Supervised Keyphrase Extraction Leveraging Candidate Clustering

Adrien Bougouin Akiko Aizawa

NII, Tokyo (JAPAN)

3 October 2010

Introduction

Problem statement

Keyphrases

- Word or multi-word expressions
- **Overview** of the content of a document

Applications

- Document indexing
- Document clustering
- Text summarization
- Query expansion
- Targeted advertising
- etc.

But. . .

Many documents do not have associated keyphrases.

Introduction

Problem statement

Keyphrases

- Word or multi-word expressions
- **Overview** of the content of a document

Applications

- Document indexing
- Document clustering
- Text summarization
- Query expansion
- Targeted advertising
- etc.

But...

Many documents do not have associated keyphrases.

Introduction

Problem statement

Keyphrases

- Word or multi-word expressions
- **Overview** of the content of a document

Applications

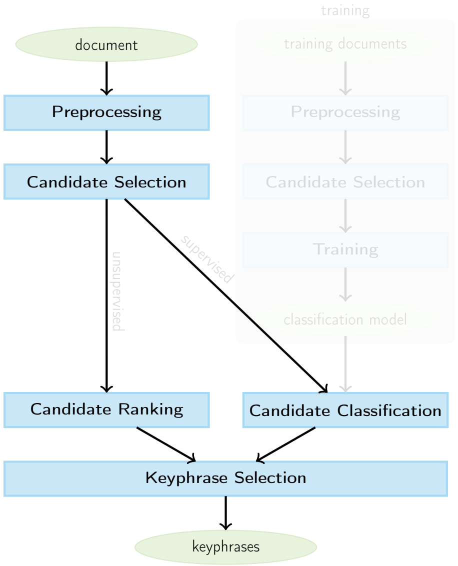
- Document indexing
- Document clustering
- Text summarization
- Query expansion
- Targeted advertising
- etc.

But...

Many documents do not have associated keyphrases.

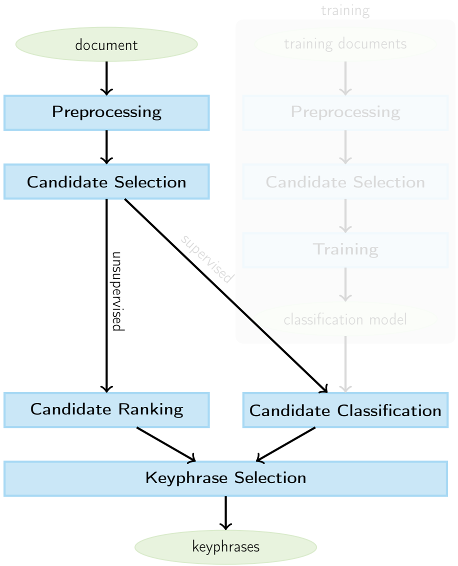
Introduction

Automatic keyphrase extraction



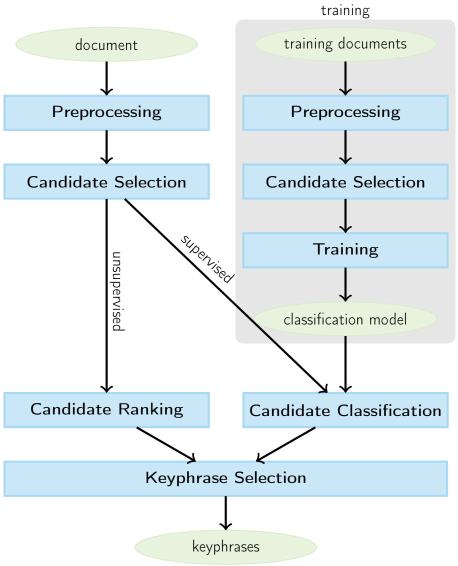
Introduction

Automatic keyphrase extraction



Introduction

Automatic keyphrase extraction



Introduction

Our work

Combining unsupervised and supervised methods:

1. Clustering candidate keyphrases into topics
2. Ranking topics **in an unsupervised way**
3. Selecting keyphrases from the best topics **in a supervised way**

Outline

1. State-of-the-art
2. Supervised TopicRank
3. Evaluation
4. Conclusion

Outline

State-of-the-art

Unsupervised methods

Mostly ranking techniques that use:

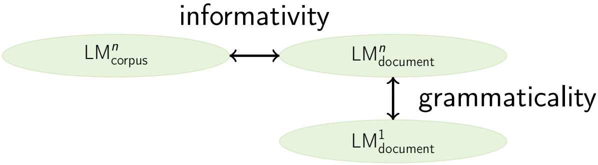
- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measures
  - refined with similar documents
  - biased with topic probabilities
  - modified to rank topics instead of words

State-of-the-art

Unsupervised methods

Mostly ranking techniques that use:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measures
  - refined with similar documents
  - biased with topic probabilities
  - modified to rank topics instead of words



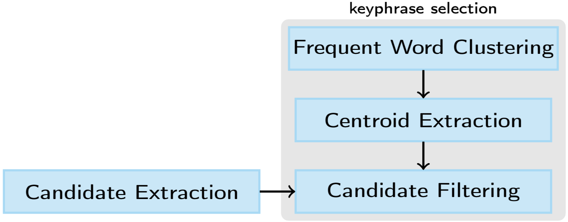
(Tomokiyo and Hurst, 2003)

State-of-the-art

Unsupervised methods

Mostly ranking techniques that use:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measures
  - refined with similar documents
  - biased with topic probabilities
  - modified to rank topics instead of words



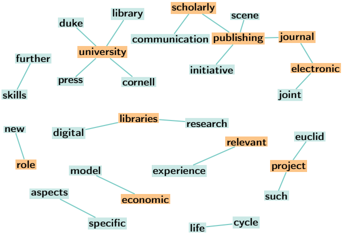
(Liu et al., 2009)

State-of-the-art

Unsupervised methods

Mostly ranking techniques that use:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measures
  - refined with similar documents
  - biased with topic probabilities
  - modified to rank topics instead of words



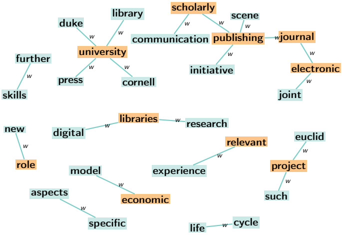
(Mihalcea and Tarau, 2004)

State-of-the-art

Unsupervised methods

Mostly ranking techniques that use:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measures
  - refined with similar documents
  - biased with topic probabilities
  - modified to rank topics instead of words



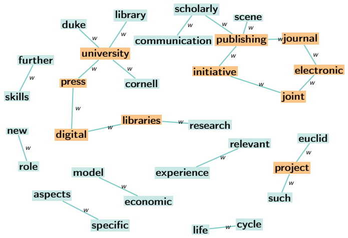
(Wan and Xiao, 2008; Tsatsaronis et al., 2010)

State-of-the-art

Unsupervised methods

Mostly ranking techniques that use:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measures
  - refined with similar documents
  - biased with topic probabilities
  - modified to rank topics instead of words



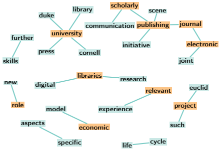
(Wan and Xiao, 2008)

State-of-the-art

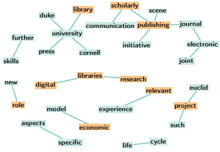
Unsupervised methods

Mostly ranking techniques that use:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measures
  - refined with similar documents
  - biased with topic probabilities
  - modified to rank topics instead of words

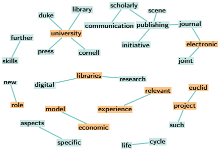


topic 1



topic 2

. . .



topic n

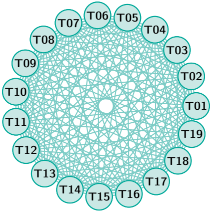
(Liu et al., 2010)

State-of-the-art

Unsupervised methods

Mostly ranking techniques that use:

- language models
- clusters
- or **graphs** of word co-occurrences
  - weighted with co-occurrence number or semantic measures
  - refined with similar documents
  - biased with topic probabilities
  - modified to rank topics instead of words



(Bougouin et al., 2013, TopicRank)

