Topic Signature Language Models for Ad-hoc Retrieval

Xiaohua ZHOU, Xiaohua HU, and Xiaodan ZHANG

***Abstract*-- Semantic smoothing, which incorporates synonym and sense information into the language models, is effective and potentially significant to improve retrieval performance. The previously implemented semantic smoothing models, such as the translation model which statistically maps document terms to query terms, and a number of other works that have followed have shown good experimental results. However, these models are unable to incorporate contextual information. Thus, the resulting translation might be fairly general and contains mixed topics. To overcome this limitation, we propose a novel context-sensitive semantic smoothing (CSSS) method that decomposes a document into a set of weighted context-sensitive topic signatures and then translate those topic signatures into query terms. The language model with such a context-sensitive semantic smoothing is referred to as the topic signature language model in this paper. In detail, we implement two types of topic signatures depending on whether an ontology exists in the application domain or not. One is concept extracted by an ontology-based approach; the other is multiword phrase extracted by Xtract if there is no ontology available in the application domain. The translation probabilities from each topic signature to individual terms are estimated through the EM algorithm. Document models based on topic signature translation are then derived. The new smoothing method is evaluated on TREC 2004/2005 Genomics Track based on the ontology-based concept, and TREC Ad hoc Track (Disk 1, 2 and 3) using multiword phrases. Both experiments show significant improvements in terms of average precision and overall recall over the two-stage language model (TSLM) as well as the language model with context-insensitive semantic smoothing (CISS).**

***Index Terms*-- Information Retrieval, Language Models, Semantic Smoothing, Topic Signature, Concept Pair, Multiword Phrase.**

I. Introduction

The language modeling approach to information retrieval (IR), initially proposed by Ponte and Croft [21], 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 due to the sparsity of training data. On one hand, because the query terms may

Manuscript received October 18, 2006. 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).

X. Zhou is with the College of Information Science & Technology, Drexel University, Philadelphia, PA 19104 USA (e-mail: xiaohua.zhou@ drexel.edu).

X. Hu is with the College of Information Science & Technology, Drexel University, Philadelphia, PA 19104 USA (telephone: 215-895-0551 e-mail: thu@ cis.drexel.edu).

X. Zhang is with the College of Information Science & Technology, Drexel University, Philadelphia, PA 19104 USA (e-mail: xzhang@ cis.drexel.edu).

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 collection model or even irrelevant noise. Thus, the core of the language modeling approach to IR is to "smooth" document models. Zhai and Lafferty [26, 28] propose several effective background 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 [15]. 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 is unable to incorporate sense and 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 et al. [14] and Cao et al. [4] present two other ways to train the translation probabilities between individual terms, but their approaches still suffer the same context-insensitivity problem as [2]. Thus, it is urgent to develop a framework to semantically smooth document models in the language modeling (LM) retrieval framework.

In this paper, we propose a novel context-sensitive semantic smoothing (CSSS) method based on topic decomposition. A document is decomposed into a set of weighted topic signatures and then those topic signatures are translated into individual terms for the purpose of document expansions. We define a topic signature as either an ontology-based concept or an automated multiword phrase. Because a concept or a multiword phrase itself contains contextual

information and its is usually unambiguous, the translation from topic signatures to individual terms should have higher accuracy and result in better retrieval performance, compared to the semantic translations between single words. 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 develop an ontology-based algorithm to extract concept-based topic signatures and adopt an existing algorithm referred to as Xtract [23] to identify phrase-based topic signatures. Furthermore, we develop an EM-based algorithm to estimate probabilities of translating each topic signature into individual terms in the vocabulary. The new smoothing method is tested on collections from two different domains in order to show its robustness. The extraction of concepts needs domain ontology. Thus we evaluate the effectiveness of concepts on TREC Genomic Track 2004/2005. The extraction of multiword phrases does not need any external human knowledge and can be applied to any public domains. Therefore we test the effectiveness of multiword phrases on TREC Disk 1, 2, and 3. The experimental results show that significant improvements are obtained over the two-stage language model (TSLM) [28] as well as the language model with context-insensitive semantic smoothing (CISS).

The contribution of this paper is three-fold. First, it proposes a new document representation using a set of weighted terms and topic signatures. The new scheme also explores the relationship between individual terms and more complicated topic signatures. Second, it develops an EM-based algorithm to estimate the semantic relationships between topic signatures and individual terms and further uses those semantic relationships to smooth the document model, which is referred to as context sensitive semantic smoothing (CSSS) in this paper. The smoothed document models can be used not only for text retrieval, but also for many other text

mining applications such as text clustering. 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 related to topic signatures. In Section 3, we first formally define topic signatures and present the methods for the topic signature extraction; then we describe in details the method of context-sensitive semantic smoothing. Section 4 shows the experimental results on TREC 2004/2005 Genomics Track collections, where topic signatures are implemented as concept pairs. Section 5 shows the experimental results on TREC Disk 1, 2 and 3, where multiword phrases are used as topic signatures. Section 6 concludes our paper.

II. RELATED WORK

The idea of topic decomposition and translation for language modeling IR is not new. It was used for query expansion as well as document expansion in literature. Song and Bruza adopted information flow (IF) for query expansion in [24]. The context of a concept is represented by a HAL vector; the degree of one concept inferring another can then be computed through vector operators. Song and Bruza also invented a heuristic approach to combine multiple concepts, which enabled information inference from a group of concepts (premises) to one individual concept (conclusion). Thus, their query expansion technique was somehow context-sensitive. However, it could not be used to expand (smooth) document models. Besides, the degree to which one individual concept could be inferred from another combined concept was not theoretically motivated; its robustness needs to be further validated.

Similarly, Bai et al. [1] used significant term pairs to expand query models. The combination of two terms is helpful to disambiguate their context and thus can capture more sense of the query. The expanded query model based on significant term pairs looked like as follows:

$$p(w \mid Q) = (1-\lambda) \sum_{q_i, q_j \in Q} p_R(w \mid q_i q_j) p(q_i q_j \mid Q) + \lambda p_{ML}(w \mid Q) \quad (1)$$

Here the first term is a unigram query model for smoothing purpose and the second term (query expansion) is based on topic decomposition and translation. The topic decomposition term $p(q_i q_j \mid Q)$ is simply assumed to be uniformly distributed. The topic translation term $p_R(w \mid q_i q_j)$ is estimated based on term co-occurrence statistics. The coefficient λ controls the influence of the expansion component. Like the information flow approach, this approach is also inappropriate for document model expansions because the distribution of term pairs in a document is obviously not uniform. Besides, the co-occurrence based estimation algorithm tends to assign higher probability values to general terms than specific terms.

