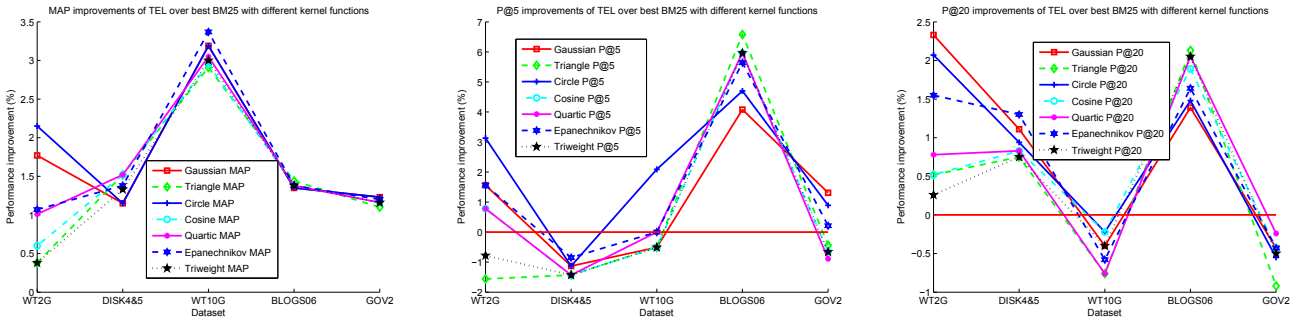

Figure 1: Performance improvements of TEL over best BM25 with different kernel functions.

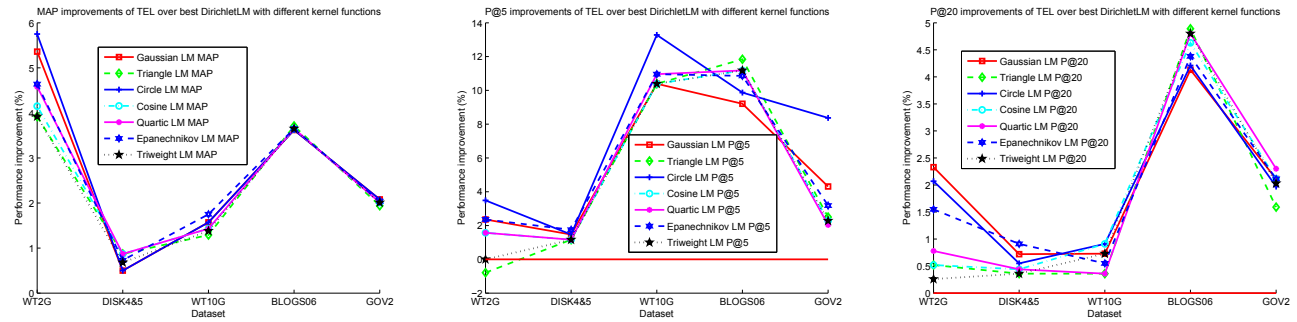

Figure 2: Performance improvements of TEL over best DirichletLM with different kernel functions.

where

$$K_{TL} = k_1 * \left(1 - b + \frac{b * |D|}{AvgDL}\right) + RN(t,D) * tf(t,D) \quad (9)$$

We integrate our model into BM25 to form the Term Location Score (TEL) as follows:

$$TEL(t,D) = ((1-\alpha) * TF(t,D) + \alpha * TL(t,D)) * IDF(t) \quad (10)$$

where α controls the contribution of our model.

4. EXPERIMENTAL RESULTS

We conduct experiments on f ve standard TREC collections: WT2G, DISK4&5, WT10G, BLOGS06, and GOV2. These datasets vary in both size and content, where WT2g contains 247,491 general Web documents (TREC'99 Web track), DISK4&5 is comprised of 528,155 newswire documents from sources such as the Financial Times and the Federal Register (TREC'97-99 Ad hoc track), WT10G has 1,692,096 general web documents (TREC'00-01 Web track), BLOGS06 consists of 3,215,171 feeds from late 2005 to early 2006 with associated permalink and homepage documents (TREC'06-08 Blog track), and GOV2 holds 25,178,548 documents crawled from .gov sites (TREC'04-06 Terabyte track).

