Short Text Feature Enrichment

--Using Link Analysis on Topic-Keyword Graph

Peng Wang

Joint work with H. Zhang, B. Xu, H.W. Hao, and Chenglin Liu

Computational-Brain Research Center
Institute of Automation, Chinese Academy of Sciences

CASIA

Outline

- Introduction
- Our Method
  - Short Text Modeling with Topic Model
  - Re-rank Keywords under Topics
  - Construct Topic-keyword Graph
  - Extract Candidate Keywords
  - Expand Short Text
- Evaluation
- Conclusions

Outline

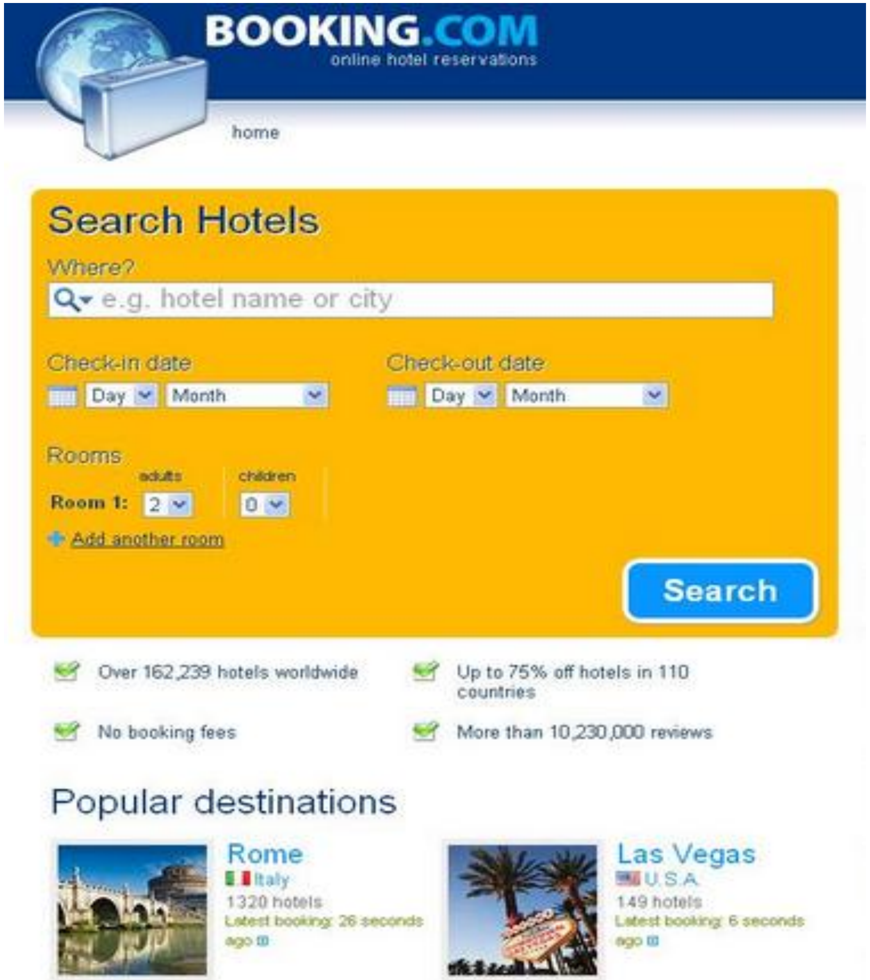
- Introduction
- Our Method
  - Short Text Modeling with Topic Model
  - Re-rank Keywords under Topics
  - Construct Topic-keyword Graph
  - Extract Candidate Keywords
  - Expand Short Text
- Evaluation
- Conclusions

Introduction

Bai du 知道

新浪微博 weibo.com

YAHOO! NEWS



Introduction (contd.)

- The Explosion of
  - e-commerce, online communication, & online publishing, products ordering…
- Typical Examples
  - Web search snippets
  - Short messages, advertising messages
  - Book/movie summaries
  - Product descriptions & customer reviews
  - News feeds
  - Forum/chat messages
  - *Sina* micro-blog, twitter.
  - Descriptions of entities: people, companies, hotels, etc

Introduction (contd.)