Berger and Lafferty proposed the statistical translation model for the first time in [2]. With this model, a term in a document is statistically mapped to query terms as described in the formula below:

$$p(q \mid d) = \sum_w t(q \mid w) l(w \mid d) \quad (2)$$

where *t(q|w)* is the translation probability from document term w to query term q and *l(w|d)* is the unigram document model. The translation model achieved significant improvement over the simple language model on two TREC collections [2]. However, the model only captures the semantic relationship between individual words and is unable to incorporate the contextual information into the translation procedure. In addition, the training of translation probability requires a large number of real query-document pairs, which are very difficult to obtain. For this reason, Berger and Lafferty used synthetic data in the experiment. Besides, a document often contains a considerable number of unique terms and thus the model expansion through

translation is computationally intensive.

The cluster language model [16] may be the first trial of topic decomposition and translation for document model expansions. Liu and Croft [16] incorporated cluster information into document model estimation:

$$p(w \mid d) = \frac{N_d}{N_d + u} p_{ML}(w \mid d) + (1 - \frac{N_d}{N_d + u}) p(w \mid cluster) \quad (3)$$

N_d is the length of the document and μ is a parameter for smoothing. The document clusters are very similar to our topic signatures in the sense that both use a set of documents with similar context rather than a single document to estimate a more accurate topic model. However, in their cluster model, a document is associated with a single cluster, which may become problematic for especially long documents, whereas a document can have multiple topic signatures in our model. Furthermore, the clustering for a large collection is extremely inefficient. Last, lots of 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 these problems.

Latent topic models such as pLSI [13] assume that a document is generated by a set of topic models with certain distribution. Each topic model is further about the distribution of words in a given vocabulary. With topic model assumption, a document is modeled as follows:

$$p(w \mid d) = \sum_{i=1}^{k} p(t_i \mid d) p(w \mid t_i) \quad (4)$$

Here k is the total number of topics in the corpus. The parameter $p(w \mid t_i)$ is the probability of topic t_i generating word w. The parameter $p(t_i \mid d)$ is the probability of document d being generated by topic t_i. Within the framework of latent topic models, a document can be associated

with multiple topics and thus it overcomes the limitation of the cluster language models. Hoffman evaluated the pLSI model for retrieval tasks within the framework of vector space model [13]. The pLSI model significantly outperformed the LSI model as well as the standard raw term matching method. However, the size of four testing collections is far from the representative of realistic IR environments and the baseline model is also far from the state of the art, making the effectiveness of the pLSI model on retrieval unclear.

The idea of topic signature is actually very similar to the latent topic. The major difference lies in their implementations, i.e. the estimation of parameters. The number of free parameters $p(t_i \mid d)$ and $p(w \mid t_i)$ in the latent topic models is mainly in proportion to the number of documents for a large collection, which will cause serious overfitting problem when Expectation Maximum (EM) algorithm [8] is used for model estimations. The estimation process also lacks scalability because all parameter should be estimated simultaneously. The worst is that when a new document is coming, there is no way to estimate the topic mixture $p(t_i \mid d)$. In our approach, we explicitly extract topic signatures from documents in the corpus. Thus, we can estimate each topic signature model $p(w \mid t_i)$ separately. Furthermore, we can simply use maximum likelihood estimator to approach $p(t_i \mid d)$ no matter the document is new or not. In short, the estimation of the topic signature language model is very efficient and scalable as well as applicable to new testing documents.

Wei and Croft [25] proposed a LDA-based document model for ad-hoc retrieval. Unlike the pLSI model where topic mixture is conditioned on each document, the LDA model samples topic mixture from a conjugate Dirichlet prior that remains same for all documents [3]. This change can solve the overfitting problem and the problem of generating new document in pLSI. To make up the possible information loss, the LDA model is further interpolated with a simple

language model. The final document model is:

$$p(w \mid d) = \lambda\left(\frac{N_d}{N_d + u} p_{ML}(w \mid d) + \left(1 - \frac{N_d}{N_d + u}\right) p(w \mid coll)\right) + (1 - \lambda)\sum_{i=1}^{k} p(t_i \mid d) p(w \mid t_i) \qquad (5)$$

The LDA model improved the retrieval performance of both simple language model and the cluster language model on five TREC collections [25]. The LDA model is estimated through Gibbs sampling which is computationally intensive. Thus, compared to the topic signature language model, the LDA model still suffers from the computing intensity as well as lack of scalability.

III. TOPIC SIGNATURE LANGUAGE MODELS

In this section, we describe topic signature language models in details. First, we define two types of topic signatures and introduce the extraction algorithms. Second, a statistical model (i.e. a distribution of words) is estimated for each topic the corresponding topic signature represents. Third, topic signature models are used for the document expansion (smoothing). Last, we discuss the scalability and complexity of the estimation of the topic signature language model.

A. Context-Sensitive Topic Signatures

The implementation of topic signatures plays a crucial role in our context-sensitive semantic smoothing approach. First, the topic signature must be context-sensitive and thus it should contain at least two terms, unless word sense is adopted. Second, sub-terms of a topic signature should have syntactic relation. Otherwise, we cannot count their frequency in a document and it becomes difficult to estimate their distributions. Third, it should be easy and efficient to extract topic signatures from texts. Following these criteria, we recommend two types of topic

signatures. One is the ontology-based concept and the other is the multiword phrase. In this sub-section, we formally define these two types of topic signatures and briefly introduce the extraction algorithms.

1) Ontology-based Concept as Topic Signature

In our previous work [32], we implemented topic signatures as concept pairs inspired by Harabagiu and Lacatusu's topic representations [10]. Formally, a topic signature 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. Because two concepts in a pair help to determine the context for each other, the meaning of a concept pair is often unambiguous and its semantic translation to individual concepts is very specific and accurate. However, the combination of two concepts causes a large vocabulary space which makes it inefficient to index large collections. The distribution of concept pairs is also quite sparse and thus it is difficult to obtain sufficient data for many concept pairs to estimate their translation probabilities to individual concepts. Aware of the unambiguousness of a concept in an ontology, we simply use ontology-based concept as topic signatures in this paper.

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 synonymy 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 [30].

In general, the extraction of concepts from texts is still a challenging problem. Fortunately, in

the domain of biology and medicine, a large ontology called UMLS [35] is developed, which makes the task of concept extractions easier. The extraction of biological concepts is a hot topic in bioinformatics and a survey of those methods can be found in [19]. However, most approaches segment a sequence of words into phrases, but do not further map the identified phrases into concepts. For this reason, we adopt MaxMatcher [31], a dictionary-based biological concept extraction tool, for UMLS concept extractions.

In order to increase the extraction recall while remaining the precision, MaxMatcher uses approximate matches between the word sequences in text and the concepts defined in a dictionary or 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. In the example shown in Figure 1, the underlined phrases are extracted concept names followed by the corresponding concept ID and semantic type. The details of the algorithm for MaxMatcher can be found in our previous work [31]. MaxMatcher has been evaluated on the GENIA corpus [36]. The precision and recall reached 71.60% and 75.18%, respectively, using approximate match criterion.

