Context-Sensitive Semantic Smoothing for the Language Modeling Approach to Genomic IR

Xiaohua Zhou, Xiaohua Hu, Xiaodan Zhang, Xia Lin, Il-Yeol Song
College of Information Science & Technology, Drexel University
xiaohua.zhou@drexel.edu, {thu, xzhang, xlin}@ischool.drexel.edu, song@drexel.edu

Abstract
Semantic smoothing, which incorporates synonym and sense information into the language models, is effective and potentially significant to improve retrieval performance. The implemented semantic smoothing models, such as the translation model which statistically maps document terms to query terms, and a number of works that have followed have shown good experimental results. However, these models are unable to incorporate contextual information. Thus, the resulting translation might be mixed and fairly general. To overcome this limitation, we propose a novel context-sensitive semantic smoothing method that decomposes a document or a query into a set of weighted context-sensitive topic signatures and then translate those topic signatures into query terms. In detail, we solve this problem through (1) choosing concept pairs as topic signatures and adopting an ontology-based approach to extract concept pairs; (2) estimating the translation model for each topic signature using the EM algorithm; and (3) expanding document and query models based on topic signature translations. The new smoothing method is evaluated on TREC 2004/05 Genomics Track collections and significant improvements are obtained. The MAP (mean average precision) achieves a 33.6% maximal gain over the simple language model, as well as a 7.8% gain over the language model with context-insensitive semantic smoothing.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval Models—language models

General Terms
Algorithm, Experimentation, Performance

Keywords
Language Models, Information Retrieval, Genomic Information Retrieval, Semantic Smoothing, Topic Signature, Concept Pair

1. Introduction
The language modeling approach to information retrieval (IR), initially proposed by Ponte and Croft [14], has been popular with the IR community in recent years due to its solid theoretical foundation and promising empirical retrieval performance. In essence, this approach centers on the document model estimation and the query generative likelihood calculation for ranking according to the estimated model. However, it is challenging to estimate an accurate document model. On one hand, because the query terms may not appear in the document, we need to assign a reasonable non-zero probability to the unseen terms. On the other hand, we need to adjust the probability of the seen terms to remove the effect of the background model or even irrelevant noise. Thus, the core of the language modeling approach to IR is to "smooth" the models. Zhai and Lafferty [16, 18] propose several effective smoothing techniques that interpolate the document model with the background collection model.

A potentially more significant and effective method is semantic smoothing that incorporates synonym and sense information into the language model [10]. Berger and Lafferty [2] incorporate a kind of semantic smoothing into the language model by statistically mapping document terms onto query terms using a translation model trained from synthetic document-query pairs. The translation model is context-insensitive (i.e., it cannot incorporate sense and other contextual information into the language model), however, and therefore the resulting translation may be mixed and fairly general. For example, the term "*mouse*" without context may be translated to both "*computer*" and "*cat*" with high probabilities. Jin [9] and Cao [3] present two other ways to train the translation models, but they still have the same context-insensitivity problem as [2].

Lafferty and Zhai [10] introduce another more generic and flexible language model called KL-divergence retrieval model as a special case of their risk minimization retrieval framework. KL-divergence retrieval estimates the query model as well as the document model. Like the document model estimation, a typical method for query model estimation is to statistically translate the terms in the original query into other terms [1, 10]. In this paper, we also refer to the translation-based query model estimation as *semantic smoothing*. Similarly, if the translation model is context-insensitive, the resulting query model may be very general. Thus, it is urgent to develop a framework to semantically smooth query and document models in the language modeling (LM) retrieval framework.

In this paper, we propose a novel context-sensitive semantic smoothing method based on topic decomposition. A query or a document is decomposed into a set of weighted topic signatures and those topic signatures are translated into individual concepts for the purpose of query or document expansions. We define a topic signature as a pair of two topic concepts that are related to each other syntactically and semantically. Because two related concepts help to determine context for each other, the signature-

SIGIR'06, August, 6-11, 2006, Seattle, Washington, USA.

based translation should have higher accuracy and result in better retrieval performance. For example, "*mouse"* in conjunction with *"computer"* could be a topic signature and the signature might be translated to "*keyboard*" with a high probability, but to "*cat*" with a low probability due to additional contextual constraints.

We incorporate the context-sensitive semantic smoothing into language models through (1) adopting an ontology-based approach to extract concepts and signatures from queries and documents; (2) developing an EM-based method to train the signature translation model; and (3) expanding document and query language models based on topic signature translations. The new smoothing method is tested on TREC04/05 Genomic collections. The experimental results show that significant improvements are obtained over the simple language model as well as the model with context-insensitive semantic smoothing. The contribution of this paper is three-fold. First, it proposes a new document representation using a set of weighted concepts and topic signatures. Second, it expands document and query language models through context-sensitive semantic smoothing. Third, it empirically proves the effectiveness of context-sensitive semantic smoothing for language modeling IR.

The remainder of this paper is organized as follows. In Section 2, we review previous work on context-sensitive semantic smoothing, finalize the representation for topic signatures, and implement the topic signature extraction. In Section 3, we present the expanded language models with context-sensitive semantic smoothing. In Section 4, we test the new model on TREC04/05 Genomics Track collections. Section 5 concludes our paper.

