

# Automatic Keyphrase Extraction via Topic Decomposition

Zhiyuan Liu, Wenyi Huang, Yabin Zheng and Maosong Sun

**Presenter: Wenyi Huang**

Department of Computer Science and Technology  
State Key Lab on Intelligent Technology and Systems  
National Lab for Information Science and Technology  
Tsinghua University

Oct 9, 2010



- What is keyphrase extraction?
- Method
  - Supervised
    - Learning algorithms for keyphrase extraction (Turney, 2000)
  - Unsupervised
    - TFIDF
    - TextRank: Bringing order into texts (Rada Mihalcea and Paul Tarau. 2004)



- What is keyphrase extraction?
- Method
  - Supervised
    - Learning algorithms for keyphrase extraction (Turney, 2000)
  - Unsupervised
    - TFIDF
    - TextRank: Bringing order into texts (Rada Mihalcea and Paul Tarau. 2004)



- What is keyphrase extraction?
- Method
  - Supervised
    - Learning algorithms for keyphrase extraction (Turney, 2000)
  - Unsupervised
    - TFIDF
    - TextRank: Bringing order into texts (Rada Mihalcea and Paul Tarau. 2004)



- What is keyphrase extraction?
- Method
  - Supervised
    - Learning algorithms for keyphrase extraction (Turney, 2000)
  - Unsupervised
    - TFIDF
    - TextRank: Bringing order into texts (Rada Mihalcea and Paul Tarau. 2004)



- What is keyphrase extraction?
- Method
  - Supervised
    - Learning algorithms for keyphrase extraction (Turney, 2000)
  - Unsupervised
    - TFIDF
    - TextRank: Bringing order into texts (Rada Mihalcea and Paul Tarau. 2004)



- What is keyphrase extraction?
- Method
  - Supervised
    - Learning algorithms for keyphrase extraction (Turney, 2000)
  - Unsupervised
    - TFIDF
    - TextRank: Bringing order into texts (Rada Mihalcea and Paul Tarau. 2004)



- What is keyphrase extraction?
- Method
  - Supervised
    - Learning algorithms for keyphrase extraction (Turney, 2000)
  - Unsupervised
    - TFIDF
    - TextRank: Bringing order into texts (Rada Mihalcea and Paul Tarau. 2004)





## What about topic ?

**Relevance** Good keyphrases should be relevant to the major topics of the given document.

**Coverage** An appropriate set of keyphrases should also have a good coverage of a document's major topics.



## What about topic ?

**Relevance** Good keyphrases should be relevant to the major topics of the given document.

**Coverage** An appropriate set of keyphrases should also have a good coverage of a document's major topics.



# Building Topic Interpreters

**Method** Latent Dirichlet Allocation (LDA)

**Datasets** Wikipedia snapshot at March 2008

word	prob.	word	prob.	word	prob.
DRUGS	.069	MIND	.081	DOCTOR	.074
DRUG	.060	THOUGHT	.066	DR.	.063
MEDICINE	.027	REMEMBER	.064	PATIENT	.061
EFFECTS	.026	MEMORY	.037	HOSPITAL	.049
BODY	.023	THINKING	.030	CARE	.046
MEDICINES	.019	PROFESSOR	.028	MEDICAL	.042
PAIN	.016	FELT	.025	NURSE	.031
PERSON	.016	REMEMBERED	.022	PATIENTS	.029
MARIJUANA	.014	THOUGHTS	.020	DOCTORS	.028
LABEL	.012	FORGOTTEN	.020	HEALTH	.025
ALCOHOL	.012	MOMENT	.020	MEDICINE	.017
DANGEROUS	.011	THINK	.019	NURSING	.017
ABUSE	.009	THING	.016	DENTAL	.015
EFFECT	.009	WONDER	.014	NURSES	.013
KNOWN	.008	FORGET	.012	PHYSICIAN	.012
PILLS	.008	RECALL	.012	HOSPITALS	.011