We compare our model against the following weighting models when they perform best on each dataset with parameters obtained as follows:

1. BM25, with $k_1 = 1.2$ and $k_3 = 8$. We adjust b in the range of [0.1, 0.9] in steps of 0.1 for each dataset to f nd the value of b that gives the best MAP for that dataset.
2. DirichletLM. We adjust μ in the range of [100, 3000] in steps of 100. We f nd the optimal value of μ for each dataset.

The proposed model TEL uses the same values as BM25 for k_1, k_3, and b, and sets $\alpha = 0.2$, $\beta = 3$, and $\gamma = 3$ for all datasets.

In the future, we would study the optimal values of the model parameters and their relations with the characteristics of the datasets. We use the TREC off cial evaluation measures in our experiments, namely the topical MAP on BLOGS06 [11], and Mean Average Precision (MAP) on all the other datasets [13]. To stress the top retrieved documents, we also include P@5 and P@20 as the evaluation measures. All statistical tests are based on Wilcoxon Matched-pairs Signed-rank test.

The experimental results are presented in Table 3. To illustrate the performance differences graphically, we plot the results in Figures 1 and 2. As shown by the two f gures, our model TEL outperforms optimized BM25 and DirichletLM in MAP over all datasets with all kernel functions, and outperforms the two optimized baseline models over most of the datasets in $P@5$ and $P@20$ with different kernel functions. The performance improvements of our model TEL against DirichletLM are greater than those against BM25. According to the two f gures, each kernel function has its advantage on some datasets. There is no single kernel function that outperforms others on all the datasets.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we extend BM25 and reward the terms based on their locations in the sentences with kernel functions. Experimental study shows that the proposed model performs signif cantly better than BM25 and DirichletLM on MAP over all datasets, and significantly better on P@5, p@10, and p@20 over most datasets.

In the future, more experiments will be conducted to further investigate the proposed model. We would investigate non-symmetric kernel functions and kernel functions with negative values since the placement of the terms at the beginning of the sentences is different from that at the end of the sentences as indicated in the f rst illustra-

