

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. For example, “mouse” may be translated into both “computer” and “cat” with high probabilities. 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 EM; 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.

Background: Why Semantic Smoothing for IR?

A Searching Scenario

Suppose you are typing a keyword “car” into Google search box for retrieving web pages related to cars. Should Google return to you the pages containing word “auto” but not “car”? If yes, how can Google make it?

A Heuristic Approach: Query Expansion

Query Expansion is a technique that automatically expands a user’s original query with synonyms or related keywords in back-end for the purpose of improving IR performance, especially the IR recall.

The key to query expansion is how to find synonyms or related keywords automatically. The augment of inappropriate keywords may make the performance even downward.

A Formal Approach: Statistical Translation

Query expansion can only exactly augment or not augment a keyword into the query. However, the real distance of any two keywords are fuzzy, not just binary, zero or one. Instead, the statistical translation language model, first proposed by Berger and Lafferty[1], can capture such kind of fuzzy distance. It statistically maps any terms in a document into query terms and the resulting summated score is used to indicate the relevance of the document to the query as described in the formula below:

$$p(q|d) = \sum_j \sum_w t(q_j|w)l(w|d)$$

Where:

$p(q|d)$ is the relevance score of document d to the query q .

$t(q_j|w)$ is the probability of translating word w into query term q_j .

$l(w|d)$ is the probability of document d generating word w .

The key issue of the translation model is to estimate the translation probability from training data set. Almost all automatic estimates are based on word occurrence.

Problems of Semantic Smoothing

Query: Keyboard

The document about the animal of mouse could be returned for the above query according to the left side translation model.

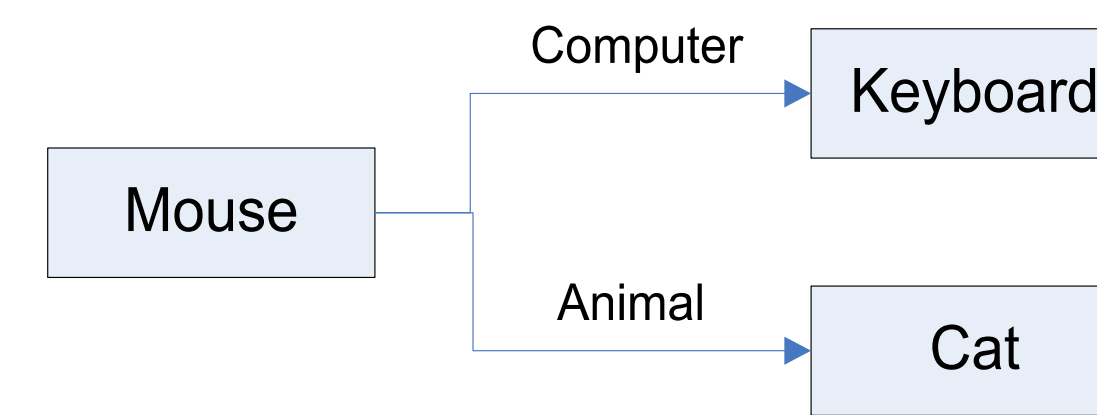


Figure 1. Without any contextual constraints, the term “mouse” may be translated to both “keyboard” and “cat” with high probabilities. Therefore, the resulting translation will be mixed and fairly general. Consequently, the IR performance will be compromised.

Solution: Context-sensitive Semantic Smoothing

Solution Highlight:

- ◆ How to Define Context?
- ◆ How to Extract Contextual Information?
- ◆ How to Estimate Context-sensitive Translation Probability?
- ◆ How to Incorporate Context-sensitive Semantic Smoothing?