**Figure:** An example of probabilistic topic model



# Topic-Decomposed PageRank

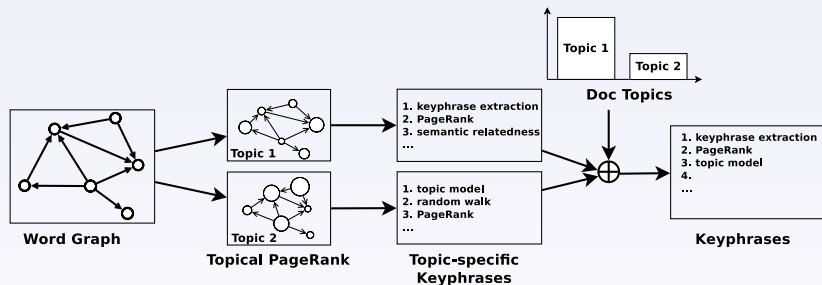
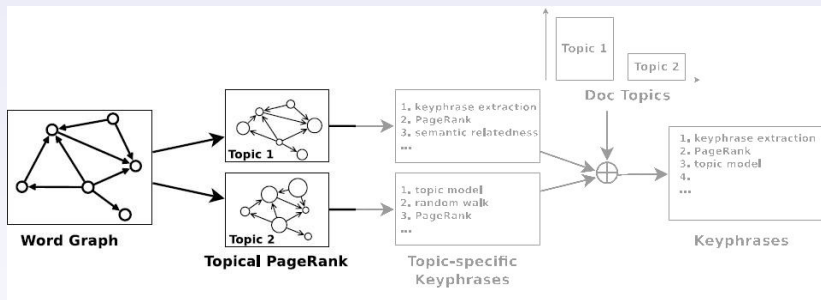


Figure: Topical PageRank for Keyphrase Extraction. (TPR)



# Calculate Ranking Scores by TPR

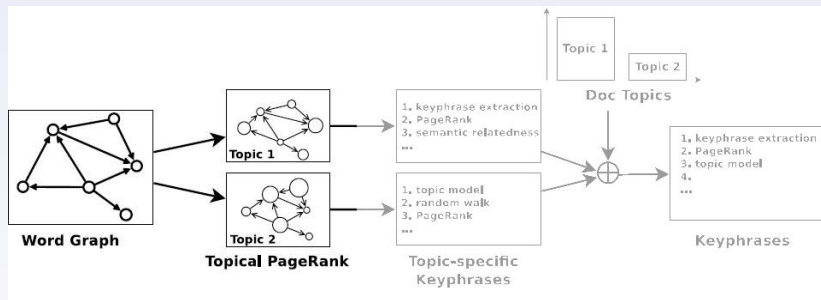


$$R_z(w_i) = \lambda \sum_{j:w_j \rightarrow w_i} \frac{e(w_j, w_i)}{O(w_j)} R_z(w_j) + (1 - \lambda) p_z(w_i). \quad (1)$$

- $p_z(w) = pr(w|z)$ , probability of word  $w$  given topic  $z$ .
- $p_z(w) = pr(z|w)$ , probability of topic  $z$  given word  $w$ .
- $p_z(w) = pr(w|z) \times pr(z|w)$ , product of hub and authority.



# Calculate Ranking Scores by TPR

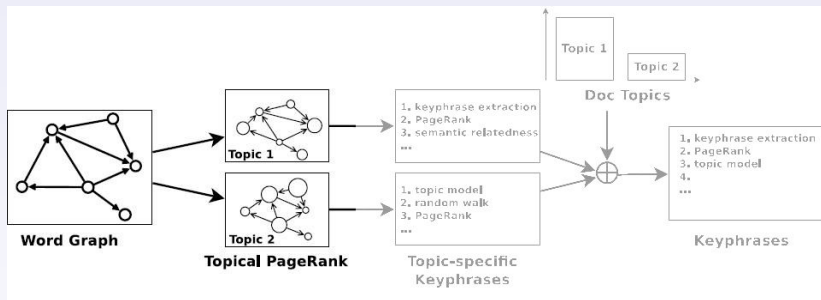


$$R_z(w_i) = \lambda \sum_{j:w_j \rightarrow w_i} \frac{e(w_j, w_i)}{O(w_j)} R_z(w_j) + (1 - \lambda) p_z(w_i). \quad (1)$$

- $p_z(w) = pr(w|z)$ , probability of word  $w$  given topic  $z$ .
- $p_z(w) = pr(z|w)$ , probability of topic  $z$  given word  $w$ .
- $p_z(w) = pr(w|z) \times pr(z|w)$ , product of hub and authority.