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

Word Index: *recent, epidemiological, study, research, activity, reveal, obesity, independent, risk, factor, periodontal, disease*
Concept Index: *C0002783, C0028754, C0031090*

Fig. 1. The demonstration of concept extraction and indexing. Stop words are removed and words are stemmed.

2) Multiword Phrase as Topic Signature

The use of phrases has a long history in information retrieval. A typical method for utilizing phrases will identify phrases within queries (e.g., “star war”, “space program”), scan documents to identify query phrases, and score the document if it contains query phrases [20]. The recognition of query phrases within documents can be done in one of the following three manners [20]:

- Boolean: it is also called conjunctive phrases [5]. All subterms of a query phrase cooccur in a document.
- Adjacent: Exact same form as the query phrase.
- Proximity: All subterms of a query phrase occur in close proximity in a document.

In this paper, we utilize multiword phrases in a different manner. We treat phrases frequently occurring in a given collection as topic signatures and try to find a set of individual words to represent the topic signature (the multiword phrase). Then we can expand a document language model by statistically mapping topic signatures into query terms (individual words). For this purpose, we identify multiword phrases within only documents. The definition of phrase in this paper is roughly equivalent to the definition of query phrases in traditional phrase models. It is sort of rigid noun phrase or collocations. It contains two or more individual words which are adjacent to each other in sequence. It often begins with an adjective or a noun and ends with a noun. The semantics of a phrase usually has the following types:

- Organization: International Business Machine Corp.
- Person: George Bush, Ronald Regan
- Location: United States, Los Angels
- Subject: Space Program, Star War

We use a slightly modified version of Xtract [23] to extract phrases in documents. Xtract is designed to extract three types of collocations: predicative relations, rigid noun phrases, and phrasal templates. It begins with extracting significant bigrams using statistical techniques, and then expands 2-Grams to N-Grams, and finally adds syntax constraint to the collocations. In Fagan's notion of phrases [5] [9], the phrases extracted by Xtract are kind of "syntactic" phrases because it imposes both statistical and syntactic constraint on phrases. In the original version,

two words are defined as a bigram if and only if they cooccur within a sentence and their lexical distance is less than five words. Because we are only interested in rigid noun phrases, the first word should be an adjective or a noun, the second word should be a noun, and their distance threshold is set to four words, in our implementation.

Xtract uses four parameters, strength (k_0), spread (U_0), peak z-score (k_1), and percentage frequency (T), to control the quantity and quality of the extracted phrases. In general, the bigger those parameters are, the higher quality but less quantity phrases Xtract produce. Smadja recommended a setting (k_0, k_1, U_0, T) = (1, 1, 10, 0.75) to achieve good results. In the experiment, we set those four parameters to (1, 1, 4, 0.75). Xtract is an effective approach to phrase extraction. The estimated precision is about 80%, which is good enough for our IR use. It is also very efficient. For example, it takes only two hours to extract phrases from the AP89 collection (84,678 documents) using our Java version implementation while Annie (a named entity recognition component of GATE [6]) takes about twelve hours to recognizes entities from the same collection.

Example Sentence:
How the many changes in the former Soviet Union (now the Commonwealth of Independent States) will affect the future of their space program remains to be seen.

Word index: *change, form, soviet, union, commonwealth, independent, state, affect, future, space, program, remain, see*
Multiword Phrase Index: *Soviet Union, independent state, space program*

Fig. 2. The demonstration of multiword phrase extraction and indexing. Stop words are removed and words are stemmed.

In the experiment, we also tried another two types of multiword phrases in order to increase phrase coverage. One is named entities (person, location, and organization) identified by GATE [6]. The other is WordNet noun phrases [18]. However, the extra phrases did not bring further improvement of IR performance. A possible explanation is that both GATE entities and WordNet noun phrases are purely "syntactic" phrases and those extra phrases (not extracted by Xtract) are often infrequent in our testing collections. In our phrase language model, infrequent

phrases (topic signature) result in little effect on document expansions.

B. Topic Signature Model Estimates

Suppose we have indexed all documents with individual terms and topic signatures (see Figure 3). 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 terms in the vocabulary. If all terms 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 terms address the issue of other topic signatures while some are background terms of the whole collection. We use the generative model proposed in [27] to remove 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 $p(w|C)$),

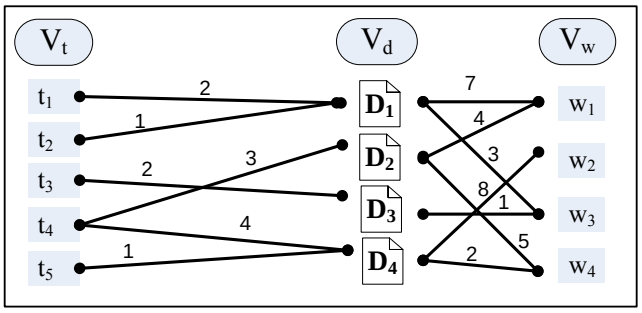


Fig. 3. Illustration of document indexing. V_t, V_d and V_w are topic signature set, document set and word set, respectively. The number on each line denotes the frequency of corresponding topic signature or word in the document.

$$p(w|\theta_{t_k},C) = (1-\alpha)p(w|\theta_{t_k}) + \alpha p(w|C) \qquad (6)$$

Here the coefficient α is accounting for the background noise and θ_{t_k} refers to the parameter set of the topic model associated with the topic signature t_k. In all the experiments in this paper, the background coefficient α is set to 0.5. Under this mixture language model, the log likelihood of generating the document set D_k is:

$$\log p(D_k|\theta_{t_k},C) = \sum_w c(w,D_k)\log p(w|\theta_{t_k},C) \qquad (7)$$

Here $c(w, D_k)$ is the document frequency of term w in D_k, i.e., the cooccurrence count of w and t_k in the whole collection. The topic model for t_k can be estimated using the EM algorithm [8]. The EM update formulas are:

$$\hat{p}^{(n)}(w) = \frac{(1-\alpha)p^{(n)}(w \mid \theta_{t_k})}{(1-\alpha)p^{(n)}(w \mid \theta_{t_k}) + \alpha p(w \mid C)} \qquad (8)$$

$$p^{(n+1)}(w \mid \theta_{t_k}) = \frac{c(w, D_k)\hat{p}^{(n)}(w)}{\sum_i c(w_i, D_k)\hat{p}^{(n)}(w_i)} \qquad (9)$$

Our topic signature model is significantly different from previous ones described in [2] [4] [14] [15] in two aspects. First, previous models take an individual term as the topic signature, and are unable to incorporate contextual information into the translation procedure. Our model uses phrases as the topic signatures. Since multiword phrases are unambiguous in most cases, the resulting translation will be more specific.