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (TF)
- Inverse document frequency (IDF)
- $TF \times IDF$
- Generic sections

$\Rightarrow$ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (*TF*)
- Inverse document frequency (*IDF*)
- $TF \times IDF$
- Generic sections

⇒ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (*TF*)
- Inverse document frequency (*IDF*)
- *TF* $\times$ *IDF*
- Generic sections

$\Rightarrow$ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (*TF*)
- Inverse document frequency (*IDF*)
- $TF \times IDF$
- Generic sections

⇒ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (TF)
- Inverse document frequency (IDF)
- $TF \times IDF$
- Generic sections

⇒ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (*TF*)
- Inverse document frequency (*IDF*)
- $TF \times IDF$
- Generic sections

⇒ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (*TF*)
- Inverse document frequency (*IDF*)
- $TF \times IDF$
- Generic sections

⇒ Currently the best performing methods

State-of-the-art
Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (*TF*)
- Inverse document frequency (*IDF*)
- *TF* × *IDF*
- Generic sections

⇒ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (TF)
- Inverse document frequency (IDF)
- $TF \times IDF$
- Generic sections

⇒ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (*TF*)
- Inverse document frequency (*IDF*)
- *TF* $\times$ *IDF*
- Generic sections

⇒ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (TF)
- Inverse document frequency (IDF)
- $TF \times IDF$
- Generic sections

⇒ Currently the best performing methods

State-of-the-art
Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (TF)
- Inverse document frequency (IDF)
- $TF \times IDF$
- Generic sections

⇒ Currently the best performing methods

State-of-the-art

Supervised methods

Train various classifiers:

- Naive Bayes (Witten et al., 1999)
- MaxEnt (Sujian et al., 2003)
- Support Vector Machines (SVMs) (Zhang et al., 2006)
- Decision trees (Ercan and Cicekli, 2007)
- Multilayer perceptrons (Sarkar et al., 2010)

with many different features:

- Length
- First position
- Part-of-Speeches
- Frequency (TF)
- Inverse document frequency (IDF)
- $TF \times IDF$
- Generic sections

$\Rightarrow$ Currently the best performing methods

State-of-the-art

Two categories / Two visions

Unsupervised vision

How important is a given phrase regarding the others?
$\Rightarrow$ extract the most important phrases

Supervised vision

How does a given phrase fit the keyphrase caracteristics in a global context?
$\Rightarrow$ extract the phrases most likely to be keyphrases

Then...

Why not combining both supervised and unsupervised approaches?

State-of-the-art

Two categories / Two visions

Unsupervised vision

How important is a given phrase regarding the others?
$\Rightarrow$ extract the most important phrases

Supervised vision

How does a given phrase fit the keyphrase caracteristics in a global context?
$\Rightarrow$ extract the phrases most likely to be keyphrases

Then...

Why not combining both supervised and unsupervised approaches?

State-of-the-art

Two categories / Two visions

Unsupervised vision

How important is a given phrase regarding the others?
$\Rightarrow$ extract the most important phrases

Supervised vision

How does a given phrase fit the keyphrase caracteristics in a global context?
$\Rightarrow$ extract the phrases most likely to be keyphrases

Then...

Why not combining both supervised and unsupervised approaches?

Outline

Supervised TopicRank

TopicRank: brief overview

1. Select the `(<NOUN>|<ADJ>)+` as candidates
2. Cluster candidates that "belong to the same topic"
   - stem overlap similarity
3. Build a complete graph of topics
   - edges weighted by a sementic strength
4. Apply PageRank's "voting concept"
   - Important topics contribute more to the importance of the topics they are strongly connected to
5. Extract keyphrases from the most important topics
   - one keyphrase per topic → the first candidate in the document

Supervised TopicRank

TopicRank: brief overview

1. Select the `(<NOUN>|<ADJ>)+` as candidates
2. Cluster candidates that "belong to the same topic"
   - stem overlap similarity
3. Build a complete graph of topics
   - edges weighted by a sementic strength
4. Apply PageRank's "voting concept"
   - Important topics contribute more to the importance of the topics they are strongly connected to
5. Extract keyphrases from the most important topics
   - one keyphrase per topic → the first candidate in the document