# Calculate Ranking Scores by TPR

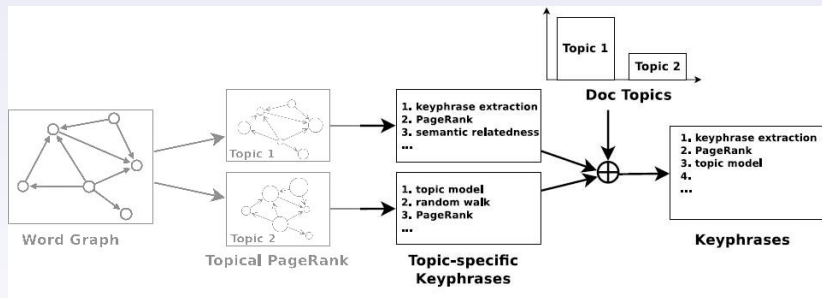


$$R_z(w_i) = \lambda \sum_{j:w_j \rightarrow w_i} \frac{e(w_j, w_i)}{O(w_j)} R_z(w_j) + (1 - \lambda) p_z(w_i). \quad (1)$$

- $p_z(w) = pr(w|z)$ , probability of word  $w$  given topic  $z$ .
- $p_z(w) = pr(z|w)$ , probability of topic  $z$  given word  $w$ .
- $p_z(w) = pr(w|z) \times pr(z|w)$ , product of hub and authority.



# Extract Keyphrases Using Ranking Scores



**Candidate Phrases** noun phrases (Hulth, 2003)

(adjective) \* (noun) +

**Doc topic distribution**  $pr(z|d)$  for each topic  $z$ .

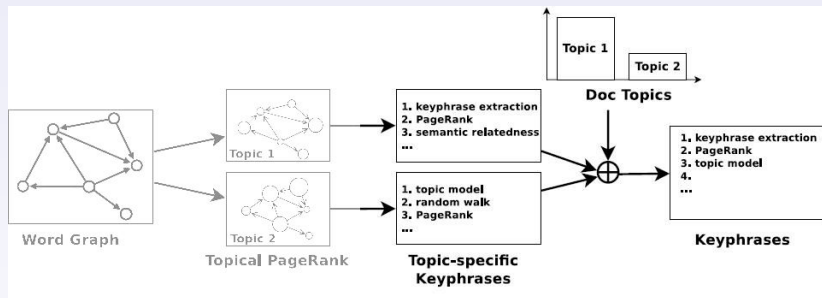
**Phrase Score**

$$R(p) = \sum_{z=1}^K R_z(p) \times pr(z|d).$$





# Extract Keyphrases Using Ranking Scores



**Candidate Phrases** noun phrases (Hulth, 2003)

(adjective) \* (noun) +

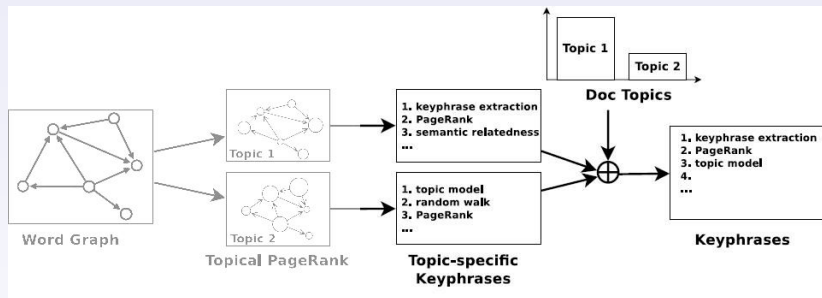
**Doc topic distribution**  $pr(z|d)$  for each topic  $z$ .

**Phrase Score**

$$R(p) = \sum_{z=1}^K R_z(p) \times pr(z|d).$$



# Extract Keyphrases Using Ranking Scores



**Candidate Phrases** noun phrases (Hulth, 2003)

(adjective) \* (noun) +

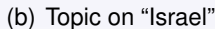
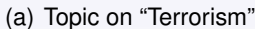
**Doc topic distribution**  $pr(z|d)$  for each topic  $z$ .