From three examples shown in Table I, we can see that the phrase-word translations are quite coherent and specific. Take the example of the phrase "space program". If we estimate the translation models for its constituent terms "space" and "program" separately, both translation models (see Figure 4) contain mixed topics and are fairly general. Some terms such as *NASA*, *astronaut*, *moon*, *satellite*, *rocket*, and *Mar*, which is related to the subject of space

TABLE I

Examples of topic signature models. The three multiword phrases are automatically extracted from the collection of AP89 by *Xtract*. We only list the top 20 topical words for each phrase. It is worth noting that the word "third" is removed from the index as a stop word and thus it does not appear in the translation result of the third phrase.

Space Program		**Star War**		**Third World Debt**	
Term	Prob.	Term	Prob.	Term	Prob.
space	0.101	star	0.088	debt	0.072
program	0.071	war	0.066	Brady	0.039
NASA	0.048	missile	0.06	loan	0.038
shuttle	0.043	strategy	0.051	world	0.038
astronaut	0.041	defense	0.051	treasury	0.037
launch	0.040	nuclear	0.043	bank	0.035
mission	0.038	space	0.034	Nicholas	0.034
flight	0.037	initialize	0.033	debtor	0.030
earth	0.037	Pentagon	0.032	trillion	0.027
moon	0.035	weapon	0.031	reduction	0.027
orbit	0.032	bomber	0.031	forgive	0.025
satellite	0.031	budget	0.028	monetary	0.025
Mar	0.030	stealthy	0.025	Mexico	0.025
explorer	0.028	program	0.025	economy	0.023
station	0.028	spend	0.024	billion	0.023
rocket	0.027	armed	0.023	reduce	0.022
technology	0.026	fiscal	0.022	burden	0.022
project	0.025	Reagan	0.021	lend	0.021
science	0.023	cut	0.021	creditor	0.021
budget	0.023	Bush	0.019	secretary	0.020

program very much, do appear in the phrase translation model, but in neither of the two subterm translation models.

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 [14], 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 [15] deals with translations in a different fashion. However, the resulting query model is fairly general and the computation of the inverse matrix is prohibitive to large collections. Cao [4] takes into account word semantics when computing term associations, but he ignores the sense of words.

Space:
space 0.245; shuttle 0.057; launch 0.053; flight 0.042; air 0.035; program 0.031; center 0.030; administration 0.026; develop 0.025; like 0.023; look 0.022; world 0.020; director 0.020; plan 0.018; release 0.017; problem 0.017; work 0.016; place 0.016; mile 0.015 base 0.014;

Program:
program 0.193; washington 0.026; congress 0.026; administration 0.024; need 0.024; billion 0.023; develop 0.023; bush 0.020; plan 0.020;money 0.020; problem 0.020; provide 0.020; writer 0.018; d 0.018; help 0.018; work 0.017; president 0.017; house .017; million 0.016; increase 0.016;

Fig. 4. The demonstration of word-word translation which is estimated by the same approach described in section 3.1. The translation results contain mixed topics and are fairly general in comparison with the result of the phrase-word translation.

We also truncate terms with extremely small translation probabilities for two purposes. First, with smaller number of translation space, the document smoothing will be much more efficient. Second, we assume terms with extremely small probability are noise (i.e. not semantically related to the given topic signature). In detail, we disregard all terms with translation probability less than 0.001 and renormalize the translation probabilities of the remaining terms.

C. Document Model Smoothing

Suppose we have indexed all documents in a given collection C with terms (individual words) and topic signatures as illustrated in Figure 3. The translation probabilities from a topic signature t_k to any individual term w, denoted as $p(w|t_k)$, are also given. Then we can easily obtain a

document model below:

$$p_t(w \mid d) = \sum_k p(w \mid t_k) p_{ml}(t_k \mid d) \qquad (10)$$

The likelihood of a given document generating the topic signature t_k can be estimated with

$$p_{ml}(t_k \mid d) = \frac{c(t_k, d)}{\sum_i c(t_i, d)} \qquad (11)$$

where $c(t_i, d)$ is the frequency of the topic signature t_i in a given document d.

We refer to the above model as translation model after Berger and Lafferty's work [2]. As we discussed in the previous sub-section, the translation from context-sensitive topic signatures to individual terms would be very specific. Thus, the smoothed (expanded) document models will be more accurate. However, not all topics in a document can be expressed by topic signatures (i.e., multiword phrases). Take the example of AP88-90. A document in this collection contains 179 unique words, but only contains 32 multiword phrases on the average (see Table II). If only translation model is used, there will be serious information loss. A natural extension is to interpolate the translation model with a unigram language model. We use two-stage method [28] to smooth the unigram language model:

$$p(Q \mid D) = \prod_{q \in Q} \{(1-\gamma) \frac{tf(q, D) + \mu p(q \mid C)}{\mid D \mid + \mu} + \gamma p(q \mid C)\} \quad (12)$$

where $p(q|C)$ is the collection background model. γ and μ are two coefficients for tuning. We also refer to this smoothed unigram model as simple language model or baseline language model in this paper.

The final document model for retrieval use is described in equation (13). It is a mixture model with two components: a simple language model and a translation model.

$$p_{bt}(w \mid d) = (1 - \lambda) p_b(w \mid d) + \lambda p_t(w \mid d) \qquad (13)$$

The translation coefficient (λ) is to control the influence of two components in the mixture model. With training data, the translation coefficient can be trained by optimizing the retrieval performance measure such as average precision. In the experiments in this paper, we train the optimal translation coefficient on one collection and then apply the learned translation coefficient to other collections.

D. Scalability and Complexity

In comparison to simple language models [18] and traditional probabilistic language models such as Okapi [22], the topic signature language model needs the following extra computational cost: (1) the extraction of topic signatures from documents in offline mode, (2) the estimation of topic models for each topic signature in offline mode, and (3) document model expansions based on topic signature translations in online mode. Fortunately, the additional computation is scalable very well and its complexity is acceptable in practice. Furthermore, the issue of scalability and complexity is significantly improved over the statistical translation model [2] and the LDA-based document model [25].

The extraction of topic signatures is time-consuming compared with individual term extraction. However, it does not cause serious problem because it can be executed in the offline and incremental mode. In the experiment, the dragon toolkit [34] is used for document indexing. The dragon toolkit implements a Java version of Xtract [23] for multiword phrase extraction. Take the example of indexing AP collection in Disk 1, 2 and 3 (about 240K news articles) on a Linux server. It takes about 15 minutes to index individual terms and 3 hours to index topic signatures (multiword phrases). From this example, we can see that the indexing time for topic signatures is acceptable as an offline task.

The estimation of topic models is highly computation-intensive. In general, the parameter space is in proportion to the number of documents in the corpus, the size of vocabulary, and the number of topics; the computational complexity is in proportion to the number of documents, the number of topics, and the number of iterations for convergence. Therefore, the estimation algorithms proposed in [2] and [25] are not scalable as well as time-consuming for large collections. For example, the estimation of the LDA model for AP collection using Gibbs sampling (please refer to [25] for detailed settings) costs about 72 hours whereas our approach uses only 45 minutes to estimate topic models for all topic signatures. Our approach estimate topic models for each topic signature separately, which dramatically reduce the parameter space and make the model converged with fewer iterations. Thus, our estimation approach increases the scalability and reduces the complexity.