Supervised TopicRank

TopicRank: brief overview

1. Select the `(<NOUN>|<ADJ>)+` as candidates
2. Cluster candidates that "belong to the same topic"
   - stem overlap similarity
3. Build a complete graph of topics
   - edges weighted by a sementic strength
4. Apply PageRank's "voting concept"
   - Important topics contribute more to the importance of the topics they are strongly connected to
5. Extract keyphrases from the most important topics
   - one keyphrase per topic → the first candidate in the document

1. Select the `(<NOUN>|<ADJ>)+` as candidates
2. Cluster candidates that "belong to the same topic"
   - stem overlap similarity
3. Build a complete graph of topics
   - edges weighted by a sementic strength
4. Apply PageRank's "voting concept"
   - Important topics contribute more to the importance of the topics they are strongly connected to
5. Extract keyphrases from the most important topics
   - one keyphrase per topic $\rightarrow$ the first candidate in the document

Supervised TopicRank

TopicRank: brief overview

1. Select the `(<NOUN>|<ADJ>)+` as candidates
2. Cluster candidates that "belong to the same topic"
   - stem overlap similarity
3. Build a complete graph of topics
   - edges weighted by a sementic strength
4. Apply PageRank's "voting concept"
   - Important topics contribute more to the importance of the topics they are strongly connected to
5. Extract keyphrases from the most important topics
   - one keyphrase per topic $\rightarrow$ the first candidate in the document

Supervised TopicRank

Motivations

TopicRank's results are encouraging:

- Significant improvement over state-of-the-art graph-based methods
- Possible improvement from 12.1% to 30.3% of f-score
  - find a better strategy to identify the keyphrase of a given topic

A new vision (???):

- Combination of local importance and global likelyhood
- Bigger granularity for the importance: topic
- Superised classification on smaller sets: topics

Supervised TopicRank
Motivations

TopicRank's results are encouraging:

- Significant improvement over state-of-the-art graph-based methods
- Possible improvement from 12.1% to 30.3% of f-score
  - find a better strategy to identify the keyphrase of a given topic

A new vision (???):

- Combination of local importance and global likelyhood
- Bigger granularity for the importance: topic
- Superised classification on smaller sets: topics

Supervised TopicRank

Motivations

TopicRank's results are encouraging:

- Significant improvement over state-of-the-art graph-based methods
- Possible improvement from 12.1% to 30.3% of f-score
  - find a better strategy to identify the keyphrase of a given topic

A new vision (???):

- Combination of local importance and global likelyhood
- Bigger granularity for the importance: topic
- Superised classification on smaller sets: topics

Supervised TopicRank

Motivations

TopicRank's results are encouraging:

- Significant improvement over state-of-the-art graph-based methods
- Possible improvement from 12.1% to 30.3% of f-score
  - find a better strategy to identify the keyphrase of a given topic

A new vision (???):

- Combination of local importance and global likelyhood
- Bigger granularity for the importance: topic
- Superised classification on smaller sets: topics

Supervised TopicRank

Motivations

TopicRank's results are encouraging:

- Significant improvement over state-of-the-art graph-based methods
- Possible improvement from 12.1% to 30.3% of f-score
    - find a better strategy to identify the keyphrase of a given topic

A new vision (???):

- Combination of local importance and global likelyhood
- Bigger granularity for the importance: topic
- Superised classification on smaller sets: topics

Supervised TopicRank

Motivations

TopicRank's results are encouraging:

- Significant improvement over state-of-the-art graph-based methods
- Possible improvement from 12.1% to 30.3% of f-score
  - find a better strategy to identify the keyphrase of a given topic

A new vision (???):

- Combination of local importance and global likelyhood
- Bigger granularity for the importance: topic
- Superised classification on smaller sets: topics

Supervised TopicRank

Motivations

TopicRank's results are encouraging:

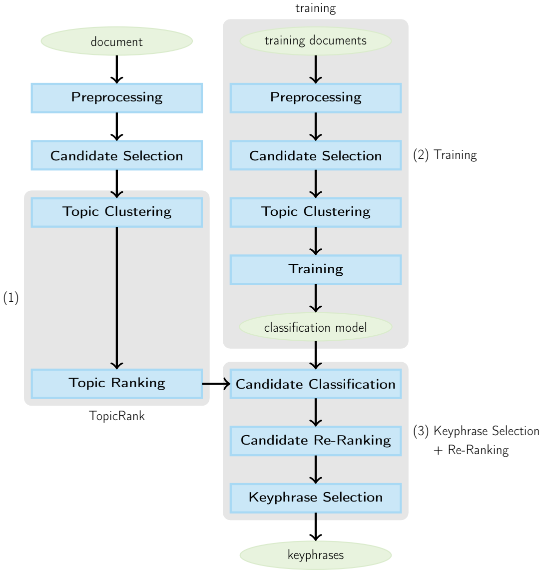
- Significant improvement over state-of-the-art graph-based methods
- Possible improvement from 12.1% to 30.3% of f-score
    - find a better strategy to identify the keyphrase of a given topic

A new vision (???):

- Combination of local importance and global likelyhood
- Bigger granularity for the importance: topic
- Superised classification on smaller sets: topics

Supervised TopicRank

Methodology



Supervised TopicRank

Training

Classifier

SVM:

- Learns a separating hyperplane between positive and negative examples
- Supports a large number of features
- Does not consider each feature to be independent

Samples

Only relevant clusters
⇒ clusters where a discrimination can actually be done

Features

Two categories of features:

- Topically independent features
- Topically dependent features

Supervised TopicRank

Training

Classifier

SVM:

- Learns a separating hyperplane between positive and negative examples
- Supports a large number of features
- Does not consider each feature to be independent

Samples

Only relevant clusters
⇒ clusters where a discrimination can actually be done

Features

Two categories of features:

- Topically independent features
- Topically dependent features

Supervised TopicRank

Training

Classifier

SVM:

- Learns a separating hyperplane between positive and negative examples
- Supports a large number of features
- Does not consider each feature to be independent

Samples

Only relevant clusters
⇒ clusters where a discrimination can actually be done

Features

Two categories of features:

- Topically independent features
- Topically dependent features

Supervised TopicRank

Training (next)

Topically independent features

- Lenght
- Structural features:
  - First position
  - In the 1st third?
  - In the 2nd third?
  - In the 3rd third?
- Distributional features:
  - TF-IDF
  - GDC (phraseness)

    $$\text{GDC}(c, d) = \frac{|c| \times \textbf{count}(c, d) \times \textbf{log}_{\textbf{10}}\textbf{count}(c, d)}{\sum_{w \in c} \textbf{count}(w, d)}$$
  - Keyphrase frequency (keyphraseness)

Topically dependent features

- Average stem overlap similarity
- Average number of completely disimilar candidates

Supervised TopicRank

Training (next)

Topically independent features

- Lenght
- Structural features:
  - First position
  - In the 1^{st} third?
  - In the 2^{nd} third?
  - In the 3^{rd} third?
- Distributional features:
  - TF-IDF
  - GDC (phraseness)

    $$\mathrm{GDC}(c, d) = \frac{|c| \times \mathbf{count}(c,d) \times \mathbf{log_{10}}\mathbf{count}(c,d)}{\sum_{w \in c} \mathbf{count}(w,d)}$$
  - Keyphrase frequency (keyphraseness)

Topically dependent features

- Average stem overlap similarity
- Average number of completely disimilar candidates

Supervised TopicRank

Training (next)

Topically independent features

- Lenght
- Structural features:
  - First position
  - In the 1st third?
  - In the 2nd third?
  - In the 3rd third?
- Distributional features:
  - TF-IDF
  - GDC (phraseness)

    $\mathrm{GDC}(c, d) = \frac{|c| \times \mathbf{count}(c,d) \times \mathbf{log_{10}} \mathbf{count}(c,d)}{\sum_{w \in c} \mathbf{count}(w,d)}$
  - Keyphrase frequency (keyphraseness)

Topically dependent features

- Average stem overlap similarity
- Average number of completely disimilar candidates

Supervised TopicRank

Training (next)