- **Challenges**
  - **Compared with normal text,**
    - Noise (class irrelevant), Irregular
    - Short and Sparsity
    - Less topic-focused
  - **Especially for BOW,**
    - ignores the textual information
    - lacks semantic knowledge
    - Less co-occurrences
- **Existing Work**
  - Phan www'2008:
    - Learn LDA based on Wikipedia as external corpus;
    - Make inference on short text collection;
    - Expand and enrich the short text;
    - better similarity measurement;

WIKIPEDIA
The Free Encyclopedia

Introduction (contd.)

- **Existing Work (contd.)**
  - Chen et al. AAAI’2011
    - Improved by learning multi-granularity topics
  - Yan et al. www’2013
    - proposed a biterm topic model;
    - model topics over the whole corpus instead of document-level.
  - Zhu et al. UAI’2011
    - present a non-probabilistic topic modeling method, which can control the sparsity.
  - Sun SIGIR’2012
    - A simple method using representative words as query to search a few of labeled samples, and the majority vote of the search results is the predictable category.
  - Gabrilovich IJCAI’2007
    - a method to improve text classification performance by enriching document representation with Wikipedia concepts

Our Framework

- Our framework based on TM & Link Analysis,

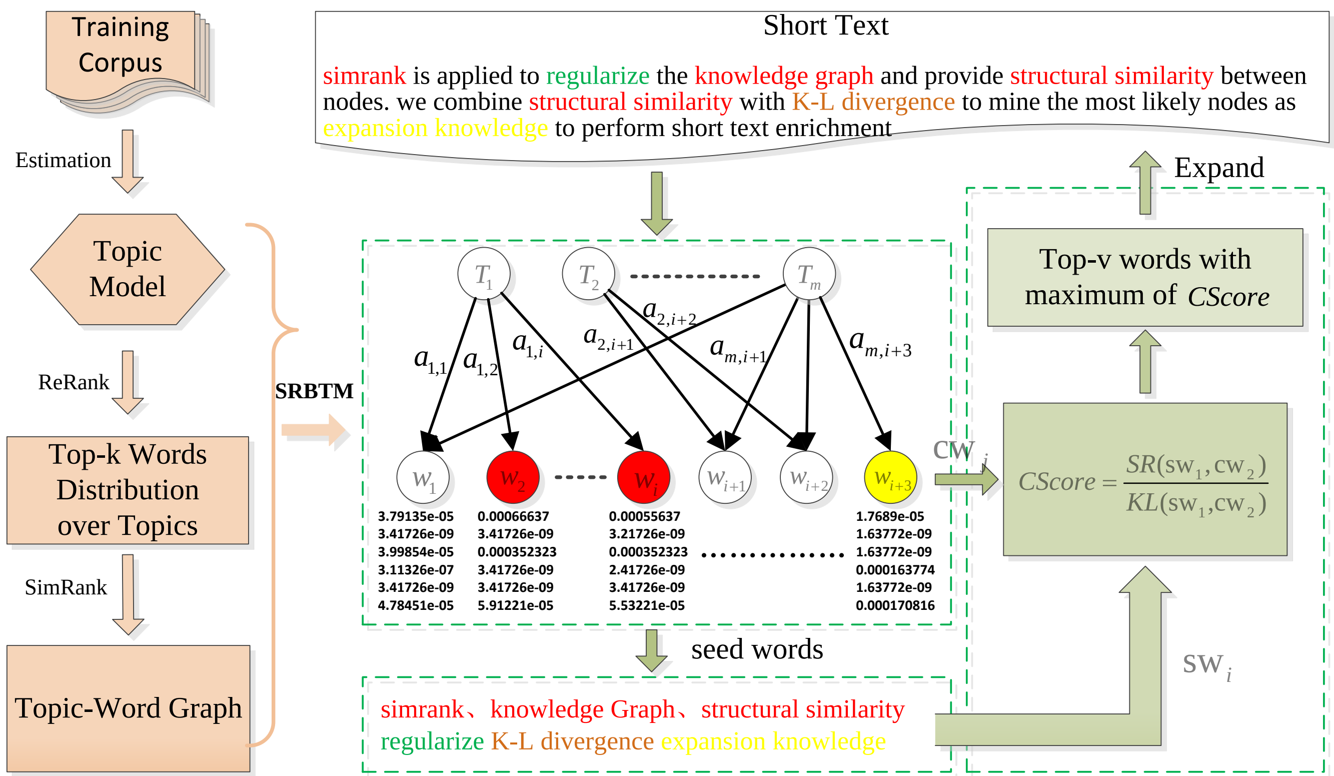


Fig 1. Method for short text expansion

Short Text Modeling with Topic Model

- Blei et.al. JMLR'2003
  - firstly proposed LDA and used it to estimate multinomial observations by unsupervised learning.
  - based on an assumption of document generation process,