TABLE II

Average numbers of unique words and topic signatures per document in four collections

Collection	avg. # of unique words	avg. # of unique topic signatures
Genomics 2004	71.3	39.2
Genomics 2005	75.2	37.6
AP89	180.1	31.8
AP88-89	178.6	31.7
WSJ90-92	196.6	35.6
SJMN91	164.2	25.3

The online document model expansion based on topic models is computationally intensive because it involves the summation of translation probabilities as shown in equation (9). The complexity is in proportion to the number of topics for a document. The number of topics is equal to number of unique terms in the statistical translation model [2], the number of latent topics in LDA-based models [25], and the number of unique topic signatures in the topic signature language model. As shown in Table II, the number of topic signatures is significantly less than the document length as well as the number of latent topics in LDA model (e.g. the optimal number of topics is 800 in [25]) in typical testing collections and thus our approach has the lowest complexity during document model expansions.

IV. EXPERIMENTS WITH ONTOLOGY-BASED CONCEPTS

A. Evaluation Metrics and Baseline Models

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. The non-interpolated average precision is defined as:

$$\frac{1}{|\mathrm{Rel}|}\sum_{D\in \mathrm{Rel}}\frac{|\{D'\in \mathrm{Rel}, r(D')\le r(D)\}|}{r(D)} \qquad (14)$$

where r(D) is the rank of document d and Rel is the set of relevant documents for a query Q. By averaging the non-interpolated average precision across all queries of a collection, we obtain the MAP for the collection.

In the experiment, we use the two-stage language model (TSLM) [28] as the first baseline. The exact formula for the two-stage model is described in equation 12. To show how strong the baseline is, we also compare the baseline to the famous Okapi model [22]. The exact formula for the Okapi model is shown below:

$$Sim(Q,D)=\sum_{q\in Q}\left\{\frac{tf(q,D)\log(\frac{N-df(q)+0.5}{df(q)+0.5})}{0.5+1.5\frac{|D|}{avg\_dl}+tf(q,D)}\right\} \qquad (15)$$

Where:

tf(q, D) is the term frequency of q in document D.

df(q) is the document frequency for q.

avg_dl is the average document length in the collection.

The major difference between the statistical translation model [2] and the proposed topic signature language model is that the latter incorporate the contextual information into the

document model expansions (smoothing). Thus, it is very natural to further compare the context-sensitive semantic smoothing (CSSS) to the context-insensitive semantic smoothing (CISS). Because it is difficult to obtain a large number of real query-document pairs, we use word-word cooccurrence data to train context-insensitive version of translation probabilities in the experiment. The parameter estimation algorithm is the same as the one for the context-sensitive version (i.e. the translation from topic signature to individual words). The retrieval model is still the mixture of a two-stage language model and a translation model as described in equation 13. But the translation component is formulated slightly differently:

$$p_t(w \mid d) = \sum_k p(w \mid w_k) p_{ml}(w_k \mid d) \qquad (16)$$

It statistically maps each individual word instead of context-sensitive topic signature in a document onto query terms.

B. Testing Collections

Our current implementation of concept-based topic signature extraction needs 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 [11] and 2005 [12]. 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, 42,251 documents for 2004 and 35,474 documents for 2005) for our experiment. The ad hoc retrieval tasks of the two tracks include 50 topics (queries), respectively. The statistics of the testing collections are shown in Table III.

TABLE III
The descriptive statistics of testing collections

Collections	Word	Concept	Rel./Doc	Q.Len/Q.#
Genomics 2004	92,362	65,257	8,268/42,251	6.4/50
Genomics 2005	80,168	57,879	4,584/35,474	6.0/49

C. Document Indexing and Query Processing

We index all documents with UMLS-based concepts and individual word as demonstrated in Figure 2. For each document, we record the frequency count of each topic signature (i.e. UMLS concept) and individual words and the basic statistics. For each topic signature and individual words, we record their frequency count in each document and the basic statistics. For word indexing, stop words are removed and each word is stemmed. For topic signatures appearing in ten or more documents, we estimate their topic models (i.e. translation probabilities) using the EM algorithms.

The query formulation is fully automated. The extraction of query terms (individual words) from topic descriptions is the same as the process of document indexing. In TREC 2004 Genomics Track, a topic was described in three sections: title, information need, and context. The information provided by section of context is a little noisy. Our pilot study showed that the baseline (both Okapi and two-stage language model) using context section achieved the performance much worse than the one without context. For this reason, we only use the title section and information need section in the experiment. In TREC 2005 Genomics Track, query #135 was removed because it contains no relevant document.

As stated in [17], the query terms 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 [17], we optimize the weight of different sections by maximizing the MAP of the baseline retrieval model. The optimal weights for the "title" section and the "information need" section are 1.0 and 0.2, respectively. In Table IV, V and VI, the sign (†) indicates the initial query is weighted.

D. Effect of Document Smoothing

We set parameters γ and μ in the two-stage language model to 0.05 and 200, respectively because the language model achieves the best performance with this configuration. 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 IV. The Okapi model is slightly better than the two-stage model, but roughly these two models are comparable to each other.

The translation coefficient (λ) in the topic signature language model is optimized by maximizing the MAP on TREC Genomics Track 04 using unweighted query. The learned optimal value is 0.3 and then we apply this learned value to other two collections. The result is shown in Table V. In order to validate the significance of the improvement, we also run paired-sample t-test. As expected, the topic signature language model outperforms the two-stage language model in terms of average precision and overall recall at the significance level of 0.01 on both TREC04 and TREC05.

TABLE IV
Comparison of the two-stage language model (TSLM) to the Okapi model. The sign† indicates the initial query is weighted.

Collection	Recall			MAP		
	TSLM	Okapi	Change	TSLM	Okapi	Change
TREC04	6544	6847	+4.6%	0.352	0.369	+4.8%
TREC04†	6680	6869	+2.8%	0.384	0.370	-3.7%
TREC05	4093	4193	+2.4%	0.265	0.270	1.9%

TABLE V
The comparison of the two-stage language model (TSLM) to the topic signature language model (i.e. context-sensitive semantic smoothing, CSSS). The sign ** and * indicate the improvement is statistically significant according to the paired-sample t-test at the level of $p<0.01$ and $p<0.05$, respectively. The sign† indicates the initial query is weighted.

Collections		TSLM	CSSS	Change
TREC04	MAP	0.352	0.422	+19.9%**
	Recall	6544	7279	+11.2%**
TREC04†	MAP	0.384	0.446	+16.2%**
	Recall	6680	7395	+10.7%**
TREC05	MAP	0.265	0.322	+21.5%**
	Recall	4093	4291	+4.8%**

To see the robustness of the topic signature language model, we change the settings of the translation coefficient. The variance of the mean average precision (MAP) with the translation coefficient λ is shown in Figure 5. When the translation coefficient ranges from 0 to 0.9, the topic signature language model performs always better than the baseline on three

collections. This shows the robustness of the new model. More interestingly, the best performance is achieved at the setting point of λ=0.3 for all four curves; 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 baseline 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.