Topically independent features

- Lenght
- Structural features:
  - First position
  - In the 1^{st} third?
  - In the 2^{nd} third?
  - In the 3^{rd} third?
- Distributional features:
  - TF-IDF
  - GDC (phraseness)

  $$\text{GDC}(c, d) = \frac{|c| \times \textbf{count}(c,d) \times \textbf{log}_{\mathbf{10}}\textbf{count}(c,d)}{\sum_{w \in c} \textbf{count}(w,d)}$$

  - Keyphrase frequency (keyphraseness)

Topically dependent features

- Average stem overlap similarity
- Average number of completely disimilar candidates

Supervised TopicRank

Training (next)

Topically independent features

- Lenght
- Structural features:
  - First position
  - In the 1st third?
  - In the 2nd third?
  - In the 3rd third?
- Distributional features:
  - TF-IDF
  - GDC (phraseness)

$$\text{GDC}(c, d) = \frac{|c| \times \textbf{count}(c,d) \times \textbf{log}_{\textbf{10}}\textbf{count}(c,d)}{\sum_{w \in c} \textbf{count}(w,d)}$$

  - Keyphrase frequency (keyphraseness)

Topically dependent features

- Average stem overlap similarity
- Average number of completely disimilar candidates

Supervised TopicRank

Keyphrase selection

- Apply the SVM classifier
- For each topic:
  - Select the candidate with the best confidence

Supervised TopicRank

Keyphrase re-ranking

Formerly, TopicRank ranks keyphrases by their topic's importance, but the topic ranking is not perfect.

$\Rightarrow$ we combine the TopicRank score to the probability that the keyphrase is actually a keyphrase

$$S(c) = \alpha \times \text{topicrank}(c) + (1 - \alpha) \times p(c)$$

$\alpha = 0.75 \Rightarrow$ more importance is given to the unsupervised topic ranking

Supervised TopicRank

Keyphrase re-ranking

Formerly, TopicRank ranks keyphrases by their topic's importance, but the topic ranking is not perfect.
$\Rightarrow$ we combine the TopicRank score to the probability that the keyphrase is actually a keyphrase

$$S(c) = \alpha \times \text{topicrank}(c) + (1 - \alpha) \times p(c)$$

$\alpha = 0.75 \Rightarrow$ more importance is given to the unsupervised topic ranking

Supervised TopicRank

Keyphrase re-ranking

Formerly, TopicRank ranks keyphrases by their topic's importance, but the topic ranking is not perfect.
$\Rightarrow$ we combine the TopicRank score to the probability that the keyphrase is actually a keyphrase

$$S(c) = \alpha \times \text{topicrank}(c) + (1 - \alpha) \times p(c)$$

$\alpha = 0.75 \Rightarrow$ more importance is given to the unsupervised topic ranking

Outline

Evaluation

Dataset

SemEval

- 244 scientific papers
  - 144 training documents
  - 100 test documents
- author- and reader-assigned keyphrases

Evaluation

Baselines

Derived baselines

Method	Features		
	Independent	Dependent	All
TopicRank+SVM	✓	✓	our method
Clustering+SVM	✓	✓	✓
SVM	✓	✗	✗

Classic baselines

- KEA
  - Naive Bayes
  - Two features: first position et TF-IDF
- TF-IDF
- TopicRank

Evaluation

Baselines

Derived baselines

Method	Features		
	Independent	Dependent	All
TopicRank+SVM	✓	✓	our method
Clustering+SVM	✓	✓	✓
SVM	✓	✗	✗

Classic baselines

- KEA
  - Naive Bayes
  - Two features: first position et TF-IDF
- TF-IDF
- TopicRank

Evaluation

Measures

- Cut-off at 10 keyphrases
- Precision
- Recall
- F-score

$$\text{f-score} = (1 + \beta^2) \times \frac{\text{precision} \times \text{recall}}{(\beta^2 \times \textit{precision}) + \textit{recall}}$$

$$\beta = 1$$

- Problem of dealing with gold standard

$\Rightarrow$ Stemmed form comparisons

Evaluation

Results

Method	Features								
	Independent			Dependent			All		
	P	R	F	P	R	F	P	R	F
TopicRank+SVM	21.5	15.1	17.6	9.0	6.5	7.5	24.2	16.7	19.6
Clustering+SVM	13.3	9.3	10.8	0.2	0.1	0.2	11.9	8.4	9.7
SVM	15.0	10.5	12.2						

