SIGIR 2008, Singapore

# A Generation Model To Unify Topic Relevance and Lexicon-based Sentiment For Opinion Retrieval

Min Zhang, Xingyao Ye
Tsinghua University, Beijing, China
z-m@tsinghua.edu.cn
July 22, 2008

IR@Tsinghua

# Introduction

- A growing interest in finding out people's opinions from web data
  - Product survey
  - Advertisement analysis …
  - Political opinion polls
  - TREC started a special track on blog data in 2006 – blog opinion retrieval
    - It has been the track that has the most participants in 2007

# Related Work

- The popular opinion identification approaches
  - Text classification [Tong 01, Liu 05, Zhang 06, etc]
  - Lexicon-based sentiment analysis [Liao 06, Mishne 06, Yang 06, Oard 06, Macdonald 07]
  - Opinion retrieval
- Opinion retrieval: To find the **sentimental relevant** documents according to a user's query
- Topicality and polarity are first fused together to form the notion of opinion retrieval by Hurst and Nigam [Hurst 04]
  - Emphasize on how to judge the existence of opinions
- First generation model on opinion ranking using the cross entropy of topics and sentiments [Eguchi 06]

$$\alpha \sum_v R_t(v) \log P_t(v) + (1-\alpha) \sum_v R_s(v) \log P_s(v)$$

# Related Work (cont.)

- One of the key problems: How to combine opinion score with relevance score of each document for ranking
- Ad hoc solutions of combining relevance ranking and opinion detection result
  - 2 steps: rank with relevance , then re-rank with sentiment score
  - Generally linear combination by experience
  - TREC blog 06 observation [Ounis 06]
    - Existing methods to sentimental document ranking provide no improvements over mere topic-relevance ranking
  - TREC blog 07
    - Better result, but still an interesting observation that the topic-relevance result outperforms most opinion-based approaches

# Generation Model For Opinion Retrieval

Information Retriever @ TsingHua University

- **The Generation Model**
  - **To find both sentimental and relevant documents with ranks**
- Topic Relevance Ranking
- Opinion Generation Model and Ranking
- Ranking function of generation model for opinion retrieval

# The Proposed Generation Model

- In existing probabilistic-based IR models, two ways to factor the “relevance” probability [Lafferty 03]
  - ***query generation*** and ***document generation***
- Document generation model: **how well the document *d* “fits” the particular query *q*,** estimate posterior probability $p(d \mid q)$

$$p(d \mid q) \propto p(q \mid d)\, p(d)$$

- When assuming a uniform document prior, the ranking function is reduced to the likelihood of generating the expected query terms from the document.

# The Proposed Generation Model (Cont.)

- In opinion retrieval, p($d$ | $q$, $\boldsymbol{s}$)
- In this work, discuss lexicon-based sentiment analysis
  - Assume
    - The latent variable $s$ is estimated with a pre-constructed bag-of-word sentiment thesaurus
    - All sentiment words $s_i$ are uniformly distributed.
  - Then $p(d|q,s) = \sum_i p(d|q,s_i)p(s_i,s)$

$$= \frac{1}{|S|}\sum_i p(d|q,s_i)$$ |$s$|: # of words in sentiment thesaurus

$$\propto \frac{1}{|S|}\sum_i p(q,s_i|d)p(d)$$

$$= \frac{1}{|S|}\sum_i p(s_i|d,q)p(q|d)p(d)$$

**topic relevance**

**given query $q$, how probably a document $d$ *generates* a sentiment word $s_i$**

# The Proposed Generation Model (Cont.)

- The final generation model

$$p(d|q,s) = I_{Op}(d,q,s) I_{rel}(d,q)$$

$I_{Op}(d,q,s) \stackrel{\text{def}}{=} \frac{1}{|S|} \sum_i p(s_i|d,q)$, ***opinion generation model*** to sentiment analysis

$I_{rel}(d,q) \stackrel{\text{def}}{=} p(q|d)p(d)$ ***document generation model*** to estimate topic relevance

  - Essentially it presents a quadratic relationship between document sentiment and topic relevance

- v.s. In previous work, linear combination

$$\mathrm{Rel}\_Op\_Score = (1-\lambda) Senti\_Score + \lambda \mathrm{Rel}\_Score$$

# Generation Model For Opinion Retrieval

- The Generation Model
  - To find both sentimental and relevant documents
- **Topic Relevance Ranking**
- Opinion Generation Model and Ranking
- Ranking function of generation model for opinion retrieval

# Topic Relevance Ranking Model $I_{rel}(d,q)$

Information Retriever @ Tsinghua University

- The Binary Independent Retrieval (BIR) model is one of the most famous ones in this branch
  - Heuristic ranking function BM25

$$ScoreI_{rel}(d,q) := \sum_{w \epsilon q \cap d} \left( ln \frac{N - df(w) + 0.5}{df(w) + 0.5} \times \frac{(k_1+1) \times c(w,d)}{k_1((1-b) + b\frac{|d|}{avdl} + c(w,d)} \times \frac{(k_3+1) \times c(w,q)}{k_3 + c(w,q)} \right)$$

  - $c(w,d)$ is the count of word $w$ in the document $d$,
    $c(w,q)$ is the count of word $w$ in the document $q$,
    $N$ is the total number of documents in the collection,
    $df(w)$ is the number of documents that contain word $w$
    $|d|$ is the length of document $d$,
    $avdl$ is the average document length,
    $k_1$(1.0 to 2.0),$b$ (usually 0.75) and $k_3$ (0 to 1000) are constants.