2. Context-Sensitive Topic Signatures

2.1 Previous Work

Liu and Croft [12] propose a cluster language model and achieve great empirical improvement over the baseline model. Unlike our method, which decomposes a document or a query into a set of small topic signatures, their method aggregates similar documents into clusters and then treats each cluster as a big document for ranking purposes. All documents in the relevant clusters are returned to the users. The two models are similar in the sense that both want to obtain a set of documents with similar context rather than a single document in order to estimate a more accurate and smoothed model. The major difference is that a document only belongs to a cluster in the cluster model whereas a document can have multiple topic signatures in our model. Furthermore, many decisions need to be made empirically for clustering, based on the domain knowledge and the collection (e.g. the number of clusters, clustering algorithm, static clustering or query-specific clustering), while the topic signature model does not have this problem.

Song and Bruza introduce an information flow (IF) based query expansion technique in [15]. A HAL vector is used as the context of a concept. The degree of one concept inferring another can then be heuristically computed. They also invent a heuristic approach to combine multiple concepts, which enables information inference from a group of concepts (premises) to one individual concept (conclusion). Therefore, their query expansion technique is somehow context-sensitive. Bai et al. [1] slightly adapt the above approach to the KL-divergence retrieval framework. Both of them achieve significant improvement over the simple language model. IF can be computed simply from term co-occurrence data without any external knowledge and is thus of value in practice. The major drawback of this approach is that it is unable to trace the information flow back to the documents or queries. Therefore, it is difficult to estimate an IF document model or query model (i.e., computing the generative probability of the premise of an IF from a query or a document). For a short query, the uniform distribution assumption, as made in [1], may not be a problem. But for a document, it is obviously not reasonable. In other words, it can not be used to expand document models. Besides, the degree to which one individual concept could be inferred from another combined concept is not theoretically motivated; its robustness needs to be further validated.

2.2 Topic Signature Representation

The choice of topic signature representation plays a crucial role in our context-sensitive semantic smoothing method. First, the topic signature must be context-sensitive and thus the signature should contain at least two terms, unless word sense is adopted. Second, terms within a signature should have syntactic relation. Otherwise, we cannot count their frequency in documents or queries and it becomes difficult to estimate signature document models and query models. Third, it should be easy to extract topic signatures from the text. Last, we hope all terms within a signature have semantic relations inspired by the idea of [3], where WordNet semantic relationships are considered.

Harabagiu and Lacatusu proposed in [6] that topics could be represented by a set of weighted binary relations between topic concepts; a relation could be either syntax-based or entity-event paired without syntactic constraint. However, for the convenience of model estimations, only syntax-based relation is allowed in this paper. In addition, we impose semantic constraints on two concepts in order to reduce the noise. Thus, we end up with the definitions of topic signature below.

Definition 1 A ***topic signature (t)*** is defined with two order-free components as in $t(w_i, w_j)$, where w_i and w_j are two concepts related to each other syntactically and semantically. For simplicity, $t(w_i, w_j)$ is also denoted as t_{ij}. The implementation of the syntactic and semantic relationships between two concepts is determined by specific applications.

Definition 2 A ***concept (w)*** is a unique meaning in a domain. It represents a set of synonymous terms in the domain. For example, *C0020538* is a concept about the disease of hypertension in UMLS Metathesaurus (http://www.nlm.nih.gov/research/umls); it also represents a set of synonymous terms including *high blood pressure*, *hypertension*, and *hypertensive disease*. Therefore, concept-based indexing and searching helps to relieve the synonym and polysemy problems in IR, especially genomic IR, where a term (e.g., a gene or a protein) might have many synonyms while also representing different concepts in different context [20].

The benefit of using concept pairs as topic signatures is four-fold. First, two-topic concepts help to determine the context for IR use while not producing too many concept combinations. Second, a number of existing approaches are available to extract binary relationships, which are similar to concept pairs. For example, in the area of NLP, especially in bioinformatics, pattern-based methods for binary relation extractions have been extensively studied in recent years [13]. Third, a concept pair itself is very similar to a short query. In TREC 2005 Genomics Track [8], structured concept pairs are directly used as ad hoc retrieval topics. In [19], concept pairs are treated as an index unit as well as a search unit. Last, documents are full of various concept pairs and it is possible to make a robust estimation of document models by linearly combining a set of topic signature models.

2.3 Topic Signature Extraction

In general, the extraction of topic signatures is done in two steps: the topic concept extraction and the concept pair extraction. The extraction of biological concepts and their binary relationships is a hot topic in bioinformatics and a survey of those methods can be found in [13]. However, on one hand, our extraction is for IR use and we are dealing with large corpora statistically; thus the extraction methods need not be perfect [5]. On the other hand, we are more interested in semantic correspondences between topic concepts than syntactic patterns. Thus, we use a generic ontology-based approach to extract topic signatures.

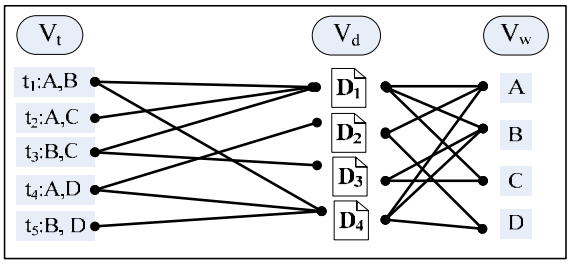


Figure 1. Illustration of document indexing. V_t, V_d and V_w are topic signature set, document set and concept set, respectively.

We adopt MaxMatcher [20] for concept extractions. The concept extraction by MaxMatcher is equivalent to maximizing the weighted overlap between the word sequences in text and the concepts in an ontology, such as the UMLS Metathesaurus. It outputs concept names as well as unique IDs representing a set of synonymous concepts. The unique concept IDs are used as an index in our experiments. MaxMatcher distinguishes between major concepts and sub concepts in the manner defined below.