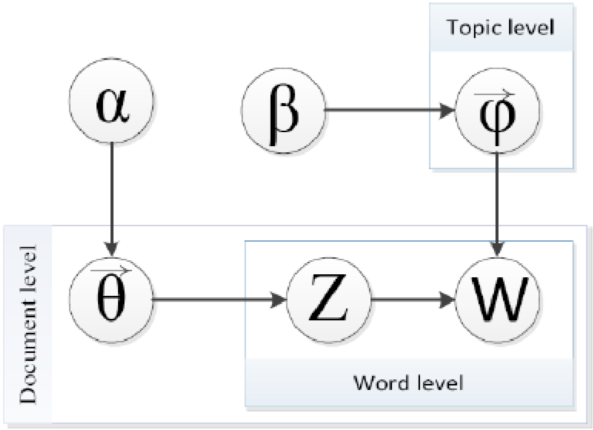


Fig.2 Grapical model of LDA

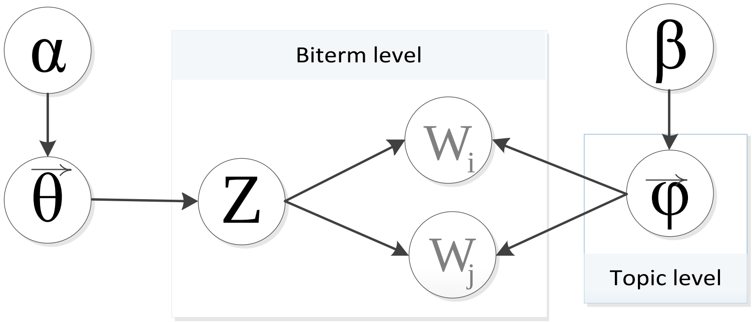


Fig.3 Grapical model of BTM

- Yan et al. www'2013
  - Based on LDA, proposed BTM especially for short text.
  - Directly model the word co-occurrences in the whole corpus to make full use of the global information.

Short Text Modeling with Topic Model

➢ To extract topics, we learn parameters of BTM.

- words from each topic are related,
- representation more topic-focused,
- to alleviate sparsity and noise .

Table 2. Most likely words of some topics

Topic0:music band rock album song songs released
Topic1:species food animals animal plants humans
Topic2: energy mass field quantum particles force
Topic3: india indian hindu pakistan sanskrit century
Topic4: blood body brain heart cells muscle syndrome
Topic5: water carbon oil chemical gas process oxygen
Topic6: government party president constitution
Topic7: power energy solar electric electrical
Topic8: ystem data code software computer
Topic9: horse opponent horses body hand match
Topic10: south africa united country islands world

Table 1. Variables in BTM

Para.	**Details**
M	number of bi-terms
α, β	Para. for Dirichlet
$\bar{\theta}$	topic distribution
z	index of a topic
$\overline{\varphi_{i,z}}$	i th word distribution
V	vocabulary size
K	the number of topics
$\mathbf{\Phi}$	a $K \times V$ matrix
BT	corpus with M biterms