# Generation Model For Opinion Retrieval

Information Retriever @ TsingHua University

- The Generation Model
  - To find both sentimental and relevant documents
- Topic Relevance Ranking
- **Opinion Generation Model and Ranking**
- Ranking function of generation model for opinion retrieval

# Opinion Generation Model $I_{op}(d,q,s)$

- $I_{op}(d,q,s)$ focus on the problem that given query $q$, how probably a document $d$ generates a sentiment expression $s$.

$$I_{Op}(d,q,s) \stackrel{\text{def}}{=} \frac{1}{|S|}\sum_i p(s_i|d,q) \propto \sum_i p(s_i|d,q)$$

- This opinion generation model is on the branch of query generation
- Different from general query generation model
  - |S| is quite large (~ thousands)

    (v.s. in general models, # of terms in the query is usually small )
  - Sparseness problem → smoothing

# Opinion Generation Model
## – Parameter Estimation (smoothing)

- $$p(s_i|d,q) = \begin{cases} p_{seen}(s_i|d,q) \\ p_{unseen}(s_i|d,q) \end{cases} = \begin{cases} p_s(s_i|d,q) & if\ s_i\ is\ seen \\ \alpha_d p(s_i|C,q) & otherwise \end{cases}$$
- By Zhai & Lafferty's study, **Jelinek-Mercer smoothing** is more effective when the "queries" are long and more verbose.
  - In this proposed opinion generation model, the "queries" are sentiment words

$$p_s(s_i|d,q) = (1-\lambda)p_{ml}(s_i|d,q) + \lambda p(s_i|C,q), \quad \alpha_d = \lambda$$

$p_{ml}(s_i|d,q)$: the maximum likelihood estimation of $p(s_i|d,q)$

# Opinion Generation Model
## – Parameter Estimation (smoothing)

- Recall $I_{op}(d,q,s) \stackrel{\text{def}}{=} \frac{1}{|S|}\sum_i p(s_i|d,q) \propto \sum_i p(s_i|d,q)$
- $\sum_i p(s_i|d,q)$

$$= \sum_{s_i \in d} p(s_i|d,q) + \sum_{s_i \notin d} p(s_i|d,q)$$

$$= \sum_{s_i \in d} p_S(s_i|d,q) + \sum_{s_i \notin d} \alpha_d p(s_i|C,q)$$

$$= \sum_{s_i \in d}[(1-\lambda)p_{ml}(s_i|d,q) + \lambda p(s_i|C,q)] + \sum_{s_i \notin d} \lambda p(s_i|C,q)$$

$$= \sum_{s_i \in d}(1-\lambda)p_{ml}(s_i|d,q) + \underline{\lambda \sum_i p(s_i|C,q)}$$

$$= \sum_{s_i \in d}(1-\lambda)p_{ml}(s_i|d,q) + \lambda$$

- We use the co-occurrence of $s_i$ and $q$ inside $d$ within a window $W$ as the ranking measure of $p_{ml}(s_i|d,q)$, then

$$ScoreI_{op}(d,q,s) := \sum_{s_i \in d}(1-\lambda)\frac{co(s_i,q|W)}{c(q,d)|W|} + \lambda$$

$co(s_i,q|w)$: the frequency of $s_i$ which is co-occurred with $q$ within $W$

$c(q,d)$: the query term frequency in the document, |W|: window size

# Generation Model For Opinion Retrieval

- The Generation Model
  - To find both sentimental and relevant documents
- Topic Relevance Ranking
- Opinion Generation Model and Ranking
- **Ranking function of generation model for opinion retrieval**

# Ranking function of generation model for opinion retrieval

Information Retriever @ TsingHua University

- **The final ranking function of the proposed generation model**

$$p(d \mid q, s) \overset{rank}{=} ScoreI_{op}(d, q, s) \times ScoreI_{rel}(d, q)$$

$$= (\sum_{S_i \in d} (1-\lambda) \frac{co(s_i, q \mid W)}{c(q, d) \cdot |W|} + \lambda) \times ScoreI_{rel}(d, q)$$

$$p(d \mid q, s) \overset{rank}{=} ScoreI_{rel}(d, q) \qquad \lambda = 1$$

- **To reduce the impact of unbalance between #(sentiment words) and #(query terms) → logarithm normalization**

$$p(d \mid q, s) \overset{rank}{=} [(1-\lambda) \log(\sum_{S_i \in d} \frac{co(s_i, q \mid W)}{c(q, d) \cdot |W|} + 1) + \lambda] \times ScoreI_{rel}(d, q)$$

# Experimental Setup – Data Set

- Data set
  - TREC blog 06 and TREC blog 07data
    - Permalinks, homepages and feeds
  - 100,649 blogs during 2.5 months
  - Only use permalinks in this work
  - 50 + 50 topics
  - Short queries (only <title> field)
- Strategy: find top 1000 relevant documents, then re-rank the list with proposed model

# Experimental Setup – Models

- General linear combination (Shown as ***Linear Comb.***)

$$\mathrm{Rel\_Op\_Score} = (1-\lambda)\,Senti\_Score + \lambda\,\mathrm{Rel\_Score}$$

- Our proposed generation model with Jelinek-Mercer smoothing (Shown as ***generation model***)

$$p(d\mid q,s) \overset{rank}{=} ScoreI_{op}(d,q,s)\times ScoreI_{rel}(d,q)$$
$$= \Big(\sum_{S_i\in d}(1-\lambda)\frac{co(s_i,q\mid W)}{c(q,d)\cdot\mid W\mid}+\lambda\Big)\times ScoreI_{rel}(d,q)$$