Definition 3 A concept syntactically embedded in another concept is called a ***sub concept***, otherwise it is called a ***major concept***. However, the membership of a concept is context-dependent; a sub concept in one text could be a major concept in another. For example, "*blood pressure*" is a sub concept for the text "*high blood pressure*," but is a major concept for text "*the blood pressure is…*" We index both sub concepts and major concepts in the experiment.

We developed a coarse approach to extract topic signatures in order to show the robustness and effectiveness of the signature-based semantic smoothing. A pair of two topic concepts will be treated as a topic signature if they meet the following three requirements: (1) both of them are major concepts; (2) they appear in the same clause of an English sentence; and (3) their semantic types are compatible according to the domain ontology. For example, two proteins could be semantically compatible in UMLS (e.g., protein-protein interaction).

Example:
A recent epidemiological study (C0002783, research activity) revealed that obesity (C0028754, disease) is an independent risk factor for periodontal disease (C0031090, disease).
Concept Index: *C0002783, C0028754, C0031090*
Topic Signature Index: *(C0028754, C0031090)*

In the above example, the underlined phrases are extracted concept names followed by the corresponding concept ID and semantic type. The concept pair of obesity and periodontal disease is a topic signature while the concept epidemiological study has no relationships with other concepts because it is in a separate clause.

3. Context-Sensitive Semantic Smoothing

3.1 Signature Translation Model Estimates

Suppose we have indexed all documents with concepts and topic signatures (see Figure 1). For each topic signature t_k, we have a set of documents (D_k) containing that topic signature. Intuitively, we can use the document set D_k to approximate the translation model for t_k, i.e., determining the probability of translating the signature to concepts in the vocabulary. If all concepts appearing in the document set center on the topic signature t_k, we can simply use maximum likelihood estimates and the problem is as simple as frequency counting. However, some concepts address the issue of other topic signatures while some are background concepts of the collection. We use the generative model proposed in [17] to remove the noise. Assume the set of documents containing t_k is generated by a mixture model (i.e., interpolating the translation model with the background collection model $p(w|C)$),

$$p(w|D_k) = (1-\alpha)p(w|\theta_{t_k}) + \alpha p(w|C) \quad (3.1)$$

where α is a coefficient accounting for the background noise and θ_{t_k} refers to the translation model of the topic signature t_k.. Then we estimate the translation model using the EM algorithm [4]:

$$\hat{p}^{(n)}(w) = \frac{(1-\alpha)p^{(n)}(w|\theta_{t_k})}{(1-\alpha)p^{(n)}(w|\theta_{t_k}) + \alpha p(w|C)} \quad (3.2)$$

$$p^{(n+1)}(w|\theta_{t_k}) = \frac{c(w,D_k)\hat{p}^{(n)}(w)}{\sum_i c(w_i,D_k)\hat{p}^{(n)}(w_i)} \quad (3.3)$$

where $c(w,D_k)$ is the frequency count of concept w in D_k.

Our topic signature translation model is significantly different from previous translation models [2, 3, 9, 10] in two aspects. First, previous translation models take an individual term as the topic signature, and are unable to incorporate contextual information into the model. Our model uses a group of terms with syntactic and semantic relation to each other as the topic signatures. Consequently, the resulting translation will be more specific.

Second, the method for model estimation is different. Berger and Lafferty [2] use document-query pairs to train translation probabilities. However, it is unlikely to obtain a large amount of real data. For this reason, they use synthetic data for model estimation. The title language model, proposed in [9], uses title-document pairs to train translation probabilities. The major drawback of the title model is that only a small portion of terms in the vocabulary would appear in the title. The Markov chain model [10] deals with translations in a different fashion. However, the resulting query model is fairly general and the computation of the inverse matrix will be prohibitive to large collections. Cao [3] takes into account word semantics when computing term associations, but he ignores the sense of the words; this model is roughly equivalent to our context-insensitive version of semantic smoothing introduced in Section 4.6.

3.2 Document Model Smoothing

Lafferty and Zhai introduced the KL-Divergence retrieval model as a special case of their risk minimization retrieval framework [10]. This retrieval model estimates query models as well as document models; the relevance of a document to a query is

equivalent to measuring the KL-divergence distance between the query model and the document model:

$$R(d;q) \propto \sum_{w} p(w \mid q) \log p(w \mid d) \qquad (3.4)$$

The introduction of query models makes the language modeling approach more flexible. Almost all previous language models for IR are the special cases of this new retrieval model. The context-sensitive semantic smoothing technique proposed in this paper works with the KL-divergence retrieval model.

A simple unigram document model can be easily obtained using the maximum likelihood estimate. To avoid assigning zero probability to unseen terms and to reduce the noise, it could be simply interpolated with a background collection model $p(w \mid C)$:

$$p_b(w \mid d) = (1-\alpha) p_{ml}(w \mid d) + \alpha p(w \mid C) \qquad (3.5)$$

where α is a coefficient accounting for the background model. We use this simple mixture language model as the baseline in the comparative study and refer to it as **DM0**.

With the availability of context-sensitive topic signatures, a document model can be expanded by statistically mapping the topic signatures in the document to query terms. That is,

$$p_t(w \mid d) = \sum_{k} p(w \mid \theta_{t_k}) p_{ml}(t_k \mid d) \qquad (3.6)$$

The topic signature document model (i.e., the generative probability of topic signatures in a document) can be computed using a maximum likelihood estimate:

$$p_{ml}(t_k \mid d) = \frac{c(t_k, d)}{\sum_i c(t_i, d)} \qquad (3.7)$$

Where $c(t_i, d)$ is the frequency of signature t_i in document d. We refer to this translation-based document model as **DM1.** The form of DM1 is same as the translation model described in [2]. However, the topic signature in DM1 is more generic. It could be individual terms, as used in [2], or context-sensitive concept pairs as used in this paper, or any other objects that can express a topic.