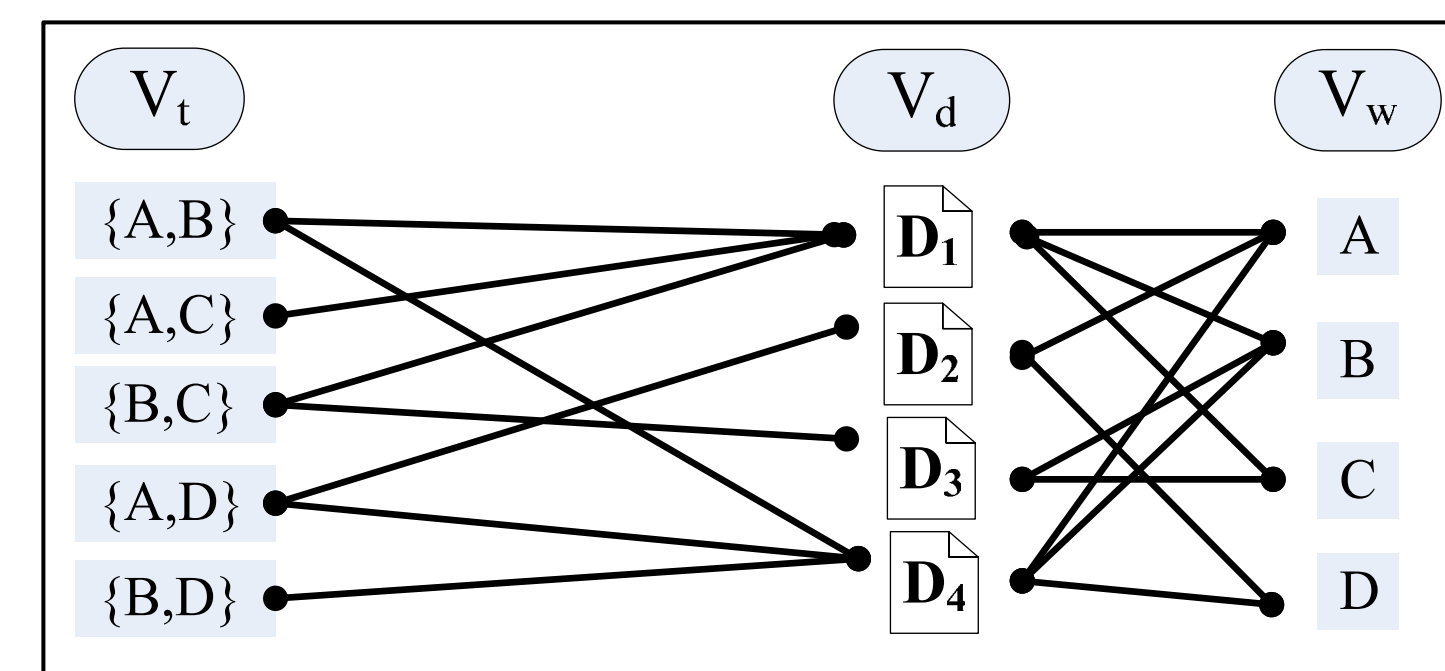
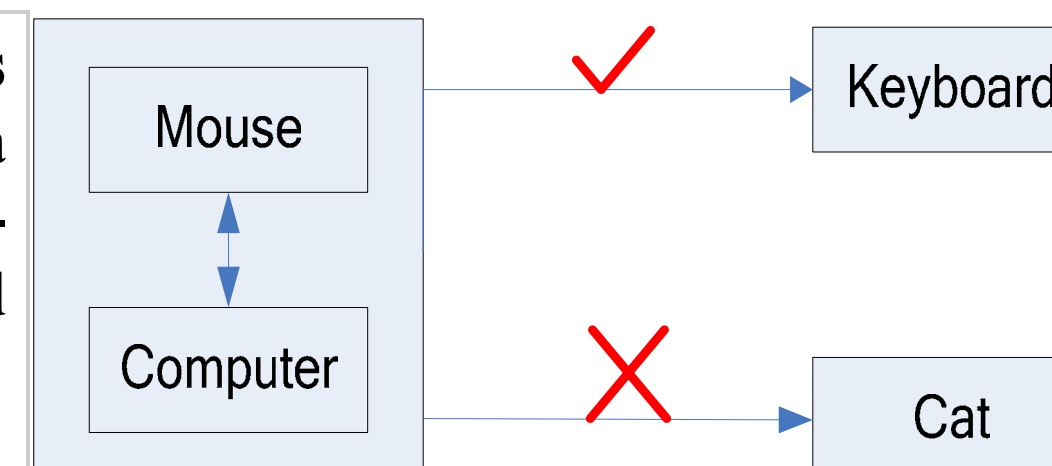


Figure 2. Illustration of document indexing. V_t , V_d , and V_w are topic signature set, document set and concept set, respectively. Topic signatures will be translated into individual concepts statistically.