Model	Eval Metric	WT2G	DISK4&5	WT10G	BLOGS06	GOV2
$BM25$	MAP	*0.3167*	*0.2176*	*0.2134*	*0.3195*	*0.3008*
	$P@5$	*0.5120*	**0.4680**	*0.3918*	*0.6380*	*0.6094*
	$P@20$	*0.3870*	*0.3613*	**0.2776**	*0.6095*	**0.5406**
$DirichletLM$	MAP	*0.3059*	*0.2190*	*0.2168*	*0.3125*	*0.2983*
	$P@5$	*0.5080*	*0.4560*	*0.3531*	*0.6080*	*0.5919*
	$P@20$	*0.3870*	*0.3627*	*0.2745*	*0.5935*	*0.5272*
TEL Gaussian	MAP	0.3223*+ (+1.77%,+5.36%)	0.2201 (+1.15%,+0.50%)	0.2202* (+3.19%,+1.57%)	0.3238+ (+1.35%,+3.62%)	**0.3045***+ (+1.23%,+2.08%)
	P@5	0.5200* (+1.56%,+2.36%)	0.4627+ (-1.13%,+1.47%)	0.3898+ (-0.51%,10.39%)	0.6640* (+4.08%,+9.21%)	**0.6174***+ (+1.31%,+4.31%)
	P@20	**0.3960** (+2.33%,+2.33%)	0.3653* (+1.11%,+0.72%)	0.2765 (-0.40%,+0.73%)	0.6180 (+1.39%,+4.13%)	0.5383 (-0.43%,+2.11%)
TEL Triangle	MAP	0.3179 (+0.38%,+3.92%)	**0.2209*** (+1.52%,+0.87%)	0.2196* (+2.91%,+1.29%)	**0.3241***+ (+1.44%,+3.71%)	0.3041*+ (+1.10%,+1.94%)
	P@5	0.5040 (-1.56%,-0.79%)	0.4613+ (-1.43%,+1.16%)	0.3898+ (-0.51%,+10.39%)	**0.6800*** (+6.58%,+11.84%)	0.6067+ (-0.44%,+2.50%)
	P@20	0.3890* (+0.52%,+0.52%)	0.3640*+ (+0.75%,+0.36%)	0.2755+ (-0.76%,+0.36%)	**0.6225** (+2.13%,+4.89%)	0.5356 (-0.92%,+1.59%)
TEL Circle	MAP	**0.3235***+ (+2.15%,+5.75%)	0.2201 (+1.15%,+0.50%)	0.2202* (+3.19%,+1.57%)	0.3238+ (+1.35%,+3.62%)	**0.3045***+ (+1.23%,+2.08%)
	P@5	**0.5280*** (+3.13%,+3.49%)	0.4627+ (-1.13%,+1.47%)	**0.4000*** (+2.09%,+13.28%)	0.6680* (+4.70%,+9.87%)	0.6148*+ (+0.89%,+8.37%)
	P@20	0.3950 (+2.07%,+2.07%)	0.3647* (+0.94%,+0.55%)	0.2770 (-0.22%,+0.91%)	0.6185 (+1.48%,+4.21%)	0.5376 (-0.55%,+1.97%)
TEL Cosine	MAP	0.3186 (+0.60%,+4.15%)	**0.2209*** (+1.52%,+0.87%)	0.2197* (+2.95%,+1.43%)	0.3239*+ (+1.38%,+3.65%)	0.3043*+ (+1.16%,+2.01%)
	$P@5$	0.5160* (+0.78%,+1.57%)	0.4613+ (-1.43%,+1.16%)	0.3898+ (-0.51%,+10.39%)	0.6760* (+5.96%,+11.18%)	0.6054+ (-0.66%,+2.28%)
	$P@20$	0.3890* (+0.52%,+0.52%)	0.3643*+ (+0.83%,+0.44%)	0.2770+ (-0.22%,+0.91%)	0.6210 (+1.89%,+4.63%)	0.5383 (-0.43%,+2.11%)
TEL Quartic	MAP	0.3199*+ (+1.01%,+4.58%)	**0.2209*** (+1.52%,+0.87%)	0.2199* (+3.05%,+1.43%)	0.3239*+ (+1.38%,+3.65%)	0.3043*+ (+1.16%,+2.01%)
	P@5	0.5160* (+0.78%,+1.57%)	0.4613+ (-1.43%,+1.16%)	0.3918+ (+0.00%,+10.96%)	0.6760* (+5.96%,+11.18%)	0.6040+ (-0.89,+2.04%)
	P@20	0.3900* (+0.78%,+0.78%)	0.3643*+ (+0.83%,+0.44%)	0.2755 (-0.76%,+0.36%)	0.6220 (+2.05%,+4.80%)	0.5393 (-0.24%,+2.30%)
TEL Epanechnikov	MAP	0.3201*+ (+1.07%,4.64%)	0.2206 (+1.38%,+0.73%)	**0.2206*** (+3.37%,+1.75%)	0.3239+ (+1.38%,+3.65%)	0.3044*+ (+1.20%,+2.04%)
	P@5	0.5200* (+1.56%,+2.36%)	0.4640+ (-0.85%,1.75%)	0.3918+ (+0.00%,+10.96%)	0.6740* (+5.64%,+10.86%)	0.6107*+ (+0.21%,+3.18%)
	P@20	0.3930 (+1.55%,+1.55%)	**0.3660*** (+1.30%,0.91%)	0.2760 (-0.58%,+0.55%)	0.6195 (+1.64%,+4.38%)	0.5383 (-0.43%,+2.11%)
TEL Triweight	MAP	0.3179 (+0.38%,+3.92%)	0.2205* (+1.33%,+0.68%)	0.2198* (+3.00%,+1.38%)	0.3239*+ (+1.38%,+3.65%)	0.3043*+ (+1.16%,+2.01%)
	P@5	0.5080 (-0.78%,0.00%)	0.4613+ (-1.43%,+1.16%)	0.3898+ (-0.51%,+10.39%)	0.6760* (+5.96%,+11.18%)	0.6054+ (-0.66%,+2.28%)
	P@20	0.3880* (+0.26%,+0.26%)	0.3640*+ (+0.75%,+0.36%)	0.2765+ (-0.40%,+0.73%)	0.6220 (+2.05%,+4.80%)	0.5379 (-0.50%,+2.03%)