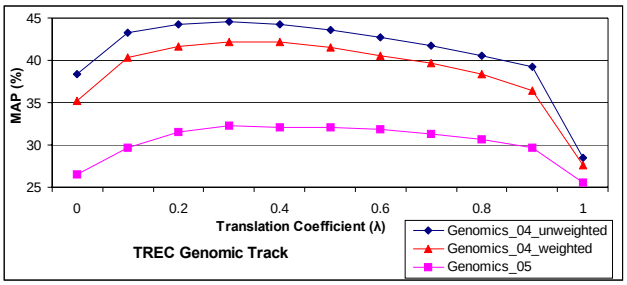


Fig. 5. The variance of MAP with the translation coefficient (λ), which controls the influence of the translation model.

E. Context-Sensitive vs. Context-Insensitive

Basically, the context-insensitive semantic smoothing (CISS) is based on the word-word translation as did in [2], [4], [14] and [15]. The comparison of CISS to CSSS is presented in Table VI. For each collection, we tune the translation coefficient (λ) to maximize the MAP. The optimal λ is about 0.3 for all three collections. Firstly, we can see that CISS significantly outperforms the two-stage language model on all three collections. The gain of the context-sensitive model over the baseline language model is consistent with the conclusions of previous work, such as [2], [4], [14] and [15]. However, CISS is slightly less effective than CSSS, as expected.

Secondly, the improvement of CSSS over CISS seems not much on genomics track. On genomics track 2005, there is almost no improvement. A possible explanation is that most document terms are biological terms such as protein, gene and cell names. The meaning of these terms (e.g. p53, brca1 and orc1) is usually unambiguous even if without additional contextual constraints. Thus, the word-word translations could be very specific and accurate.

TABLE VI

Comparison of the context-sensitive semantic smoothing (CSSS) to the context-insensitive semantic smoothing (CISS) on MAP. The rightmost column is the change of CSSS over CISS. The sign ** and * indicate the improvement is statistically significant according to the paired-sample t-test at the level of $p<0.01$ and $p<0.05$, respectively.

Collections		TSLM	CISS	vs. TSLM	CSSS	vs. CISS
Genomics 2004	MAP	0.352	0.408	+15.9%**	0.422	+3.4%*
	Recall	6544	7176	+9.7%**	7279	+1.4%*
Genomics 2004†	MAP	0.384	0.432	+12.5%**	0.446	+3.2%*
	Recall	6680	7359	+10.2%**	7395	+0.5%
Genomics 2005	MAP	0.265	0.322	+21.5%**	0.322	+0.0%
	Recall	4093	4283	+4.6%**	4291	+0.2%

V. Experiments with Multiword Phrases

A. Testing Collections

In this section, we evaluate the effectiveness of automated multiword phrases as topics signatures. Compared to ontology-based concepts, the extraction of multiword phrases does not need any external human knowledge and could be applied to any public domain. The model is validated on six TREC ad hoc collections from disc1, disc2 and disc3. We select these collections for three reasons. First, these collections are well studied and may published results are available to compare. Second, the content of these collections is all about general news stories on which the Xtract is supposed to work very well on the automated phrase extraction. Third, compared to the vocabulary in genomic collections, the vocabulary of news stories is more ambiguous and thus the context-sensitive semantic smoothing

TABLE VII

The descriptive statistics of ten testing collections

Collections	Word	Phrase	Rel./Doc	Q.Len/Q.#
AP89/1-50	145,349	114,096	3,301/84,678	3.4/47
AP88&89/51-100	204,970	127,736	6,101/164,597	3.4/49
AP88&89/101-150	204,970	127,736	4,822/164,597	4.0/50
WSJ90-92/101-150	135,864	75,687	2,049/74,520	3.8/48
WSJ90-92/151-200	135,864	75,687	2,041/74,520	4.6/49
SJMN91/51-100	173,727	95,986	2,322/90,257	3.4/48

is supposed to take the advantage over the context-insensitive semantic smoothing. The descriptive statistics of these testing collections are shown in Table VII.

B. Document Indexing and Query Processing

We obtain two separate indices, word index and phrase index, for each collection. For word indexing, each document is processed in a standard way. The document is tokenized and stemmed (using porter-stemmer) and stop words are removed. We use a 319-word stop list compiled by van Rijsbergen. Xtract [23] is employed to extract phrases from documents. For phrases appearing in more than ten documents, we estimate their translation probabilities to single-word terms.

The query formulation is fully automated. For each collection, we remove all queries (topics) which contain no relevant documents. Early TREC topics are often described in multiple sections including title, description, narrative, and concept. As many other studies did, we use only the section of title. The extraction of query terms from topic descriptions is the same as the process of word indexing. That is, each topic is tokenized and stemmed and stop words are removed. The average length of queries and total number of queries for each collection is listed in Table VII.

C. Effect of Document Smoothing

We set the parameters γ and μ in the two-stage language model to 0.5 and 750, respectively in the experiment because almost all collections achieve the optimal MAP at this setting point. Interestingly, the Okapi model and the two-stage language model have very similar

TABLE VIII
The comparison of the two-stage language model (TSLM) to the Okapi model.

Collection/Topics	Recall			MAP		
	TSLM	Okapi	Change	TSLM	Okapi	Change
AP89/1-50	1621	1618	-0.2%	0.187	0.187	0.0%
AP88-89/51-100	3428	3346	-2.4%	0.252	0.239	-5.2%
AP88&89/101-150	3055	3087	+1.0%	0.219	0.220	+0.5%
WSJ90-92/101-150	1510	1488	-1.5%	0.239	0.249	+4.2%
WSJ90-92/151-200	1612	1624	+0.7%	0.314	0.304	-3.2%
SJMN91/51-100	1350	1348	-0.1%	0.190	0.184	-3.2%

retrieval performance in the experiment as shown in Table VIII. This is also a kind of indication that both baseline models are well tuned.

TABLE IX

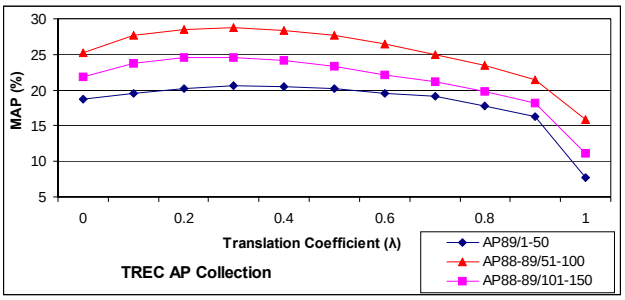
The effect of document expansions based on phrase-word translation. The sign ** and * indicate the improvement is statistically significant according to the paired-sample t-test at the level of $p<0.01$ and $p<0.05$, respectively.