Context Definition

A pair of two concepts, referred to as *topic signature*, serves as the context of a concept. The term “Mouse” in conjunction with “Computer” will be translated into “Keyboard”, but not “Cat”.



Concept and Topic Signature Extraction

MaxMatcher [4] is used to extraction concept from texts.

A topic signature is defined as a pair of two concepts if they:

- ◆ Both of them are major concepts.
- ◆ They appear in the same clause of an English sentence
- ◆ Their semantic types are compatible according to domain ontology

Context Sensitive Translation Probability Estimates

The probability of translating a topic signature t_k to a concept w can be estimated by the Expectation Maximum (EM) algorithm [3] with the following update formulas:

$$\hat{p}^{(m)}(w) = \frac{(1-\alpha)p^{(m)}(w|\theta_k)}{(1-\alpha)p^{(m)}(w|\theta_k) + \alpha p(w|C)}$$

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

Where: D_k is the set of documents containing topic signature t_k ; $c(w, D_k)$ is the frequency count of concept w in D_k ; α is the background noise coefficient; C denotes the background model.

Document Model and Query Model Smoothing

Document Model Smoothing

$$p_i(w|d) = \sum_k p(w|t_k) \frac{c(t_k, d)}{\sum_i c(t_i, d)}$$

Where: $c(t_k, d)$ is the frequency count of topic signature t_k in document d .

Query Model Smoothing (Pseudo Relevance Feedback)

$$p_f(w|q) = \sum_k p_{ks}(w)p(t_k|\theta_F)$$

Where:

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

The signature feedback model $p(t_k|\theta_F)$ could be estimated using the EM algorithm [3] using feedback documents (top-ranked documents):

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

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

Where: F is the feedback document set; $c(t_k, F)$ is the frequency count of topic signature t_k in F ; α is the background noise coefficient; C denotes the background model.

Evaluation

Evaluation Highlight:

- ◆ Testing Collections? TREC Genomic Track 2004/05
- ◆ Evaluation Measure? Average Precision and Recall
- ◆ Evaluation Logic?

Context-sensitive Model vs. Baseline Language Model
Context-sensitive Model vs. Context-insensitive Model
Document Smoothing vs. Query Smoothing vs. Both

Table 1. The comparison of the baseline language model to document smoothing model and query smoothing model. The number of relevant documents for TREC04 and TREC05 are 8266 and 4585, respectively. The asterisk (*) indicates the initial query is weighted. MAP means mean average precision.

Collection	Base	Document Smooth		Query Smooth		
		Abs.	Change	Abs.	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%

Table 2. The interaction effect of document and query smoothing. “Max” is the maximum effect achieved by document smoothing or query smoothing. “Both” is the result of using both smoothing techniques. “Change^[1]” is the improvement of “Both” over “Base”. “Change^[2]” is the improvement of “Both” over “Max”

Collection		Base	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%

Table 3. Comparison of the context-sensitive semantic document smoothing to the context-insensitive semantic document smoothing on MAP. The rightmost column is the change of Sensitivity Model over Insensitivity Model.

Collection	Base	Context Insensitivity		Context Sensitivity		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%

Conclusions

Findings From the Experiment:

- ◆ The effect of the document smoothing and query smoothing as well as their interaction effect, based on context-sensitive translation, are all positive, in comparison with baseline language models.
- ◆ The effect of the context-sensitive document smoothing is superior to that of the context-insensitive.

Contributions of This Paper:

- ◆ Propose a new document representation using a set of weighted concepts and topic signatures.
- ◆ Expand document and query language models through context-sensitive semantic smoothing.
- ◆ Empirically prove the effectiveness of context-sensitive semantic smoothing for language modeling IR.

References

* The paper based on this research project has been accepted by SIGIR 2006.

- [1] Berger, A. and Lafferty J., “Information Retrieval as Statistical Translation”, SIGIR’99, 222-229.
- [2] Lafferty, J. and Zhai, C., “Document Language Models, Query Models, and Risk Minimization for Information Retrieval”, SIGIR’01, 111-119
- [3] Zhai, C. and Lafferty, J., “Model-based Feedback in the Language Modeling Approach to Information Retrieval”, CIKM 2001, 403-410
- [4] Zhou, X., Hu, X. and Zhang, X., “Using Concept-based Indexing to Improve Language Modeling Approach to Genomic IR”, ECIR 2006, 444-455