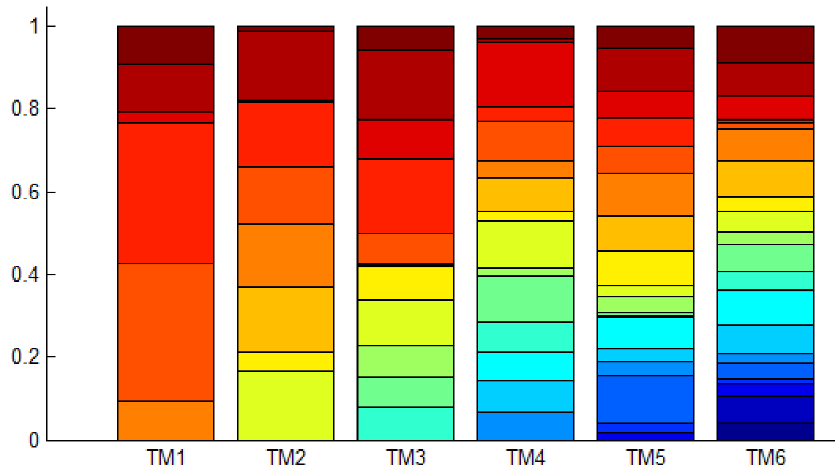


Fig.4 Topic distribution over keywords

Re-rank Keywords under Topics

- The long tail like distribution of keywords under topics,

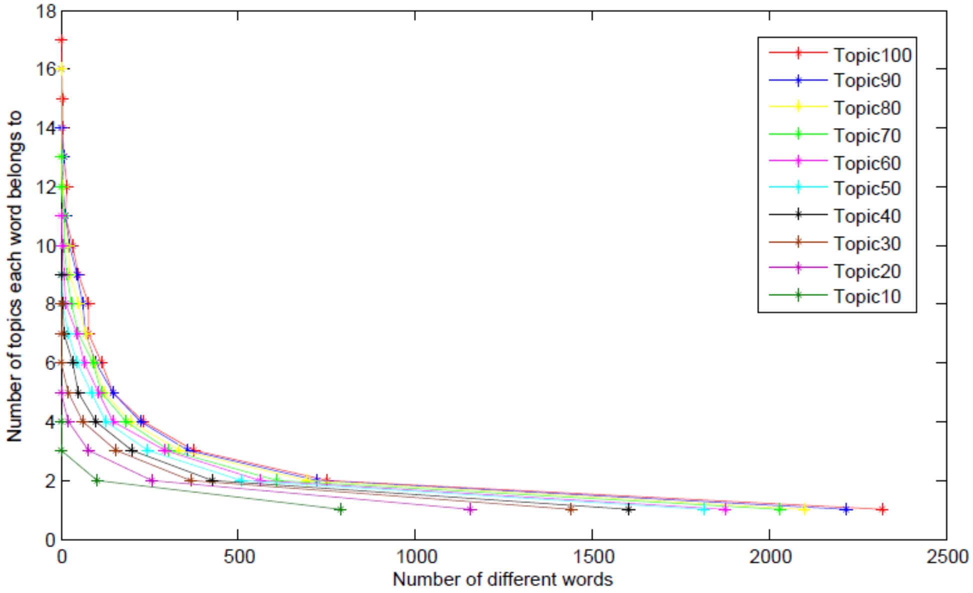
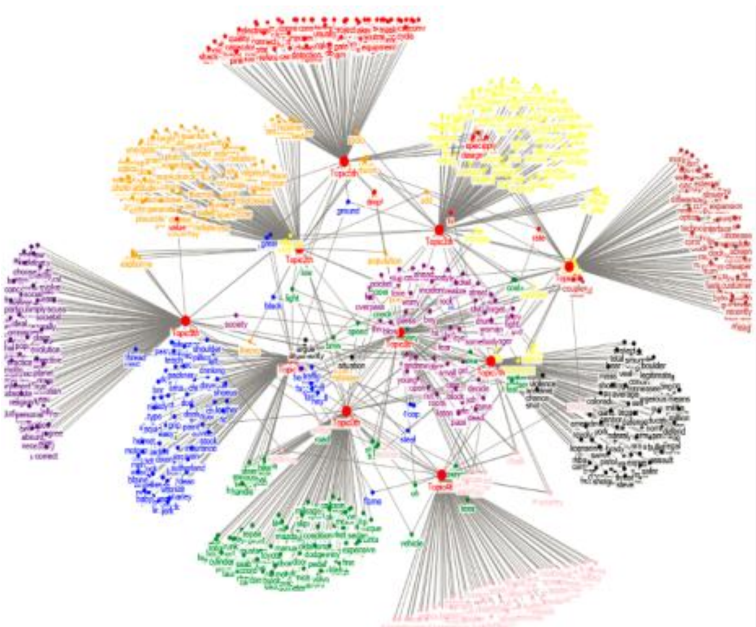


Fig.5 The keywords distribution under topics

$$SAS = \frac{e^{\hat{\varphi}_{z,i}}}{\sum_{m=1}^{M} e^{\hat{\varphi}_{z,i}}} \qquad (1)$$

where $\hat{\varphi}_{z,i}$ is the probability distribution of the *i*th word under topic *k*.