- Our proposed generation model with Jelinek-Mercer smoothing and logarithm normalization (Shown as ***Generation, log***)

$$p(d\mid q,s) \overset{rank}{=} [(1-\lambda)\log\Big(\sum_{S_i\in d}\frac{co(s_i,q\mid W)}{c(q,d)\cdot\mid W\mid}+1\Big)+\lambda]\times ScoreI_{rel}(d,q)$$

Information Retriever @ Tsinghua University

# Experimental Setup – Sentimental Lexicons

| | Thesaurus Name | Size | Description |
|---|---|---|---|
| 1 | HowNet | 4621 | English translation of pos/neg Chinese words from HowNet |
| 2 | WordNet | 7426 | Selected words from WordNet with seeds |
| 3 | Intersection | 1413 | Words appeared in both 1 and 2 |
| 4 | Union | 10634 | Words appeared in either 1or 2 |
| 5 | General Inquirer | 3642 | All words in the positive and negative categories |
| 6 | SentiWordNet | 3133 | Words with a positive or negative score above 0.6 |

# Experimental Results And Discussion

- Effectiveness of Sentimental Lexicons
- Selection of Window Size
- Opinion Retrieval Model Comparison
- Per-topic Analysis
- Case Study

# 1. Effectiveness of Sentimental Lexicons

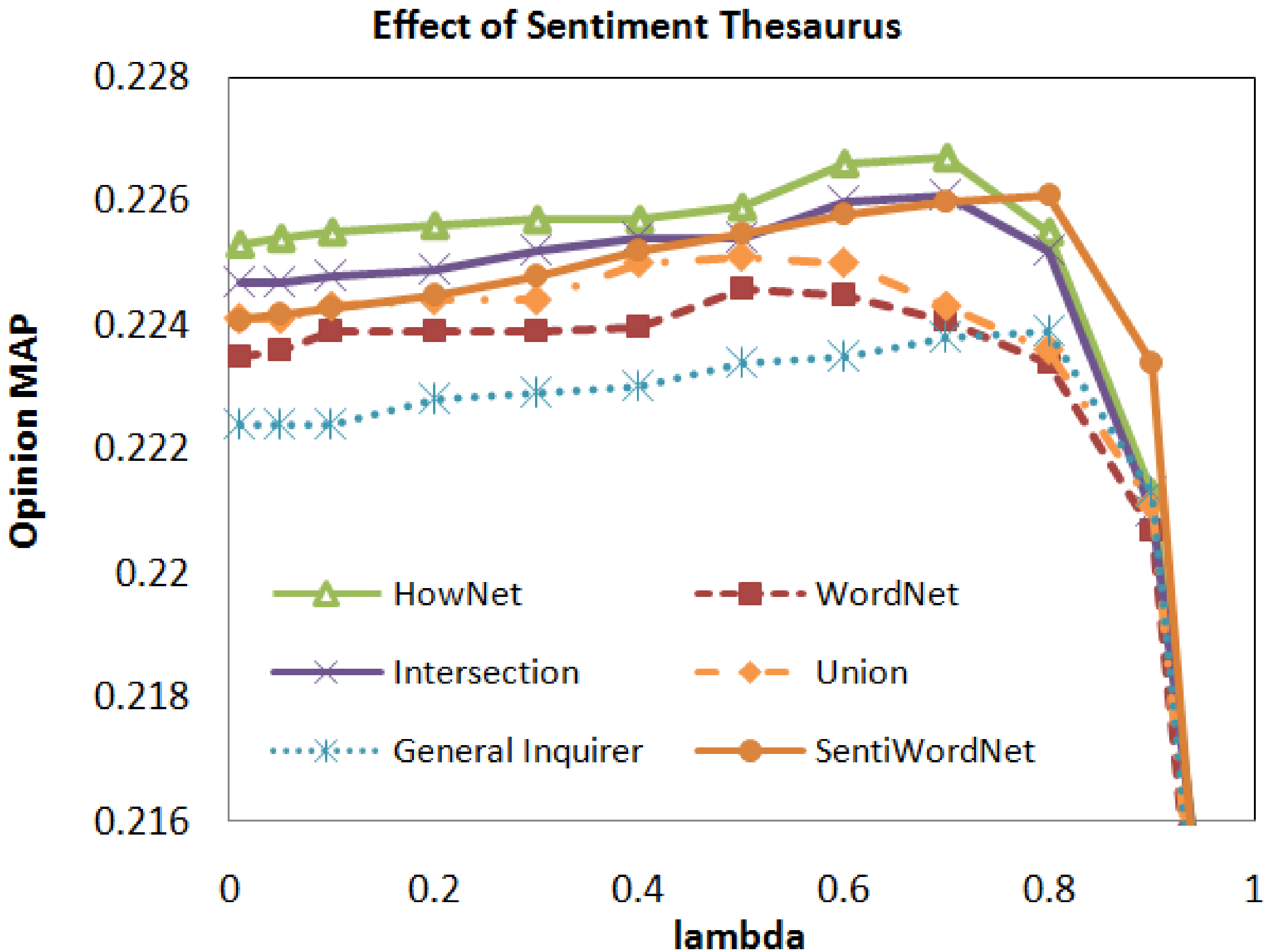
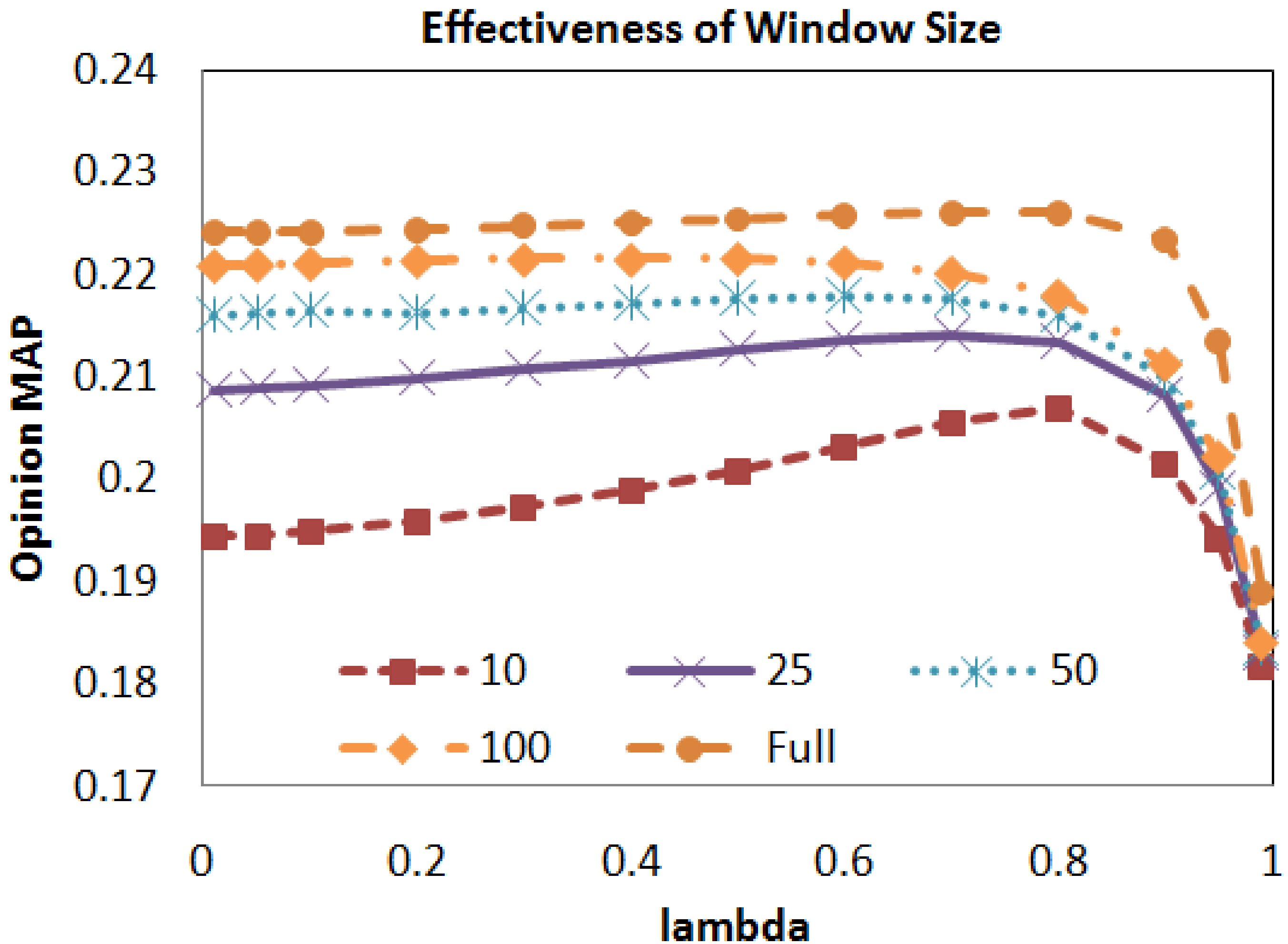