**Phrase Score**

$$R(p) = \sum_{z=1}^K R_z(p) \times pr(z|d).$$



## Arafat Says U.S. Threatening to Kill PLO Officials



# Experiments

## 1 Datasets

- NEWS: 308 news articles in DUC2001
- RESEARCH: 2,000 abstracts of research articles (Hulth, 2003)

## 2 Evaluation Metrics

- precision, recall, F-measure

$$p = \frac{c_{correct}}{c_{extract}}, \quad r = \frac{c_{correct}}{c_{standard}}, \quad f = \frac{2pr}{p+r}, \quad (3)$$

- binary preference measure (Bpref)

$$\text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \quad (4)$$

- mean reciprocal rank (MRR)

$$\text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d},$$



# Experiments

## 1 Datasets

- NEWS: 308 news articles in DUC2001
- RESEARCH: 2,000 abstracts of research articles (Hulth, 2003)

## 2 Evaluation Metrics

- precision, recall, F-measure

$$p = \frac{c_{correct}}{c_{extract}}, \quad r = \frac{c_{correct}}{c_{standard}}, \quad f = \frac{2pr}{p+r}, \quad (3)$$

- binary preference measure (Bpref)

$$\text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \quad (4)$$

- mean reciprocal rank (MRR)

$$\text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d},$$



# Experiments

## 1 Datasets

- NEWS: 308 news articles in DUC2001
- RESEARCH: 2,000 abstracts of research articles (Hulth, 2003)

## 2 Evaluation Metrics

- precision, recall, F-measure

$$p = \frac{c_{correct}}{c_{extract}}, \quad r = \frac{c_{correct}}{c_{standard}}, \quad f = \frac{2pr}{p+r}, \quad (3)$$

- binary preference measure (Bpref)

$$\text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \quad (4)$$

- mean reciprocal rank (MRR)

$$\text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d},$$



# Experiments

## 1 Datasets

- NEWS: 308 news articles in DUC2001
- RESEARCH: 2,000 abstracts of research articles (Hulth, 2003)

## 2 Evaluation Metrics

- precision, recall, F-measure

$$p = \frac{c_{correct}}{c_{extract}}, \quad r = \frac{c_{correct}}{c_{standard}}, \quad f = \frac{2pr}{p+r}, \quad (3)$$

- binary preference measure (Bpref)

$$\text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \quad (4)$$

- mean reciprocal rank (MRR)

$$\text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d},$$



# Experiments

## 1 Datasets

- NEWS: 308 news articles in DUC2001
- RESEARCH: 2,000 abstracts of research articles (Hulth, 2003)

## 2 Evaluation Metrics

- precision, recall, F-measure

$$p = \frac{c_{correct}}{c_{extract}}, \quad r = \frac{c_{correct}}{c_{standard}}, \quad f = \frac{2pr}{p+r}, \quad (3)$$

- binary preference measure (Bpref)

$$\text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \quad (4)$$

- mean reciprocal rank (MRR)

$$\text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d},$$





# Experiments

## 1 Datasets

- NEWS: 308 news articles in DUC2001
- RESEARCH: 2,000 abstracts of research articles (Hulth, 2003)

## 2 Evaluation Metrics

- precision, recall, F-measure

$$p = \frac{c_{correct}}{c_{extract}}, \quad r = \frac{c_{correct}}{c_{standard}}, \quad f = \frac{2pr}{p+r}, \quad (3)$$

- binary preference measure (Bpref)

$$\text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \quad (4)$$

- mean reciprocal rank (MRR)

$$\text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d},$$



# Experiments

## 1 Datasets

- NEWS: 308 news articles in DUC2001
- RESEARCH: 2,000 abstracts of research articles (Hulth, 2003)

## 2 Evaluation Metrics

- precision, recall, F-measure

$$p = \frac{c_{correct}}{c_{extract}}, \quad r = \frac{c_{correct}}{c_{standard}}, \quad f = \frac{2pr}{p+r}, \quad (3)$$

- binary preference measure (Bpref)

$$\text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \quad (4)$$

- mean reciprocal rank (MRR)

$$\text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d},$$



# Influences of Parameters - The Number of Topics $K$

$K$	Pre.	Rec.	F.	Bpref	MRR
50	0.268	0.330	0.296	0.204	0.632
100	0.276	0.340	0.304	0.208	0.632
500	0.284	0.350	0.313	0.215	0.648
1000	0.282	0.348	0.312	0.214	0.638
1500	0.282	0.348	0.311	0.214	0.631