Collection/Topics		TSLM	CSSS	Change
AP89 1-50	MAP	0.187	0.206	+10.2%**
	Recall	1621	1748	+7.8%**
AP88-89 51-100	MAP	0.252	0.288	+14.3%**
	Recall	3428	3771	+100%*
AP88-89 101-150	MAP	0.219	0.246	+12.3%**
	Recall	3055	3445	+12.8%**
WSJ90-92 101-150	MAP	0.239	0.256	+7.1%**
	Recall	1510	1572	+4.1%*
WSJ90-92 151-200	MAP	0.314	0.334	+6.5%**
	Recall	1612	1620	+0.5%
SJMN91 51-100	MAP	0.190	0.208	+9.5%**
	Recall	1350	1472	+9.0%**

The translation coefficient (λ) in the topic signature language model is optimized by maximizing the MAP on the collection of AP89 Topic 1-50. The optimal value is 0.3 and we then apply this learned coefficient to other five collections. Interestingly, all collections achieve the best performance at the setting point of $\lambda=0.3$. We then compare the result of the topic signature language model to the two-stage language model. The comparison is shown in Table IX. In order to validate the significance of the improvement, we also run paired-sample t-test. The incorporation of phrase-word translation improves both MAP and overall recall over the baseline model on all six collections. Except the recall on the collection of WSJ 90-92 Topic 151-200, the improvements over the two-stage language model are all statistically significant at the level of $p<0.05$ or even $p<0.01$. Considering the baseline model is already very strong, we think the topic signature language model is very promising to improve the IR performance.

To see the robustness of the topic signature language model, we also change the settings of the translation coefficient. The variance of MAP with the translation coefficient λ is shown in Figure 6. In a wide range from 0 to 0.6, the topic signature language model always performs better than the baseline on all six collections. This shows the robustness of the model. For all six curves in Figure 6, the best performance is achieved at the setting point of $\lambda=0.3$; after that point, the performance is downward. A possible explanation is that the extracted topic signatures

(multiword phrases) do not capture all points of the document, but the two-stage 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.



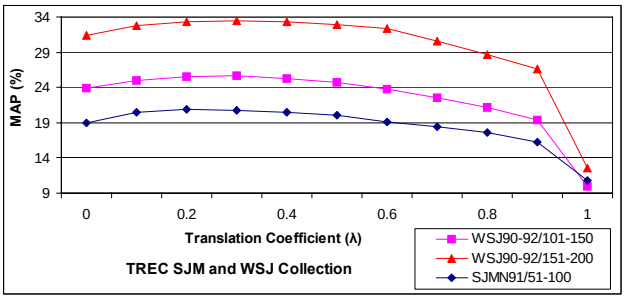


Fig 6. The variance of MAP with λ, which controls the influence of the context-sensitive translation model in the mixture phrase language model

D. Context-Sensitive vs. Context-Insensitive

In newswire collections, many terms are very ambiguous. Terms could have different meanings in different contexts. Thus, the word-word translation may be fairly general and contains mixed topics. The phrase-word translation well solves this problem since most multiword-phrases have very specific meaning and are unambiguous.

The comparison of the context-sensitive semantic smoothing (CSSS) to the context-insensitive semantic smoothing (CISS) is shown in Table X. For each collection, we tune the translation coefficient (λ) to maximize the MAP of CISS. The optimal λ is about 0.1 for all six collections, which is much smaller than the optimal value for CSSS (λ=0.3). It is also a kind of indication that the word-word translation is much noisier than the phrase-word

TABLE X

Comparison of the context-sensitive semantic smoothing (CSSS) to the context-insensitive semantic smoothing (CISS) on MAP. The rightmost column is the change of CSSS over CISS. The sign ** and * indicate the improvement is statistically significant according to the paired-sample t-test at the level of $p<0.01$ and $p<0.05$, respectively.

Collections		TSLM	CISS	vs. TSLM	CSSS	vs. CISS
AP89 1-50	MAP	0.187	0.195	+4.3%*	0.206	+5.6%
	Recall	1621	1730	+6.7%*	1748	+1.0%
AP88-89 51-100	MAP	0.252	0.272	+7.9%	0.288	+5.9%*
	Recall	3428	3735	+9.0%*	3771	+1.0%
AP88-89 101-150	MAP	0.219	0.235	+7.3%**	0.246	+4.7%
	Recall	3055	3237	+6.0%*	3445	+6.4%*
WSJ90-92 101-150	MAP	0.239	0.244	+2.1%	0.256	+4.9%*
	Recall	1510	1568	+3.8%**	1572	+0.3%
WSJ90-92 151-200	MAP	0.314	0.324	+3.2%	0.334	+3.1%
	Recall	1612	1646	+2.1%*	1620	-1.6%
SJMN91 51-100	MAP	0.190	0.199	+4.7%*	0.208	+4.5%
	Recall	1350	1427	+5.7%**	1472	+3.2%

translation. From the experimental results, we can firstly see that CISS greatly outperform the two-stage language model and most of the improvements are statistically significant. Secondly, the CSSS has considerable gain over the CISS especially on the measure of MAP.

In addition, the CSSS is computationally more efficient than the CISS. The CSSS is based on phrase-word translations while the CISS based on word-word translations. As shown in Table II, an average document in the testing collections about 180 unique works, but only about 30 unique multiword phrases. In other words, the CSSS is six times faster than the CISS for the construction of occurrence data as well as the document model expansions (smoothing).

E. Vs. Other Types of Phrases

The different types of phrases may have different impact on the retrieval performance. Fagan reported significant improvement on some collections using "statistical" phrases, but none with "syntactic" phrases in his thesis [9]. In this paper, we used kind of phrases with both "syntactic" and "statistical" constraints extracted by Xtract and got very positive results. An interesting question is then raised up:

"*Can other types of phrases (e.g. WordNet phrases and Named Entities) still get positive results with the topic signature language model?*"

To test this idea, we add WordNet noun phrases and named entities including person, organization, and location to the document index and see if the IR performance is further improved or even decreased. WordNet noun phrases are manually selected phrases. The named entities are automatically extracted by GATE purely according to syntactic rules. Thus, neither of them is constrained by statistical criteria. Take the example of AP89 collection. Before adding extra phrases, the collection has 114,096 phrases. After adding WordNet noun phrases and named entities, the number of phrases is increased by about 50K. However, the increase of

phrase coverage does not make any improvement of IR performance. The other five collections are in the similar case. Examining the extra noun phrases carefully, we find out that most of those phrases are infrequent in the testing collection. Actually, the majority of phrases frequently occurring in the collection are already extracted by Xtract. Those infrequent phrases will have little effect on the document model expansions, thus have no effect on the IR performance. Therefore, in order to make the topic signature (phrase) language model effective, we should use phrases, frequently occurring in the collection or constrained by "statistical" criteria.