Figure 3 MAP -λ curve for different sentiment thesaurus. (Blog 06 Data)

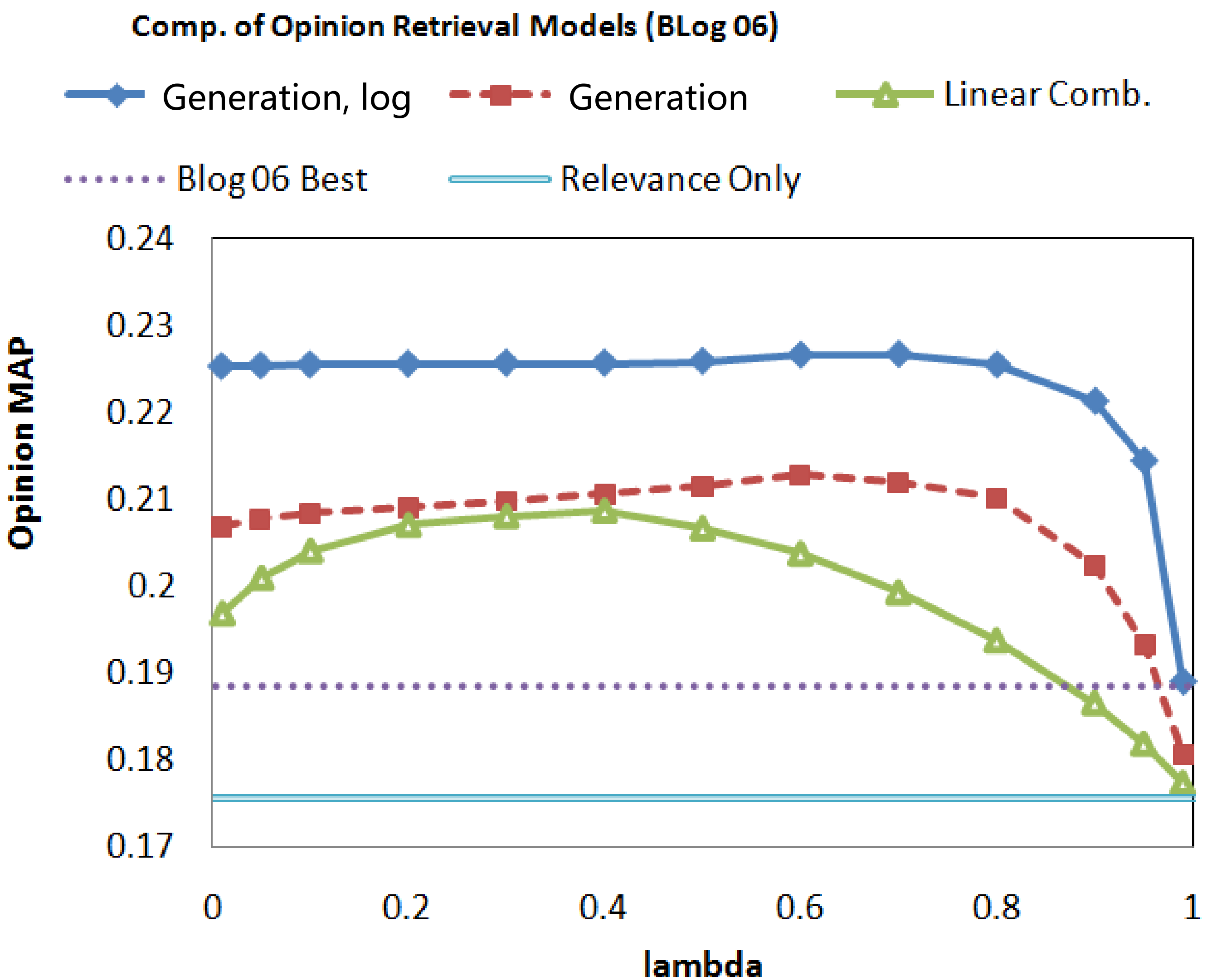
All the following experiments use sentiWordNet