Construct Topic-keyword Graph



Re-Rank

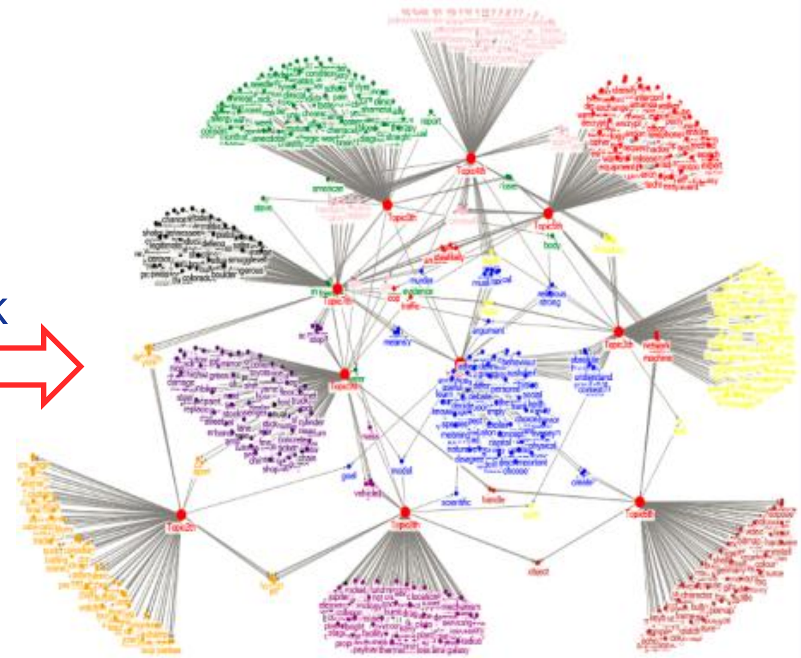
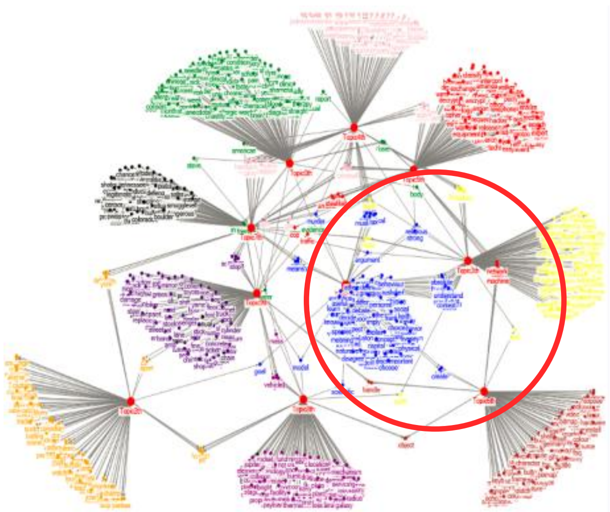


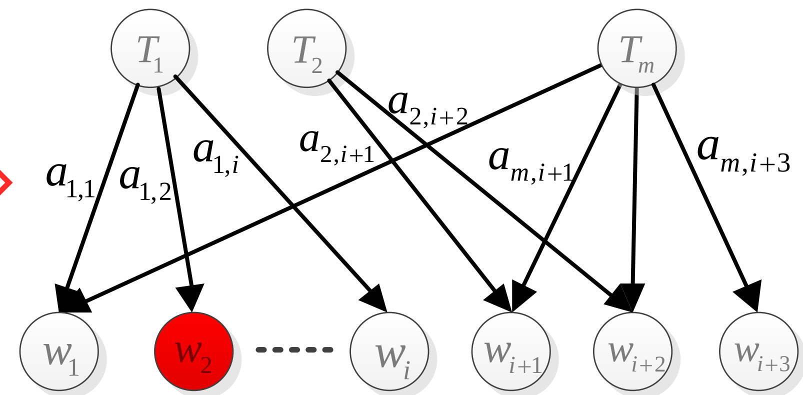
Fig.6 native topic-keyword graph

Fig.7 re-Ranked topic-keyword graph

Extract Candidate Keywords

Link Analysis:





$$s(w_a, w_b) = \begin{cases} 1, & \text{if } w_a = w_b \\ \dfrac{C}{|I(w_a)||I(w_b)|} \displaystyle\sum_{i=1}^{|I(w_a)|} \sum_{j=1}^{|I(w_b)|} s(I_i(w_a), I_j(w_b)), & \text{if } w_a \neq w_b \end{cases}, \quad (2)$$

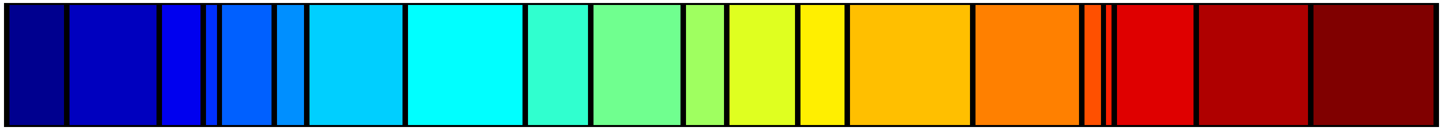
Expand Short Text

$$SR(w_a, w_b) = SAS(w_a)SAS(w_b)s(w_a, w_b) \,, \qquad (3)$$

$$KL(sw_i, cw_j) = \frac{1}{2}\left[D\left(p^{(z)}_{sw_i} \,\|\, \frac{p^{(z)}_{sw_i} + p^{(z)}_{cw_j}}{2}\right) + D\left(p^{(z)}_{cw_j} \,\|\, \frac{p^{(z)}_{cw_j} + p^{(z)}_{sw_i}}{2}\right)\right] \qquad (4)$$

$$CScore(sw_i, cw_j) = \frac{SR(sw_i, cw_j)}{KL(sw_i, cw_j)} \,, \qquad (5)$$

Keywords with *Cscore():*



Provided that,

$$\overrightarrow{w_m} = \{\textit{Modeling short text based on Wikipedia and LDA}\}$$

$$\overrightarrow{w_m} = \{\underbrace{\textit{Modeling short text based on Wikipedia and LDA}}_{\text{Original feature}}$$  $\}$

Original feature

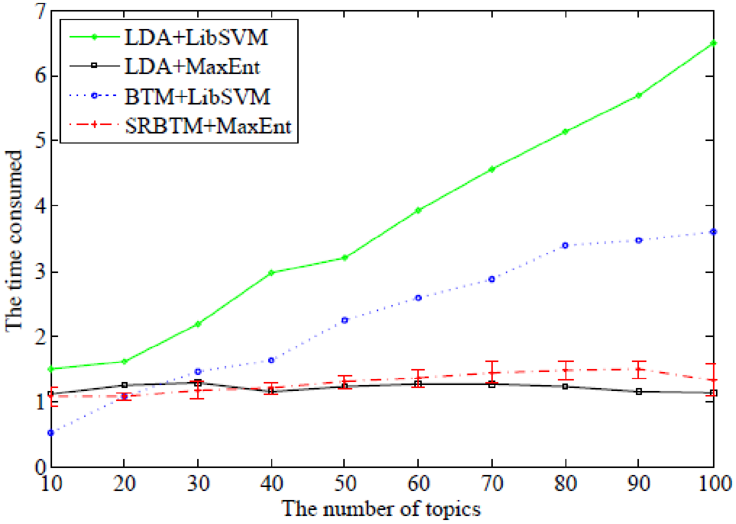
Expanded info.