A potential problem of Model DM1 is that the extracted topic signatures may not be very representative when the document is too short or the criterion of being a topic signature is too strict. Thus, the accuracy of the document model will be compromised. To overcome this limitation, we interpolate DM1 with DM0.

$$p_{bt}(w \mid d) = (1-\lambda) p_b(w \mid d) + \lambda p_t(w \mid d) \qquad (3.8)$$

This mixture model is referred to as **DM2**. The translation coefficient (λ) controls the influence of the translation component in the mixture model. The mixture model becomes DM0 when λ is zero and becomes DM1 when λ is one. In the experiment, we tune the translation coefficient to optimize the retrieval performance.

3.3 Query Model Smoothing

Like the expansion mechanism for document models, a query model can be expanded through the signature-concept translation if the query could be decomposed into a set of representative topic signatures. That is,

$$p(w \mid q) = \sum_{k} p(w \mid \theta_{t_k}) p(t_k \mid q) \qquad (3.9)$$

However, query descriptions are often very short; therefore, it is difficult to extract representative topic signatures, let alone compute the generative probability (i.e. the importance to the query). For this reason, we do not smooth query models during the initial search. Instead, we update the query model according to the top-ranked documents of the initial search, which is referred to as blind feedback or pseudo-relevance feedback. We expect that the feedback documents will give us a more precise sense of what the query is about.

Zhai and Lafferty formalized the blind feedback as a process of re-estimating the query model according to the feedback documents within the KL-divergence retrieval framework [17]. By interpolating the feedback model with the initial query model, we obtain the final query model for the feedback search. The tunable feedback coefficient γ controls the influence of the feedback model in the mixture query model.

$$p(w \mid q) = (1-\gamma) p_i(w \mid q) + \gamma p_f(w \mid q) \qquad (3.10)$$

Though top-ranked documents have a high probability of being relevant to the initial query, it does not mean all topic signatures in those documents are relevant to the query. Intuitively, topic signatures containing one or more query terms are probably more relevant to the initial query than those containing no query terms. Counting topic signatures containing at least one query term, we derive a feedback model below:

$$p_f(w \mid q) = \sum_{k: q \cap t_k \neq \varnothing} p(w \mid \theta_{t_k}) \frac{c(t_k, F)}{\sum_{i: q \cap t_i = \varnothing} c(t_i, F)} \qquad (3.11)$$

where $c(t_k, F)$ is the frequency of the topic signature t_k in the feedback document set F. The resulting feedback model, however, might be fairly general because the topic signature translation probability is simply trained from a set of documents containing that signature. To overcome this problem, only self-translation (i.e., translating a topic signature to its own concepts) is allowed. Then we obtain a heuristic feedback model called **FM0**:

$$p_f(w \mid q) = \sum_{k: q \cap t_k \neq \varnothing} p_s(w \mid t_k) \frac{c(t_k, F)}{\sum_{i: q \cap t_i = \varnothing} c(t_i, F)} \qquad (3.12)$$

where:

$$p_s(w \mid t_k) = \begin{cases} 0 & w \notin t_k \\ 1/|t_k| & w \in t_k \end{cases}$$

Model FM0 will filter out some irrelevant topic signatures using term associations, but it will still keep the effect of the background topic signatures. In other words, high-frequency topic signatures in F might be also frequent in the collection. For this reason, we use the approach introduced in [17] to remove the effect of the background collection model. This approach assumes that the topic signatures in feedback documents are generated by a mixture model (interpolating the signature feedback model with the signature collection model):

$$p(t_k \mid \theta_F, F, C) = (1-\alpha) p(t_k \mid \theta_F) + \alpha p(t_k \mid C) \qquad (3.13)$$

Thus, we get our second feedback model called **FM1**:

$$p_f(w \mid q) = \sum_{k} p_s(w \mid t_k) p(t_k \mid \theta_F) \qquad (3.14)$$

The signature feedback model $p(t_k \mid \theta_F)$ could be estimated using the EM algorithm [4] with the following update formulas.

$$\hat{p}^{(n)}(t_k) = \frac{(1-\alpha)p^{(n)}(t_k \mid \theta_F)}{(1-\alpha)p^{(n)}(t_k \mid \theta_F) + \alpha p(t_k \mid C)} \quad (3.15)$$

$$p^{(n+1)}(t_k \mid \theta_F) = \frac{c(t_k, F)\hat{p}^{(n)}(t_k)}{\sum_i c(t_i, F)\hat{p}^{(n)}(t_i)} \quad (3.16)$$

The approach proposed in [17] uses the generative model directly to estimate the feedback model, whereas we use the same approach to estimate the signature feedback model first, and then translate context-sensitive topic signatures to its own concepts. This difference may result in two advantages. First, a term in a different context will be treated the same and counted together in [17]. However, the context of topic signatures will be accounted for in our approach. Second, our approach will favor terms that frequently interact with other terms. We think it is better than simply accounting for the frequency of terms (even after removing the effect of the background model), as was done in [17], when estimating a query model. Roughly, a term with high occurrence frequency will also interact with other terms frequently, but in a fine sense, these two concepts are different.