# 2. Selection of Window Size




- Figure 4. MAP v.s. window size under different λ (Blog06)

# 3. Opinion Retrieval Model Comparison




- Figure 5. MAP-λ curve for different opinion ranking formulas

# 3. Opinion Retrieval Model Comparison (Cont.)

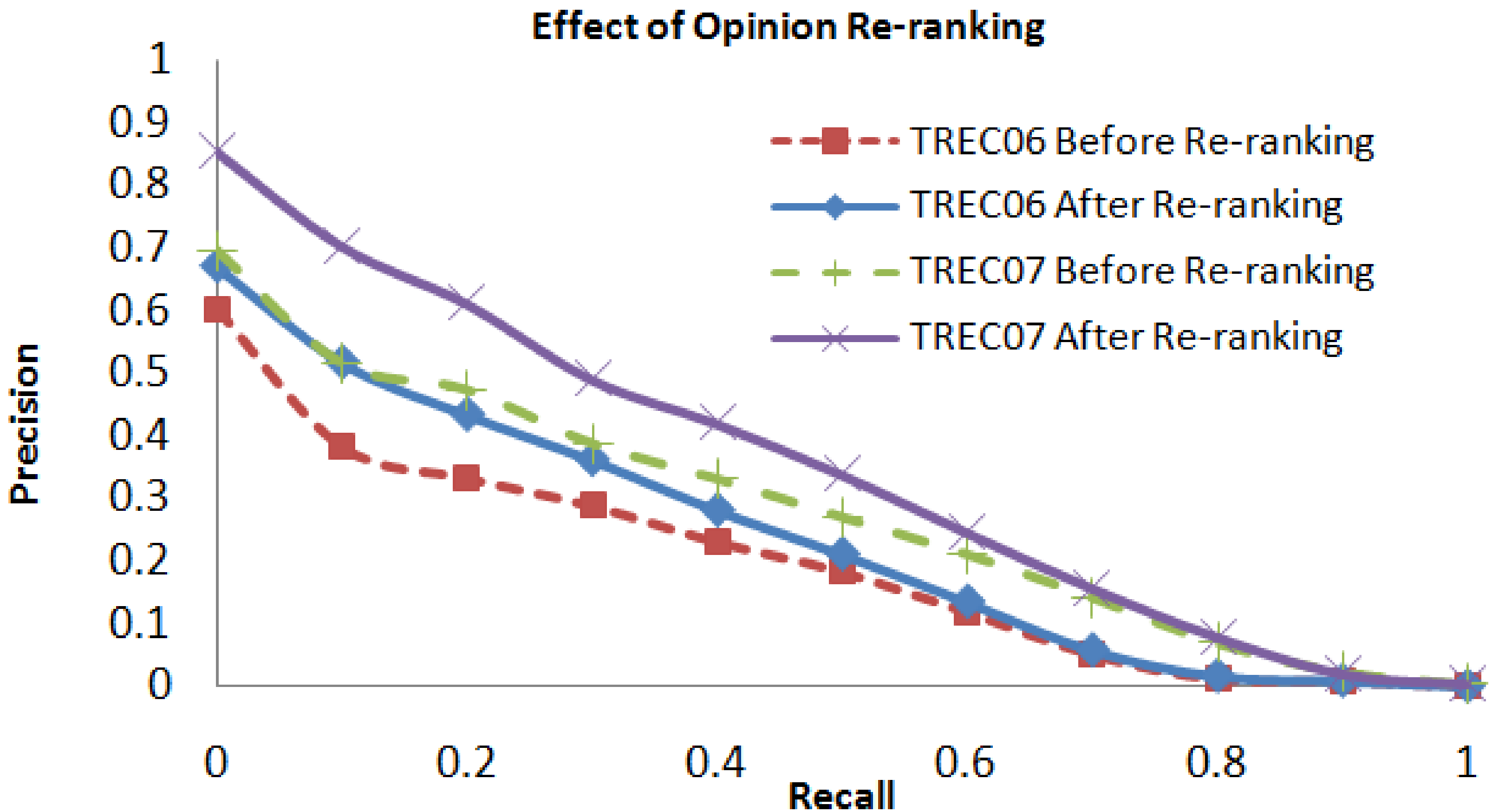
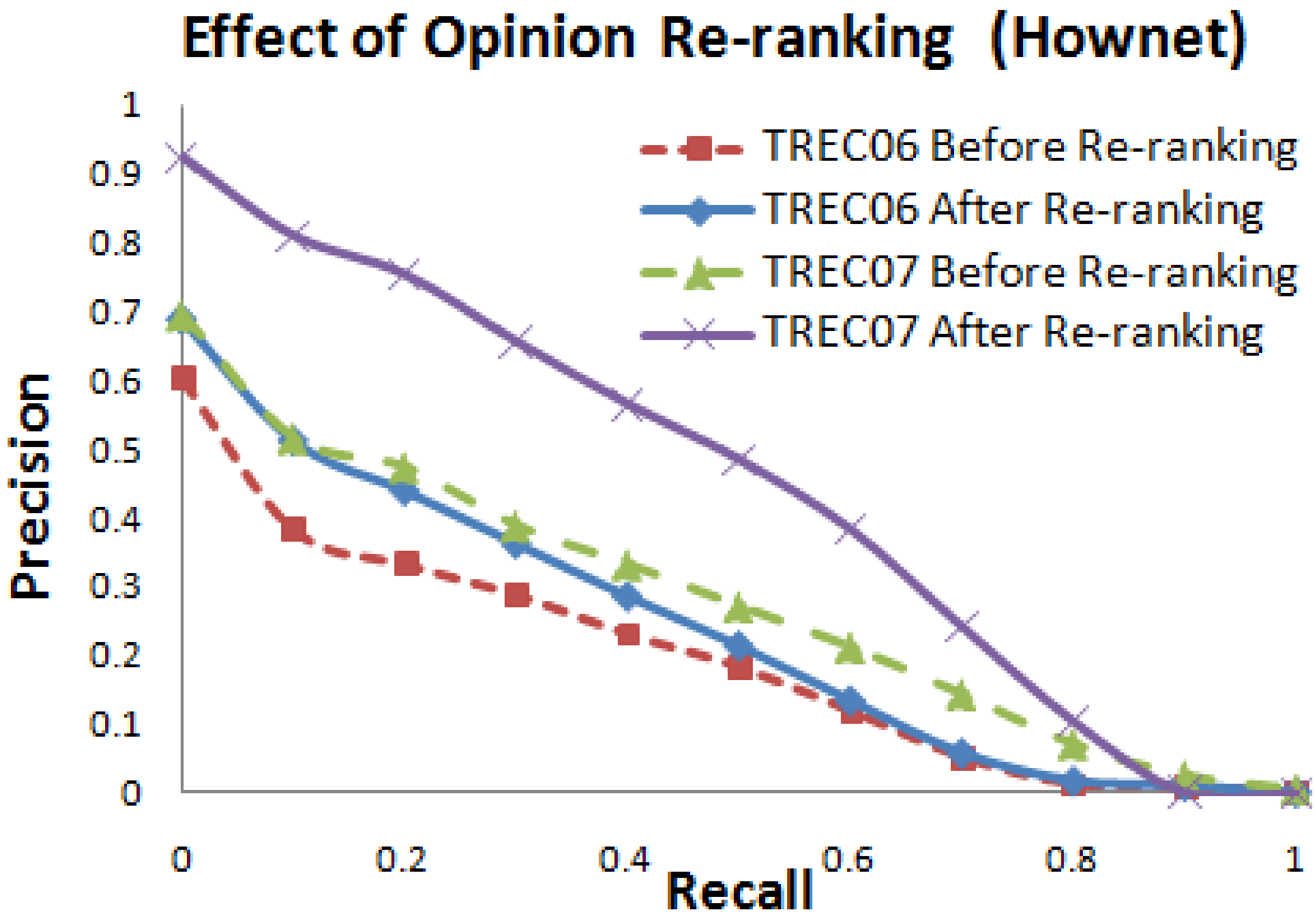



Figure 6. Precision-recall curves before and after opinion re-ranking of top 1000 relevant documents