VI. Conclusion

In this paper, we proposed a topic signature language model for ad-hoc text retrieval. This new model decomposed a document into a set of weighted context-sensitive topic signatures and then translated those topic signatures into individual query terms. Because the topic signature itself contained contextual information, the document model expansion based on topic signatures would be more accurate, compared to the model expansion based on context-insensitive term translations proposed in previous work [2] [4] [14], and thus improved the retrieval performance.

We implemented two types of topic signatures in this paper. When domain-specific ontology is available, ontology-based concepts can be used as topic signatures. Otherwise, automated multiword phrases are an alternative. We evaluated the effectiveness of ontology-based concepts on TREC Genomics Track 2004 and 2005 and the effectiveness of multiword phrases on TREC Ad hoc Track Disc1&Disc2&Disc3. The topic signature language model significantly outperformed the two-stage language model on all collections. We also implemented a context-insensitive version of semantic smoothing. It has the same framework as the topic signature

language model, but the document model expansion (smoothing) is based on context-insensitive word-word translations rather than context-sensitive signature-word translations. As expected, it is less effective than the context-sensitive semantic smoothing, though it does achieve significant improvement over the two-stage language model.

The topic signature language is the linear interpolation of the two-stage language model and the topic signature based translation model. It is required to set the translation efficient which controls the influence of the translation component in the mixture model. It is somewhat ad-hoc nature. Fortunately, the experiments showed the robustness of the model. When the translation coefficient took different values in a wide range (0-0.9 for ontology-based concepts and 0-0.6 for multiword phrases), the topic signature language model always performed better than the baseline. More interestingly, all collections achieved the best MAP at the same setting (i.e. λ=0.3). This means it is reasonable to train the optimal translation coefficient on one collection and then apply the learned coefficient to other collections in future.

We also found out two factors would affect the effectiveness of the topic signature language model. One is the degree of the ambiguity of terms in the collection. If terms (e.g. in newswire collections) are very ambiguous, the topic signature model (i.e. context-sensitive semantic smoothing) can take much advantage over the context-insensitive semantic smoothing. The other is occurrence frequency of the topic signatures in the collection. If the topic signatures infrequently occur in the collection, the model has little effect on improving the IR performance.

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 terms. 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 approach to document expansions based on topic signature translations. Third, we empirically proved the superiority of the context-sensitive semantic smoothing over context-insensitive semantic smoothing as well as simple background smoothing.

Probabilistic topical models such as pLSI [13] and LDA [25] also take the context into account and thus can handle the word polysemy problem. In this paper, we analyzed their computing complexity in the setting of IR and concluded that these two models were computationally less efficient than the topic signature language model in both offline topic model estimation stage and online document model smoothing stage. However, the comparison of the effectiveness of three models on retrieval tasks is still unclear. It should be interesting to have a comprehensively comparative study on these three models with respect to their efficiency and effectiveness for ad-hoc text retrieval.

Besides, how to optimize the mixture weights of the topic signature language model remains an opening issue. In this paper, we empirically tuned a fixed translation coefficient on training data set and achieved good results. Ideally, the translation coefficient should be conditioned on each document because the relative information provided by the translation model based topic signatures varied with different documents. In addition, the topic signature language model can also be applied to applications other than information retrieval. Traditional text mining problems such as text clustering and text classification are also based on document models. Thus, it is natural to extend the application of the new model to those areas. Our previous work [33] has successfully applied this model to agglomerative document clustering. In future, we will further evaluate its effectiveness in related areas.

REFERENCES

[1] Bai, J., Nie, J.Y., and Cao, G., "Context-Dependent Term Relations for Information Retrieval," *In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, July 2006, Sydney, Australia

[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] Blei, D., Ng, A. and Jordan, M., "Latent Dirichlet allocation," Journal of machine Learning Research, 3, 2003, pp 993-1022.

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

[5] Croft, W.B., Turtle, H.R., and Lewis, D.D., "The use of phrases and structured queries in information retrieval," *In Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1991, pp. 32—45

[6] Cunningham, H., "GATE, A General Architecture for Text Engineering," Computers and the Humanities, 2002, Vol. 36, pp. 223-254

[7] Deerwester, S., Dumais, T.S., Furnas, W.G., Landauer, K.T., and Harshman, R., "Indexing by Latent Semantic Analysis," *Journal of the American Society of Information Science*, 1990, 41(6): 391- 407

[8] 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.

[9] Fagan, J., "Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods," Ph.D. Thesis, Technical Report 87-868, Cornell University, Computer Science Department, 1987.

[10] 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

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

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

[13] Hoffman, T., "Probabilistic latent semantic indexing," *1999 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999, pp. 50-57

[14] 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

[15] 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.

[16] 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.

[17] 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.

[18] Miller, G. A., "WordNet: a lexical database for English," *Communications of the ACM*, 1995, 38(11), pp. 39-41

[19] 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.

[20] Pickens, J. and Croft, W.B., "An exploratory analysis of phrases in text retrieval". In *RIAO2000 Conference Proceedings*, pp. 1179-1195, Paris, France.

[21] 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.

[22] Robertson, S.E. et al. "Okapi at TREC-4", *In the Fourth Text Retrieval Conference*, 1993.

[23] Smadja, F., "Retrieving collocations from text: Xtract," *Computational Linguistics*, 1993, 19(1), pp. 143--177.

[24] 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.

[25] Wei, X. and Croft, W.B., "LDA-based document models for ad-hoc retrieval," *In Proceedings of the 29th ACM SIGIR Conference on Research and Development in IR*, pp. 178-185

[26] 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.

[27] 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.

[28] 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.

[29] 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

[30] 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.

[31] Zhou, X., Zhang, X., and Hu, X., "MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup," *In the 9th biennial The Pacific Rim International Conference on Artificial Intelligence (PRICAI 2006)*, Aug 9-11, 2006, Guilin, Guangxi, China, pp. 1145-1149

[32] Zhou X., Hu X., Zhang X., Lin X., Song I-Y, "Context-Sensitive Semantic Smoothing for the Language Modeling Approach to Genomic IR," *In the 29th Annual International ACM SIGIR Conference (SIGIR 2006)*, Aug 6-11, 2006, Seattle, WA, USA, pp. 70-77

[33] Zhou, X., Zhang, X., and Hu, X., "Semantic Smoothing of Document Models for Agglomerative Clustering," in the *Twentieth International Joint Conference on Artificial Intelligence(IJCAI 07)*, Hyderabad, India, Jan 6-12, 2007, pp. 2928-2933

[34] Zhou, X., Zhang, X., and Hu, X., "The Dragon Toolkit Developer Guide," Data Mining & Bioinformatics Lab, Drexel University, http://www.ischool.drexel.edu/dmbio/dragontool/tutorial.pdf

[35] UMLS, http://www.nlm.nih.gov/research/umls/

[36] GENIA Corpus, http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/