4. Experimental Results

4.1 Test Collection and Evaluation

Our current implementation of topic signature extraction relies on a domain ontology. For this reason, we validate our context-sensitive semantic smoothing method on genomic collections because UMLS could be used as the domain ontology for this area.

The testing collections are TREC Genomic Track 2004 [7] and 2005 [8]. The original collection is a ten-year subset of Medline abstracts and contains about 4.6 million abstracts. We only used the sub-collection (i.e., the human relevance-judged document pool, 48,753 documents for 2004 and 41,018 documents for 2005) for our experiment. The ad hoc retrieval tasks of the two tracks include 50 topics (queries), respectively. We use the simple language model introduced in [12] (i.e., DM0) as the baseline. To give readers the sense of how good the baseline language model is, we also report the performance of the Okapi retrieval model in Table 1. Roughly, the performance of the baseline language model is comparable to that of the Okapi model. Following the convention of TREC, we use the mean average precision (MAP) as the major performance measure and the overall recall at 1000 documents as a supplemental measure.

Table 1. Comparison of the baseline language model to the Okapi model. The Okapi formula is the same as the one in [10]. The number of relevant documents for TREC04 and TREC05 are 8266 and 4585, respectively. The asterisk indicates the initial query is weighted as described in Section 4.2.

Collection	Recall			MAP		
	SLM	Okapi	Change	SLM	Okapi	Change
TREC04	6411	6662	+3.9%	0.345	0.363	+5.2%
TREC04*	6527	6704	+2.7%	0.364	0.364	+0.0%
TREC05	4084	4124	+1.0%	0.255	0.250	-2.0%
TREC05*	4135	4134	-0.0%	0.260	0.254	-2.3%

4.2 Indexing Schema and Query Processing

We index all documents with UMLS-based concepts and topic signatures as shown in Figure 1. For each document, we record the frequency count of each concept and signature and the basic statistics. For each concept and topic signature, we record their frequency count in each document and the basic statistics. For concept indexing, we do not use any stop list. For topic signatures appearing in more than one document, we estimate their translation models using the EM algorithms detailed in Section 3.1.

The query formulation is fully automated. The extraction of query terms from topic descriptions is the same as the process of document indexing. In TREC04 Genomics Track, a topic was described in three sections: title, information need, and context. The "context" section provided the background information of the topic. Assuming the background information could be learned from blind feedback, we intentionally ignore this section during query formulation. The final formulated query contains 4.3 terms on average. TREC05 Genomics Track provided more structured queries that look like a binary relation between two topic concepts. Because the queries are too short, we also include sub-concepts in the query. The final formulated query contains 5.1 terms on average (Query #135 was removed because it contains no relevant document).

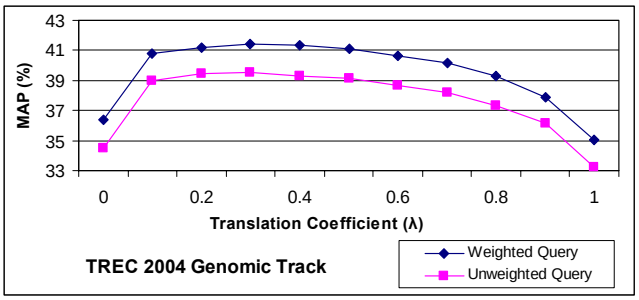
As stated in [12], the concepts in the "title" section are clearly more important than those in the remaining sections. For this reason, we weight query terms according to the sections from which they are extracted. Following the method proposed in [12], we optimize the weight of different sections by maximizing the MAP of the baseline retrieval model. The weights for the "title" section and the "information need" section are 1.0 and 0.6, respectively. In TREC 2005 Genomic Track, the topic description is presented in one section, but we found that the major concepts are more important than those sub-concepts. Similarly, we weight the query terms according to whether they are sub-concepts or not. The method for weight optimization is the same as that for query section weighting. The weights for major concepts and the sub concepts are 1.0 and 0.2, respectively. In Table 1-6, the asterisk (*) indicates the initial query is weighted.

4.3 Effect of Document Smoothing

We evaluate the document model with context-sensitive semantic smoothing (i.e., DM2). The coefficient (α) controlling the influence of the background collection model in all document models, DM0-DM2, is set to 0.05 in this paper. The translation coefficient (λ) in DM2 is optimized by maximizing MAP. The coefficient accounting for background noise is set to 0.3 when using an EM algorithm to train signature translation models. The result is shown in Table 2. The IR performance is significantly improved for both TREC04 and TREC05 after adopting semantic smoothing on document models.

Table 2. The comparison of the baseline language model (DM0) to document smoothing model (DM2) and query smoothing model (FM1).

Collection		DM0	λ=0.3		γ =0.6	
			DM2	Change	FM1	Change
TREC04	MAP	0.345	0.395	+14.5%	0.451	+30.9%
	Recall	6411	6749	+5.3%	6929	+8.0%
TREC04*	MAP	0.364	0.414	+13.7%	0.460	+26.9%
	Recall	6527	6905	+5.8%	7039	+7.8%
TREC05	MAP	0.255	0.277	+8.6%	0.279	+9.4%
	Recall	4084	4167	+2.0%	4227	+3.5%
TREC05*	MAP	0.260	0.288	+10.8%	0.287	+10.4%
	Recall	4135	4214	+1.9%	4235	+2.4%