# 3. Opinion Retrieval Model Comparison (Cont.)

Information Retriever @ TsingHua University

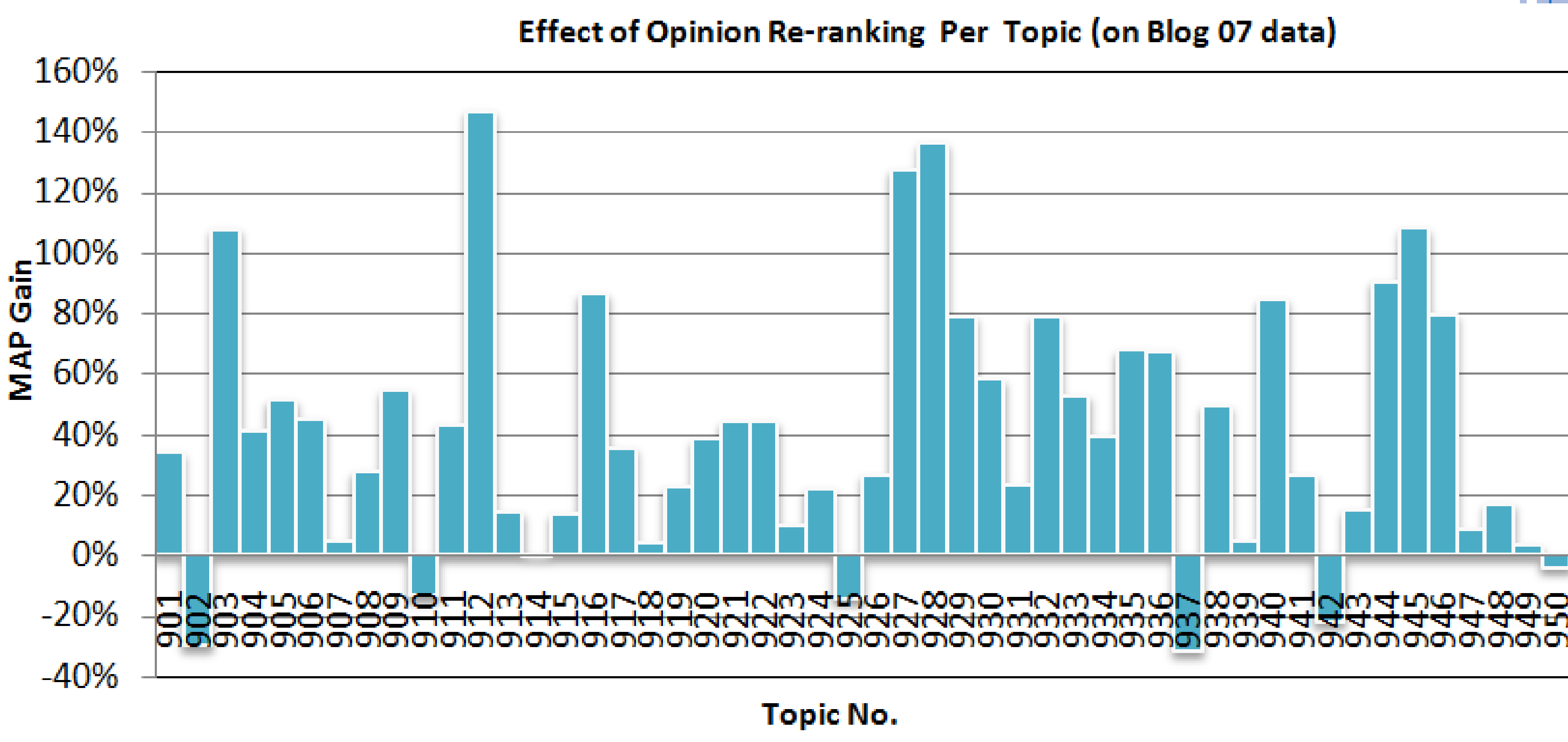


# 3. Opinion Retrieval Model Comparison (Cont.)

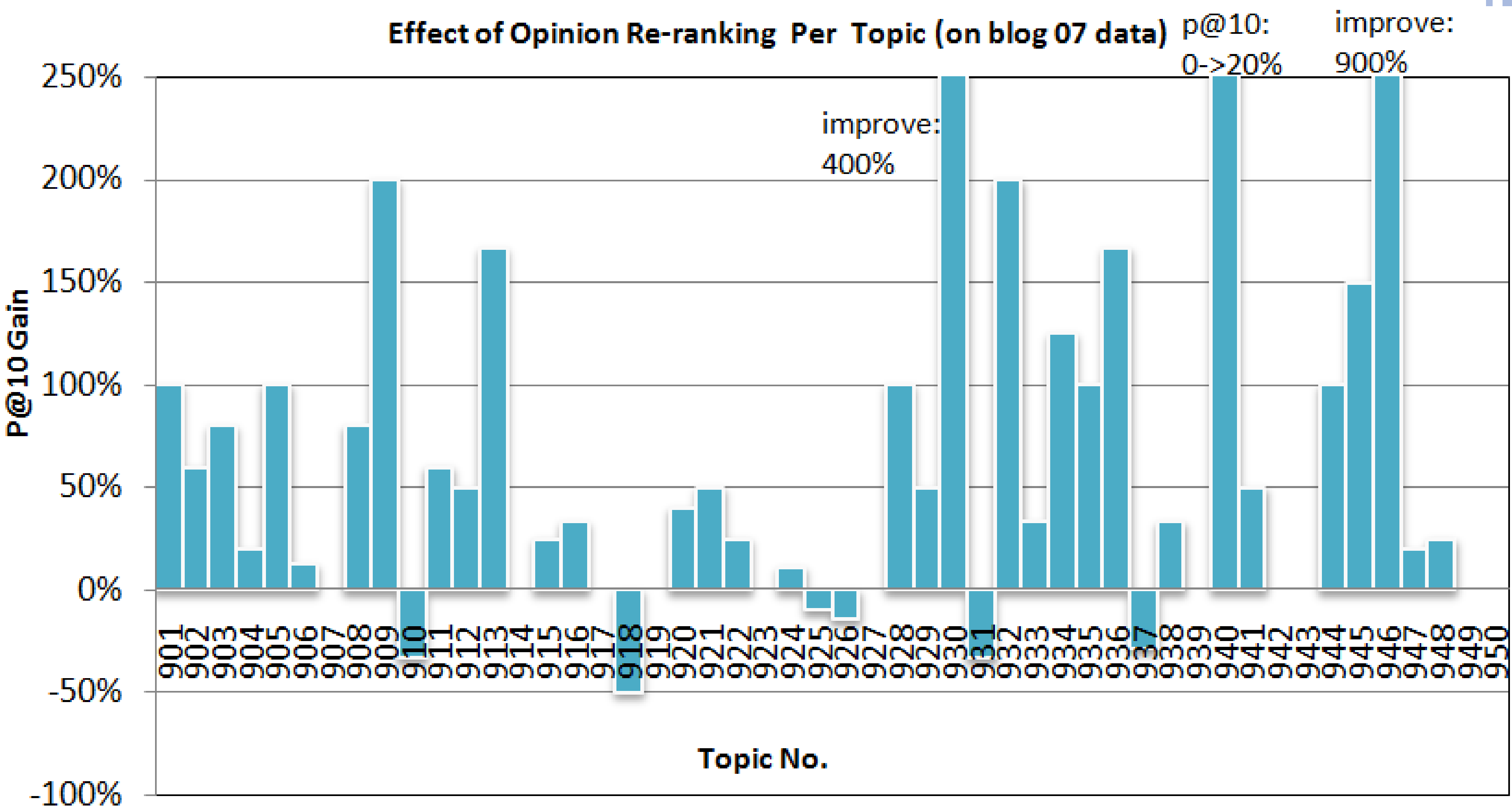
| Data Set | Method | MAP | R-Prec | P@10 |
|---|---|---|---|---|
| Blog 06 | Best run at blog 06 | 0.2052 | 0.2881 | 0.468 |
| | Best title-run at blog 06 | 0.1885 | 0.2771 | **0.512** |
| | Our Relevance Baseline(title-run) | 0.1758 | 0.2619 | 0.350 |
| | Our Unified Model | **0.2257** | **0.3038** | 0.507 |
| Blog 07 | Greatest improvement at blog 07 | 15.9% | 8.6% | 21.6% |
| | Our Relevance Baseline(title-run) | 0.2632 | 0.3249 | 0.432 |
| | Our Unified Model * | 0.3371 | 0.3896 | 0.606 |
| | improvement | **28.1%** | **19.9%** | **40.3%** |