Table 3: Comparison between TEL and two baselines BM25 and DirichletLM with different kernel functions: Parameter b of BM25 and parameter μ of DirichletLM are obtained and set individually for each dataset for their best performances, and "*" and "+" denote statistically signif cant improvements over BM25 and DirichletLM (Wilcoxon signed-rank test with $p < 0.05$), respectively. The best result obtained on each dataset is in bold. The two percentages below each value are the percentage improvement of TEL over BM25 and DirichletLM, respectively.

tive experiment. It is also worthwhile to analyze the optimal values of the model parameters and their relations with the characteristics of the datasets. Different term proximity measures would also be explored to improve the performance of our model.

6. ACKNOWLEDGMENTS

This research is supported by the research grant from the Natural Sciences & Engineering Research Council (NSERC) of Canada and NSERC CREATE Award. We thank anonymous reviewers for their thorough review comments on this paper.

7. REFERENCES

[1] H. Ann. *The Essentials of English — a writer's handbook*. New York, Pearson Education, 2003.

[2] C. Dickens. *Hard Times*. Bradbury & Evans, 1854.

[3] D. A. Evans and C. Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *ACL 1996*, pages 17–24.

[4] S.-H. Hung, C.-H. Lin, and J.-S. Hong. Web mining for event-based commonsense knowledge using lexico-syntactic pattern matching and semantic role labeling. *Expert Systems with Applications*, 37(1):341–347, 2010.

[5] F. Ibekwe-SanJuan and et al. Annotation of scientif c summaries for information retrieval. *CoRR*, 2011.

[6] Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In *RIAO'94*, pages 146–160.

[7] K. Liu and et al. Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. *Meth. of info. in med.*, 50(5):397, 2011.

[8] S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *SIGIR'04*, pages 266–272.

[9] Y. Lv and C. Zhai. Positional language models for information retrieval. In *SIGIR'09*, pages 299–306.

[10] C. F. Meyer. *Introducing English Linguistics*. Cambridge University Press, 2010.

[11] I. Ounis and et al. Overview of the TREC-2006 blog track.

[12] O. Vechtomova. Noun phrases in interactive query expansion and document ranking. *Info. Retrieval*, 9:399–420, 2006.

[13] E. Voorhees and D. Harman. *TREC: Experiment and evaluation in information retrieval*. MIT Press, 2005.

[14] D. Yang and et al. A natural language processing and semantic-based system for contract analysis. In *ICTAI 2013*, pages 707–712.

[15] J. Zhao, X. Huang, and S. Wu. Rewarding term location information to enhance probabilistic information retrieval. In *SIGIR'12*.

[16] J. Zhao, X. Huang, and Z. Ye. Modeling term associations for probabilistic information retrieval. *ACM Trans. Inf. Syst.*, 32(2), 2014.

[17] H.-T. Zheng and et al. Exploiting noun phrases and semantic relationships for text document clustering. *Info. Sci.*, 179(13):2249–2262, 2009.