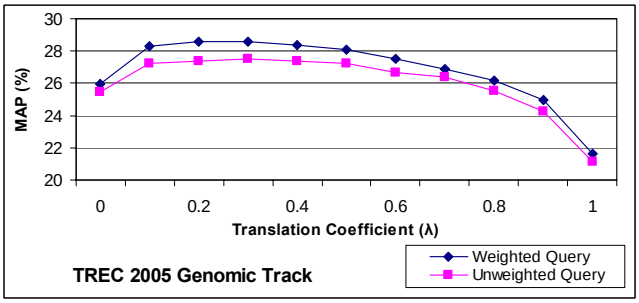
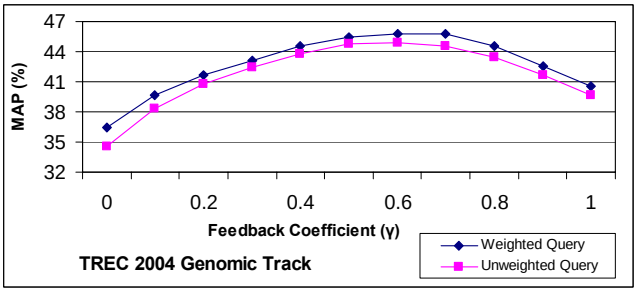


Figure 2. The variance of MAP with the translation coefficient (λ), which controls the influence of the translation model in DM2.

The variance of MAP with the translation coefficient λ is shown in Figure 2. For all four curves, the best performance is achieved at λ=0.3; after that point, the performance is downward. A possible explanation is that the extracted topic signatures do not capture all points of the document, but the basic language model captures those missing points. For this reason, when the influence of the translation model is too high in the mixture model, the performance is downward and even worse than that of the baseline. Therefore, if we can find a better topic signature representation for documents and queries, or we can refine the extraction of topic signatures, the IR performance might be further improved.

4.4 Effect of Query Smoothing

The blind feedback gives the chance to estimate an accurate query model and is thus expected to perform better than the baseline language model. We select the top 50 documents for feedback using Model FM1; the coefficient accounting for background noise is set to 0.3 when using the EM algorithm to train signature feedback models. For the efficiency of retrieval, we only expand 10 top-ranked terms and then renormalize their probability. Expanding more terms will only slightly improve the results but will seriously affect the retrieval efficiency. The feedback query model is further interpolated with the initial query model (QM0); the feedback coefficient (γ) is optimized by maximizing MAP.



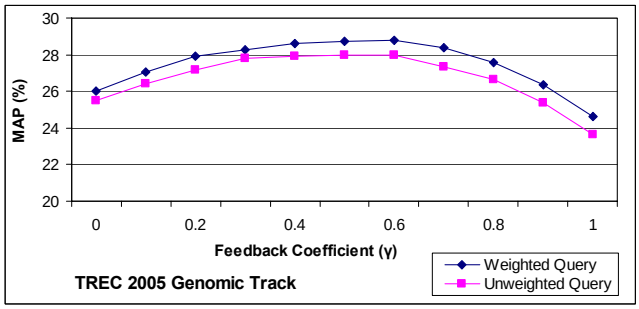


Figure 3. The variance of MAP with the feedback coefficient (γ), which controls the influence of the feedback model in blind feedback (i.e. DM0+FM1).

The comparison to the baseline language model is shown in Table 1. The feedback significantly raises the IR performance. The effect of feedback is also robust. As shown in Figure 3, the feedback model is always superior to the baseline when the feedback coefficient γ is changed from 0 to 1 for TREC04, and is better than the baseline except at γ=1.0 for TREC05.

The document smoothing and query smoothing are effective for both TREC04 and TREC05. However, the effect on TREC04 is clearly much more significant than on TREC05. A possible explanation is that TREC04 is "easier" than TREC05.

4.5 Interaction Effect of Document Smoothing and Query Smoothing

The document semantic smoothing maps related document terms to query terms while the query semantic smoothing (feedback) expands query terms to match document terms. Their effects will overlap to some degree when used together and therefore it is not expected to achieve the overall effect equal to the summation of both. Actually, the overall effect could be worse than each individual effect without careful parameter tuning. For example, "*hypertension*" and "*obesity*" can translate to each other with high probabilities; the initial query contains "*obesity*" and the feedback expands the term "*hypertension.*" In this case, if we still use document semantic smoothing in the feedback search, we may overestimate the importance of "*obesity*" to the original query and thus degrade the performance.

Table 3. The interaction effect of document smoothing (DM2) and query smoothing (FM1). "Max" is the maximum effect achieved by DM2 or FM1. "Both" is the result of DM2+FM1. "Change[1]" is the improvement of DM2+FM1 over DM0. "Change[2]" is the improvement of DM2+FM1 over "Max".

Collection		DM0	Max	Both	Change[1]	Change[2]
TREC04	MAP	0.345	0.451	0.461	+33.6%	+2.2%
	Recall	6411	6929	7026	+9.6%	+1.4%
TREC04*	MAP	0.364	0.460	0.470	+29.1%	+2.2%
	Recall	6527	7039	7079	+8.5%	+0.6%
TREC05	MAP	0.255	0.279	0.295	+15.7%	+5.7%
	Recall	4084	4227	4273	+4.7%	+1.1%
TREC05*	MAP	0.260	0.288	0.313	+20.4%	+8.7%
	Recall	4135	4235	4317	+4.4%	+1.9%

For the above considerations, we take the advantage of document semantic smoothing during the initial search and hope the top-ranked documents will be more relevant to the query and

result in a more accurate feedback model. In the feedback search, we still use document smoothing, but set its influence to a small degree to avoid overestimation. The feedback coefficient γ is set to 0.6 and the translation coefficient λ for the initial search is set to 0.3 according to the performance curves shown in Figure 2 and 3. The λ for feedback search is optimized by maximizing MAP. As expected, the optimal value ranges from 0.01 to 0.05. The final result is reported in Table 3. The interaction of document smoothing and query smoothing consistently achieves positive effect on the retrieval of TREC04 and TREC05.