*: **on Blog 07 data, use the same parameters as those on Blog 06 data,**
$\lambda$ =0.6, window = full, thesaurus: SentiWordNet

# 4. per-topic analysis (MAP Gain)



# 4. per-topic analysis (p@10 Gain)



# 5. Case Study

## Details of the best re-ranked topics examples

| Topic | Title | | Description | | |
|---|---|---|---|---|---|
| TREC 06 - 895 | Oprah | | Find opinions about Oprah Winfrey's TV show | | |
| | **MAP** | **Prec@10** | **Prec@30** | **Prec@100** | **Prec@1000** |
| Before re-ranking | 0.0687 | 0.2000 | 0.0333 | 0.1200 | 0.0640 |
| After re-ranking | 0.2721 | 0.8000 | 0.5000 | 0.3400 | 0.0640 |
| **Topic** | **Title** | | **Description** | | |
| TREC 07 - 946 | tivo | | Find opinions about TIVO brand digital video recorders | | |
| | **MAP** | **Prec@10** | **Prec@30** | **Prec@100** | **Prec@1000** |
| Before re-ranking | 0.2779 | 0.1000 | 0.3333 | 0.3900 | 0.2650 |
| After re-ranking | 0.4991 | 1.0000 | 0.9667 | 0.8300 | 0.2650 |

# Topic 946 – example 1

Information Retriever @ TsingHua University

BLOG06-20051229-025-0029161424, ranked 283 → ranked1

- TiVo's custom wireless G USB adapter has arrived. As previously reported the adapter is compatible with Series 2 models and off-loads some of the network processing, by utilizing a full MAC chipset..." Sells for $50 ……

Factorial description, no opinion

- I understand it s not the adapter but the TiVo software, but I don't care. **TiVo could fix that pretty easy but** chooses not to at the expense of our security. I run my TiVo wireless. Mainly because...
- **I guess it s a good thing indirectly, less incentive to watch TV. Personally I think that it s nice to be able to organise TV life around your life**
- **Not having WPA is a fault of the Tivo software**, not the device (AFAIK this is not a "smart" device). I run all my Tivos with Netgear WG111s... **I ll also be a bit miffed** if they turn out to get decent transfer speed with these things as Tivo just sold me the 111s **as their fastest transfer method,** ……

# Topic 946 – example 2

- <DOCNO>BLOG06-20051225-017-0000132016</DOCNO>

<content>

Ranked 5 → Ranked 306

Tivo

Monday, August 15, 2005

The Internet s tivo software Resource Search Google: Other tivo software Resources: Hacking the TiVo TiVo is a trademark of TiVo Inc. This site or software on this site is in no way affiliated with or endorsed by TiVo Inc. TiVo Community Forum - powered by vBulletin This also includes hacks that remove ads from TiVo software. 17385 2434... USATODAY.com - TiVo investors give standing ovation to Comcas..

posted by mdb @  

0 Comments:

</content>

</DOC>

# Conclusion

- Proposed a formal generation opinion retrieval model
  - Topic relevance & sentimental scores are integrated with quadratic comb.
- Opinion generation ranking functions are derived
  - Using the language modeling approach with smoothing
  - With logarithm normalization paradigm
- Discussed the roles of the sentiment lexicon and the matching window.
- It is a general model for opinion retrieval
  - Domain-independent lexicons
  - No assumption has been made on the nature of blog-structured text

# Future work

- Automatically constructing collection-based sentiment lexicons
- Understanding the nature of opinion expressing behavior on the Web
- Go beyond document re-ranking
  - Opinion-oriented index
  - Use linguistic information
  - ……

*Thanks for your attention!*

*Questions & comments?*