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
Introduction

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)
Introduction

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)
Introduction

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)
Introduction

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)
Introduction

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)
Motivation

<table>
<thead>
<tr>
<th>What about topic?</th>
</tr>
</thead>
</table>
| **Relevance**    | Good keyphrases should be relevant to the major topics of the given document.  
<p>| <strong>Coverage</strong>     | An appropriate set of keyphrases should also have a good coverage of a document’s major topics. |</p>
<table>
<thead>
<tr>
<th>What about topic?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relevance</strong></td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
</tr>
</tbody>
</table>
Building Topic Interpreters

**Method**  Latent Dirichlet Allocation (LDA)

**Datasets**  Wikipedia snapshot at March 2008

---

**Table:**

<table>
<thead>
<tr>
<th>word</th>
<th>prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRUGS</td>
<td>.069</td>
</tr>
<tr>
<td>DRUG</td>
<td>.060</td>
</tr>
<tr>
<td>MEDICINE</td>
<td>.027</td>
</tr>
<tr>
<td>EFFECTS</td>
<td>.026</td>
</tr>
<tr>
<td>BODY</td>
<td>.023</td>
</tr>
<tr>
<td>MEDICINES</td>
<td>.019</td>
</tr>
<tr>
<td>PAIN</td>
<td>.016</td>
</tr>
<tr>
<td>PERSON</td>
<td>.016</td>
</tr>
<tr>
<td>MARIJUANA</td>
<td>.014</td>
</tr>
<tr>
<td>LABEL</td>
<td>.012</td>
</tr>
<tr>
<td>ALCOHOL</td>
<td>.012</td>
</tr>
<tr>
<td>DANGEROUS</td>
<td>.011</td>
</tr>
<tr>
<td>ABUSE</td>
<td>.009</td>
</tr>
<tr>
<td>EFFECT</td>
<td>.009</td>
</tr>
<tr>
<td>KNOWN</td>
<td>.008</td>
</tr>
<tr>
<td>PILLS</td>
<td>.008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>word</th>
<th>prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIND</td>
<td>.081</td>
</tr>
<tr>
<td>THOUGHT</td>
<td>.066</td>
</tr>
<tr>
<td>REMEMBER</td>
<td>.064</td>
</tr>
<tr>
<td>MEMORY</td>
<td>.037</td>
</tr>
<tr>
<td>THINKING</td>
<td>.030</td>
</tr>
<tr>
<td>PROFESSOR</td>
<td>.028</td>
</tr>
<tr>
<td>FELT</td>
<td>.025</td>
</tr>
<tr>
<td>REMEMBERED</td>
<td>.022</td>
</tr>
<tr>
<td>THOUGHTS</td>
<td>.020</td>
</tr>
<tr>
<td>FORGOTTEN</td>
<td>.020</td>
</tr>
<tr>
<td>MOMENT</td>
<td>.020</td>
</tr>
<tr>
<td>THINK</td>
<td>.019</td>
</tr>
<tr>
<td>THING</td>
<td>.016</td>
</tr>
<tr>
<td>WONDER</td>
<td>.014</td>
</tr>
<tr>
<td>FORGET</td>
<td>.012</td>
</tr>
<tr>
<td>RECALL</td>
<td>.012</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>word</th>
<th>prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOCTOR</td>
<td>.074</td>
</tr>
<tr>
<td>DR.</td>
<td>.063</td>
</tr>
<tr>
<td>PATIENT</td>
<td>.061</td>
</tr>
<tr>
<td>HOSPITAL</td>
<td>.049</td>
</tr>
<tr>
<td>CARE</td>
<td>.046</td>
</tr>
<tr>
<td>MEDICAL</td>
<td>.042</td>
</tr>
<tr>
<td>NURSE</td>
<td>.031</td>
</tr>
<tr>
<td>PATIENTS</td>
<td>.029</td>
</tr>
<tr>
<td>DOCTORS</td>
<td>.028</td>
</tr>
<tr>
<td>HEALTH</td>
<td>.025</td>
</tr>
<tr>
<td>MEDICINE</td>
<td>.017</td>
</tr>
<tr>
<td>NURSING</td>
<td>.017</td>
</tr>
<tr>
<td>DENTAL</td>
<td>.015</td>
</tr>
<tr>
<td>NURSES</td>
<td>.013</td>
</tr>
<tr>
<td>PHYSICIAN</td>
<td>.012</td>
</tr>
<tr>
<td>HOSPITAL</td>
<td>.011</td>
</tr>
</tbody>
</table>

**Figure:** An example of probabilistic topic model
Topic-Decomposed PageRank

Figure: Topical PageRank for Keyphrase Extraction. (TPR)
Calculate Ranking Scores by TPR

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

- \( p_z(w) = pr(w|z) \), 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.