However, the interaction effect on TREC05 is much more significant than on TREC04. It is most likely because the top-ranked documents returned by the basic language model (i.e., without document semantic smoothing) on TREC04 are good enough to estimate an accurate feedback model. In general, the worse the performance of the basic language model, the more significant the interaction effect will be.

4.6 Comparison of Feedback Models

The feedback model FM0 heuristically selects the topic signatures relevant to the query using term associations. FM1 uses a formal generative model to estimate the importance of each signature to the query and thus is expected to perform better than FM0 in terms of predicting the query. The comparison of these two feedback models is shown in Table 4. Though both are effective, FM1 performs consistently better than FM0, as expected.

Table 4. Comparison of blind feedback model FM1 to FM0

Collection	Recall			MAP		
	FM0	FM1	Change	FM0	FM1	Change
TREC04	6808	6929	+1.7%	0.442	0.451	+2.0%
TREC04*	6811	7039	+3.3%	0.449	0.460	+2.4%
TREC05	4192	4227	+0.8%	0.270	0.279	+3.3%
TREC05*	4215	4235	+0.5%	0.279	0.288	+3.2%

4.7 Context-Sensitive vs. Context-Insensitive

Following the method proposed in [1] and [3], we can simply use the extracted topic signatures to estimate a context-insensitive translation model, i.e., mapping one concept to another:

$$p'_k(w) = (1-\alpha)\frac{c(t(w,w_k))}{\sum_i c(t(w_i,w_k))} + \alpha p(w \mid C) \qquad (4.1)$$

where $c(t(w_i,w_k))$ is the frequency count of signature $t(w_i,w_k)$ in the whole collection. Then we get a context-insensitive version of DM2 denoted as DM2'.

The results of the comparative experiment on DM2 and DM2' are presented in Table 5. The translation coefficient (λ) is optimized by maximizing the MAP. The optimal λ is 0.3 for DM2 and 0.01 for DM2'. The optimal λ for DM2' is extremely small, most likely due to two reasons. First, the context-insensitive smoothing does not capture the semantics of the query well, and thus the influence in the mixture model is downward. Second, a topic signature in DM2' translates to a relatively small number of concepts and thus the average translation probability is much higher than in DM2.

The context-sensitive semantic smoothing approach performs significantly better than context-insensitive semantic smoothing approaches. The gain of DM2' over the baseline language model is consistent with the conclusions of previous work, such as [1] and [3]. [1] achieved 3-4% gain using HAL relationships and [3] achieved 5-6% gain using WordNet relationship and cooccurrence relationship.

Table 5. Comparison of the context-sensitive semantic smoothing (DM2) to the context-insensitive semantic smoothing (DM2') on MAP. The rightmost column is the change of DM2 over DM2'.

Collection	DM0	DM2'		DM2		Change
	MAP	MAP	Change	Map	Change	
TREC04	0.346	0.367	+6.1%	0.395	+14.5%	+7.6%
TREC04*	0.364	0.384	+5.5%	0.414	+13.7%	+7.8%
TREC05	0.255	0.260	+2.0%	0.277	+8.6%	+6.5%
TREC05*	0.260	0.269	+3.5%	0.288	+10.8%	+7.1%

4.8 Comparison to Other Approaches

We compared our method with two state-of-the-art approaches: the query expansion using information flow [15] and the model-based feedback [17].Both of them work within the LM framework. The former is also a context-sensitive semantic smoothing approach. The latter proves empirically to be effective on other TREC collections [17]. Because the information flow can not take the advantage of weighted initial queries, we used unweighted queries for comparisons. The information flow approach did not support concept-based indexing; thus, the result from only word-based indexing was obtained for it. For each approach, we tried different parameter combinations and reported the best result in Table 6.

Our approach achieved the best result for both 2004 and 2005. It performed significantly better than the local information flow approach, possibly because we imposed semantic constraint on topic signatures and used biological concepts rather than single words as building blocks, which made the information inference more meaningful on the genomic collections. Interestingly, the incorporation of domain knowledge did not help much when using the simple language model for retrieval; the result for TREC 2005 was even slightly worse. This showed that the context-sensitive semantic smoothing using topic signatures provided an effective mechanism to incorporate domain knowledge. The result of the model-based feedback was also improved by using the concept-based indexing, but less effective than our approach, especially for TREC 2004.

Table 6. Comparison of the retrieval performance of six approaches on TREC genomic track 2004 and 2005. "Word" or "Concept" means the indexing unit used. The concept-based indexing is based on the UMLS Metathesaurus. All approaches are implemented by us.

IR Approaches	TREC 2004		TREC 2005	
	MAP	Recall	MAP	Recall
Simple Language Model (Word)	0.324	6328	0.258	4101
Simple Language Model (Concept)	0.345	6411	0.255	4084
Local Information Flow (Word)	0.378	6793	0.272	4220
Model-based Feedback (Word)	0.372	6742	0.279	4260
Model-based Feedback (Concept)	0.424	6896	0.290	4213
Topic Signature (Concept)	0.461	7026	0.295	4273

5. Conclusions and Future Work

In this paper, we propose a novel context-sensitive semantic smoothing approach that decomposes a document and a query into a set of weighted context-sensitive topic signatures and then translate those topic signatures into query terms. We validated the

approach on two genomics collections: TREC Genomic Track 2004 and 2005. The document smoothing, the query smoothing, and the interaction of both all proved to be effective and robust on the testing collections in comparison to the baseline language model. We also implemented a context-insensitive version of semantic smoothing using extracted topic signatures. As expected, it is significantly less effective than the context-sensitive semantic smoothing, though it does achieve a slight improvement over the baseline language model. Our approach was also compared to two other IR approaches, a context-sensitive smoothing approach using information flow and an effective model-based feedback approach. Our approach performed significantly better than the former and slightly better than the latter. All experiments altogether concluded that the context-sensitive smoothing using topic signatures was effective to incorporate domain knowledge for genomic IR.