- Low performance of dependent features
- Best performance overall derived baselines

Evaluation

Results

Method	Features								
	Independent			Dependent			All		
	P	R	F	P	R	F	P	R	F
TopicRank+SVM	21.5	15.1	17.6	9.0	6.5	7.5	24.2	16.7	19.6
Clustering+SVM	13.3	9.3	10.8	0.2	0.1	0.2	11.9	8.4	9.7
SVM	15.0	10.5	12.2						

- Low performance of dependent features
- Best performance overall derived baselines

Evaluation

Results (next)

Method	P	R	F
KEA	18.8	13.3	15.4
TF-IDF	13.2	8.9	10.5
TopicRank	14.9	10.3	12.1
TopicRank+SVM	24.2	16.7	19.6
$\text{TopicRank}_{\text{max}}$	37.6	25.8	30.3

- Best performance over state-of-the-art methods
- The dream is not fulfilled...

Evaluation

Results (next)

Method	P	R	F
KEA	18.8	13.3	15.4
TF-IDF	13.2	8.9	10.5
TopicRank	14.9	10.3	12.1
TopicRank+SVM	24.2	16.7	19.6
$\text{TopicRank}_{\text{max}}$	37.6	25.8	30.3

- Best performance over state-of-the-art methods
- The dream is not fulfilled…

Outline

Conclusion

- Extension of TopicRank
- Supervised keyphrase selection with TopicRank's best topics
- Results show improvement over TopicRank
- There is still a huge gap between the current performance and the best possible ones

Perspectives

- Apply topic modeling to improve TopicRank's topic clustering (LDA, etc.)
- Experiment with topic labelling methods proposed for LDA

Thank you

References

Adrien Bougouin, Florian Boudin, and Béatrice Daille. Topicrank: Graph-Based Topic Ranking for Keyphrase Extraction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 543–551, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I13-1062.

Gonenc Ercan and Ilyas Cicekli. Using Lexical Chains for Keyword Extraction. *Information Processing and Management*, 43(6):1705–1714, nov 2007. ISSN 0306-4573. URL http://dx.doi.org/10.1016/j.ipm.2007.01.015.

Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. Clustering to Find Exemplar Terms for Keyphrase Extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 (EMNLP)*, pages 257–266, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://dl.acm.org/citation.cfm?id=1699510.1699544.

References

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. Automatic Keyphrase Extraction Via Topic Decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pages 366–376, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1870658.1870694.

Rada Mihalcea and Paul Tarau. TextRank: Bringing Order Into Texts. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Kamal Sarkar, Mita Nasipuri, and Suranjan Ghose. A New Approach to Keyphrase Extraction Using Neural Networks. *International Journal of Computer Science Issues Publicity Board 2010*, 2010.

References

Li Sujian, Wang Houfeng, Yu Shiwen, and Xin Chengsheng. News-Oriented Keyword Indexing with Maximum Entropy Principle. In *Proceedings of the 17th Pacific Asia Conference*. COLIPS Publications, 2003.

Takashi Tomokiyo and Matthew Hurst. A Language Model Approach to Keyphrase Extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, pages 33–40, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. URL `http://dx.doi.org/10.3115/1119282.1119287`.

George Tsatsaronis, Iraklis Varlamis, and Kjetil Nørvåg. SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1074–1082, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1873781.1873902`.

References

Xiaojun Wan and Jianguo Xiao. Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, pages 855–860. AAAI Press, 2008. ISBN 978-1-57735-368-3. URL `http://dl.acm.org/citation.cfm?id=1620163.1620205`.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill Manning. KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of the 4th ACM Conference on Digital Libraries*, pages 254–255, New York, NY, USA, 1999. ACM. ISBN 1-58113-145-3. URL `http://doi.acm.org/10.1145/313238.313437`.

Kuo Zhang, Hui Xu, Jie Tang, and Juanzi Li. Keyword Extraction Using Support Vector Machine. In *Proceedings of the 7th International Conference on Advances in Web-Age Information Management*, pages 85–96, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-35225-2, 978-3-540-35225-9. URL `http://dx.doi.org/10.1007/11775300_8`.