Evaluation

Experimental data

- Search snippets dataset, consists of 10,060 training snippets and 2,280 test snippets from 8 categories, as shown in Table 3. On average, each snippet has 18.07 words

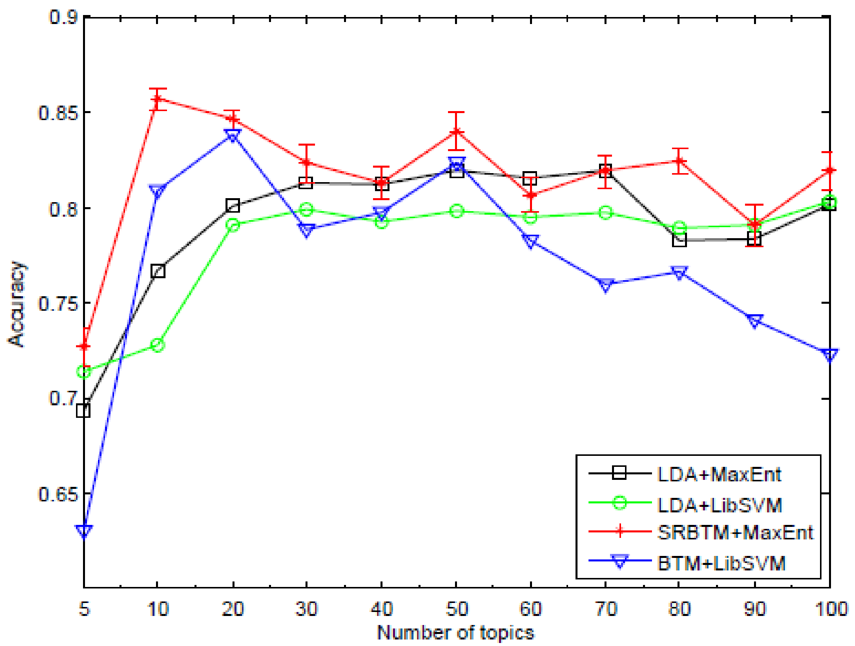
Domain	Tr_snippets	Te_snippets
Business	1200	300
Computers	1200	300
Cult.-arts-ente.	1880	330
Edu.-Science	2360	300
Engineering	220	150
Health	880	300
Politics-Society	1200	300
Sports	1120	300
Total	10060	2280

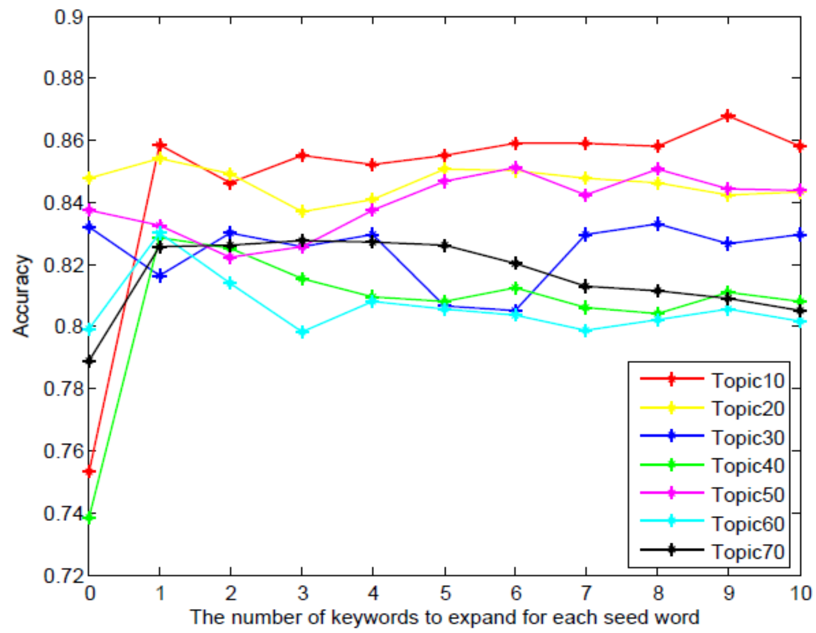


- Collected by Phan professor.
- Used in Phan and Nguyen WWW'2008, Chen et al. AAAI'2011;

Evaluation

❖ Experimental Results:





Conclusions

- Achievements
  - Re-rank the candidates keywords under topics;
  - Construct the topic-keywords graph;
  - Extract keywords using link analysis;
  - Expand the short text by Keywords.
- Remaining Issues
  - The BTM is time-consuming;
  - Noise is still may be introduced.

Thanks For Your Time !