This paper made the following contributions. First, we presented a new document representation, i.e., representing a document as a set of weighted topic signatures and concepts. In particular, we chose concept pairs as topic signatures and adopted a generic ontology-based approach to extract concepts and concept pairs. The new representation could be applied to other retrieval, summarization, and text classification techniques. Second, we proposed an EM-based method to train the context-sensitive translation model for each signature and then formalized the query and document expansions based on signature translations. Third, we empirically proved the superiority of the context-sensitive semantic smoothing over context-insensitive semantic smoothing as well as non-semantic smoothing.

Our current implementation of topic signature extraction relies on a domain ontology. For this reason, we only tested our method on two genomic collections because UMLS can be used as the domain ontology for this area. However, the proposed method could be applicable to any application domain. For future work, we will adopt other existing concept and relation extraction approaches (i.e., those without ontologies) and apply context-sensitive semantic smoothing to more IR collections in general domains.

6. Acknowledgement

This work is supported in part by NSF Career Grant IIS 0448023, NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), and PA Dept of Health Grant (No. 239667). We also thank four anonymous reviewers for their comments on the paper.

7. References

[1] Bai, J., Song, D., Bruza, P., Nie, J.Y., and Cao, G., "Query Expansion Using Term Relationships in Language Models for Information Retrieval", In *Proceedings of the ACM 14th Conference on Information and Knowledge Management (CIKM)*, November 2005, Bremen, Germany.

[2] Berger, A. and Lafferty J., "Information Retrieval as Statistical Translation", In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in IR*, 1999, pp.222-229.

[3] Cao, G., Nie, J.Y., and Bai, J., "Integrating Word Relationships into Language Models", *Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 2005, pp. 298 - 305

[4] Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, 1977, 39: 1-38.

[5] Grefenstette, G., "Use of syntactic context to produce term association lists for information retrieval", *Proceedings of the 15th annual international ACM SIGIR conference on Research and Development in Information Retriev*a, 1992, pp. 89 - 97

[6] Harabagiu, S. and Lacatusu, F., "Topic themes for multi-document summarization"*, 2005 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*, 2005, pp. 42-48

[7] Hersh, W. et al. "TREC 2004 Genomics Track Overview", the *Thirteenth Text Retrieval Conference*, 2004.

[8] Hersh, W. et al. "TREC 2005 Genomics Track Overview", the *Fourteenth Text Retrieval Conference*, 2005.

[9] Jin, R., Hauptmann, A., and Zhai, C., "Title Language Model for Information Retrieval", *2002 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*, 2002, pp. 42-48

[10] Lafferty, J. and Zhai, C., "Document Language Models, Query Models, and Risk Minimization for Information Retrieval", In *Proceedings of the 24th ACM SIGIR Conference on Research and Development in IR*, 2001, pp.111-119.

[11] Liu, X. and Croft, W.B., "Cluster-based retrieval using language models", In *Proceedings of the 24th ACM SIGIR Conference on Research and Development in IR*, 2001, pp.186-193.

[12] Miller, D., Leek, T., and Schwartz M.R., "A Hidden Markov Model Information Retrieval System", In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, 1999, pp 214-221.

[13] Mooney, R. J. and Bunescu, R. "Mining Knowledge from Text Using Information Extraction", *SIGKDD Explorations* (special issue on Text Mining and Natural Language Processing), 7, 1 (2005), pp. 3-10.

[14] Ponte, J. and Croft, W.B., "A Language Modeling Approach to Information Retrieval", In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in IR*, 1998, pp.275-281.

[15] Song, D. and Bruza P.D., "Towards Context-sensitive Information Inference", *Journal of the American Society for Information Science and Technology (JASIST)*, 2003, Vol. 54, 321-334.

[16] Zhai, C. and Lafferty, J., "A Study of Smoothing Methods for Language Models Applied to Ad hoc Information Retrieval", In *Proceedings of the 24th ACM SIGIR Conference on Research and Development in IR*, 2001, pp.334-342.

[17] Zhai, C. and Lafferty, J., "Model-based Feedback in the Language Modeling Approach to Information Retrieval", In *Proceedings of the 10th International Conference on Information and Knowledge Management*, 2001, pp.403-410.

[18] Zhai, C. and Lafferty, J., "Two-Stage Language Models for Information Retrieval**"**, *2002 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*, 2002.

[19] Zhou, X., Hu, X., Lin, X., Han, H., and Zhang, X., "Relation-based Document Retrieval for Biomedical Literature Databases", *The 11th International Conference on Database Systems for Advanced Applications (DASFAA 2006)*, 12 - 15 April, 2006, Singapore, pp. 689-701

[20] Zhou, X., Zhang, X., and Hu, X., "Using Concept-based Indexing to Improve Language Modeling Approach to Genomic IR", *The 28th European Conference on Information Retrieval (ECIR' 2006)*, 10 - 12 April, 2006, London, UK, pp. 444-455.