Wenyi Huang  Dept. CS&T, THU  Automatic Keyphrase Extraction via Topic Decomposition
Calculate Ranking Scores by TPR

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

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

- \( p_z(w) = \text{pr}(w|z) \), probability of word \( w \) given topic \( z \).
- \( p_z(w) = \text{pr}(z|w) \), probability of topic \( z \) given word \( w \).
- \( p_z(w) = \text{pr}(w|z) \times \text{pr}(z|w) \), 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

Examples

(a) Topic on “Terrorism”

(b) Topic on “Israel”

(c) Topic on “U.S.”

(d) TPR Result
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
  \[
  p = \frac{c_{\text{correct}}}{c_{\text{extract}}}, \quad r = \frac{c_{\text{correct}}}{c_{\text{standard}}}, \quad f = \frac{2pr}{p+r},
  \]
- binary preference measure (Bpref)
  \[
  \text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}.
  \]
- 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
     \[ p = \frac{c_{\text{correct}}}{c_{\text{extract}}}, \quad r = \frac{c_{\text{correct}}}{c_{\text{standard}}}, \quad f = \frac{2pr}{p+r}, \quad (3) \]
   - 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)
     \[ \text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d}, \quad (5) \]
Experiments

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

Evaluation Metrics
- precision, recall, F-measure
  \[ p = \frac{c_{\text{correct}}}{c_{\text{extract}}}, \quad r = \frac{c_{\text{correct}}}{c_{\text{standard}}}, \quad f = \frac{2pr}{p+r}, \]  
- binary preference measure (Bpref)
  \[ \text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \]
- 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
  \[ p = \frac{c_{\text{correct}}}{c_{\text{extract}}}, \quad r = \frac{c_{\text{correct}}}{c_{\text{standard}}}, \quad f = \frac{2pr}{p+r}, \quad (3) \]
- 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)
  \[ \text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d}, \quad (5) \]

Wenyi Huang  Dept. CS&T, THU
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

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

- binary preference measure (Bpref)

\[
B\text{pref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}.
\] (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

\[
p = \frac{c_{\text{correct}}}{c_{\text{extract}}}, \quad r = \frac{c_{\text{correct}}}{c_{\text{standard}}}, \quad f = \frac{2pr}{p + r}. \tag{3}
\]

- binary preference measure (Bpref)

\[
\text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \tag{4}
\]

- 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
     \[ p = \frac{c_{\text{correct}}}{c_{\text{extract}}}, \quad r = \frac{c_{\text{correct}}}{c_{\text{standard}}}, \quad f = \frac{2pr}{p + r}. \] (3)
   - binary preference measure (Bpref)
     \[ \text{Bpref} = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}. \] (4)
   - mean reciprocal rank (MRR)
     \[ \text{MRR} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{\text{rank}_d}. \] (5)
Influences of Parameters - The Number of Topics $K$

<table>
<thead>
<tr>
<th>$K$</th>
<th>Pre.</th>
<th>Rec.</th>
<th>F.</th>
<th>Bpref</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.268</td>
<td>0.330</td>
<td>0.296</td>
<td>0.204</td>
<td>0.632</td>
</tr>
<tr>
<td>100</td>
<td>0.276</td>
<td>0.340</td>
<td>0.304</td>
<td>0.208</td>
<td>0.632</td>
</tr>
<tr>
<td>500</td>
<td>0.284</td>
<td>0.350</td>
<td>0.313</td>
<td>0.215</td>
<td>0.648</td>
</tr>
<tr>
<td>1000</td>
<td>0.282</td>
<td>0.348</td>
<td>0.312</td>
<td>0.214</td>
<td>0.638</td>
</tr>
<tr>
<td>1500</td>
<td>0.282</td>
<td>0.348</td>
<td>0.311</td>
<td>0.214</td>
<td>0.631</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>Pref</th>
<th>Pre.</th>
<th>Rec.</th>
<th>F.</th>
<th>Bpref</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$pr(w</td>
<td>z)$</td>
<td>0.256</td>
<td>0.316</td>
<td>0.283</td>
<td>0.192</td>
</tr>
<tr>
<td>$pr(z</td>
<td>w)$</td>
<td>0.282</td>
<td>0.348</td>
<td>0.312</td>
<td>0.214</td>
</tr>
<tr>
<td>prod</td>
<td>0.259</td>
<td>0.320</td>
<td>0.286</td>
<td>0.193</td>
<td>0.587</td>
</tr>
</tbody>
</table>

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

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

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

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

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

Wenyi Huang  Dept. CS&T, THU  Automatic Keyphrase Extraction via Topic Decomposition
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/