**Table:** Influence of the number of topics  $K$  when the number of keyphrases  $M=10$  on NEWS.



# Influences of Parameters - Damping Factor $\lambda$

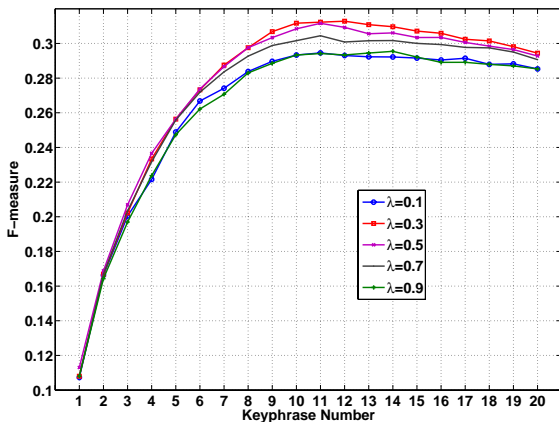


Figure: F-measure of TPR with  $\lambda=0.1, 0.3, 0.5, 0.7$  and  $0.9$  when  $M$  ranges from 1 to 20 on NEWS.



# Different Preference Values

Pref	Pre.	Rec.	F.	Bpref	MRR
$pr(w z)$	0.256	0.316	0.283	0.192	0.584
$pr(z w)$	0.282	0.348	0.312	0.214	0.638
prod	0.259	0.320	0.286	0.193	0.587

**Table:** Influence of three preference value settings when the number of keyphrases  $M = 10$  on NEWS.



# Comparing with Baseline Methods

Method	Pre.	Rec.	F.	Bpref	MRR
TFIDF	0.239	0.295	0.264	0.179	0.576
PageRank	0.242	0.299	0.267	0.184	0.564
LDA	0.259	0.320	0.286	0.194	0.518
TPR	<b>0.282</b>	<b>0.348</b>	<b>0.312</b>	<b>0.214</b>	<b>0.638</b>

**Table:** Comparing results on `NEWS` when the number of keyphrases  $M=10$ .

Method	Pre.	Rec.	F.	Bpref	MRR
TFIDF	0.333	0.173	0.227	0.255	0.565
PageRank	0.330	0.171	0.225	0.263	0.575
LDA	0.332	0.172	0.227	0.254	0.548
TPR	<b>0.354</b>	<b>0.183</b>	<b>0.242</b>	<b>0.274</b>	<b>0.583</b>

**Table:** Comparing results on `RESEARCH` when the number of keyphrases  $M=5$ .



# Comparing with Baseline Methods

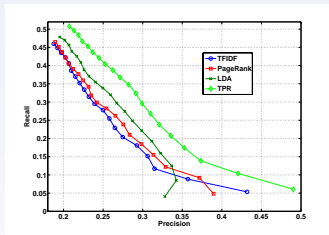


Figure: Precision-recall results on NEWS,  $M$  ranges from 1 to 20.

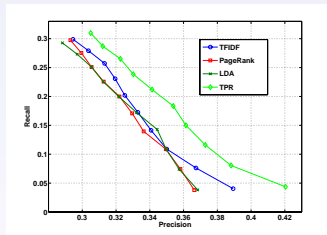


Figure: Precision-recall results on RESEARCH,  $M$  ranges from 1 to 10.



# Conclusion

- TPR outperform all baselines on both datasets
- TPR enjoys advantages of both LDA and TFIDF/PageRank methods
- Bpref and MRR serve as supplemental metrics for evaluation





# Conclusion

- TPR outperform all baselines on both datasets
- TPR enjoys advantages of both LDA and TFIDF/PageRank methods
- Bpref and MRR serve as supplemental metrics for evaluation



# Conclusion

- TPR outperform all baselines on both datasets
- TPR enjoys advantages of both LDA and TFIDF/PageRank methods
- Bpref and MRR serve as supplemental metrics for evaluation



# Thank You !

## QUESTIONS ?

My Homepage

<http://nlp.csai.tsinghua.edu.cn/~hwy/>

