Keyphrase Annotation with Graph Co-Ranking

Adrien Bougouin, Florian Boudin, Béatrice Daille

Université de Nantes, LINA

December 16th, 2016

How to infer the main content of a domain-specific document?

Keyphrases

- Single- or multi-word expressions
- Important topics/concepts
- Useful to multiple Information Retrieval tasks:
    - Document indexing
    - Text summarization
    - Query expansion
    - etc.

Introduction ⟩ Example

*Toucher : le tango des sens. Problèmes de **sémantique lexicale***

*À partir d'une hypothèse sur la sémantique de l'unité lexicale 'toucher' formulée en termes de forme schématique, cette étude vise à rendre compte de la **variation sémantique** manifestée par les emplois de ce **verbe** dans la construction **transitive** directe 'C0 toucher C1'. Notre étude cherche donc à articuler variation sémantique et invariance fonctionnelle. Cet article concerne essentiellement le mode de variation co-textuelle : en conséquence, elle ne constitue qu'une première étape dans la compréhension de la construction des valeurs référentielles que permet 'toucher'. Une étude minutieuse de nombreux exemples nous a permis de dégager des constantes impératives sous la forme des 4 notions suivantes : sous-détermination sémantique, contact, anormalité, et contingence. Nous avons tenté de montrer comment ces notions interprétatives sont directement dérivables de la forme schématique proposée.*

Reference keyphrases (French):

Français; *modélisation*; *analyse distributionnelle*; *interprétation sémantique*; ***variation sémantique***; ***transitif***; ***verbe***; syntaxe; ***sémantique lexicale***.

Reference keyphrases (English):

French; modelling; distributional analysis; semantic interpretation; **semantic variation**; **transitive**; **verb**; syntax; **lexical semantics**.

- Silence
- Domain consistency
- Free syntax *(e.g. syntax, semantic variation, transitive, etc.)*
- Risks of semantic redundancy (over generation)

Introduction ⟩ Difficulties

- **Silence** ⊙
- **Domain consistency** ⊙
- Free syntax *(e.g. syntax, semantic variation, transitive, etc.)*
- Risks of semantic redundancy (over generation)

Introduction 〉 Difficulties

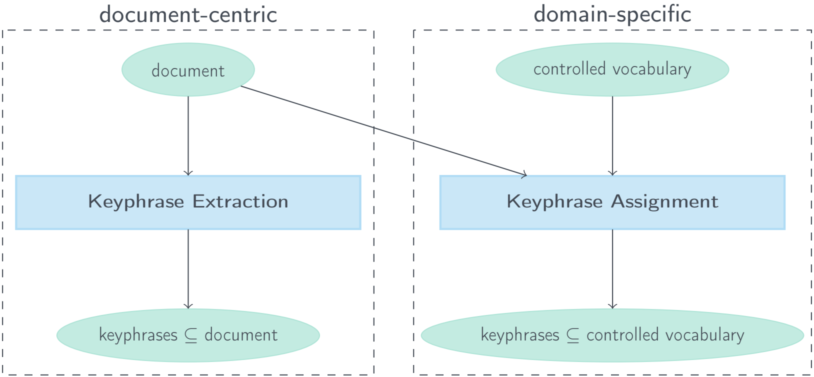
- **Silence** ⊙
- **Domain consistency** ⊙
- Free syntax *(e.g. syntax, semantic variation, transitive, etc.)* ✓*
- Risks of semantic redundancy/over generation ✓*

* (Bougouin et al., 2013, TopicRank)

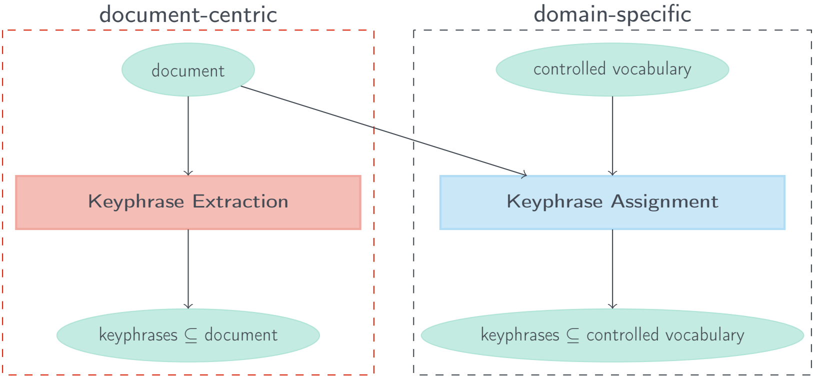
Outline

1. Related Work
2. Proposal
3. TopicCoRank
4. Evaluation
5. Conclusion

Related Work



Related Work



Supervised keyphrase extraction

- Relying on reference keyphrases
- Learning to determine keyphrase likelihood
- Mixing statistical and linguistic features

Unsupervised keyphrase extraction

- Looking for the most important keyphrase candidates
- Using mainly statistics
- Linking keyphrase candidates to each other

Related Work 〉 Keyphrase extraction (TopicRank)

Graph-based approach detecting documents most important topics and extracting keyphrases from these topics

1	Keyphrase candidate selection	`/(NOUN | ADJ)+/`
2	Keyphrase candidates topical clustering	*lexical clustering*
3	Topic graph construction	*complete graph*
4	Graph-based topic ranking	*Google's PageRank*
5	Keyphrase extraction from the important topics	*one per topic*

Related Work 〉 Keyphrase extraction (TopicRank)

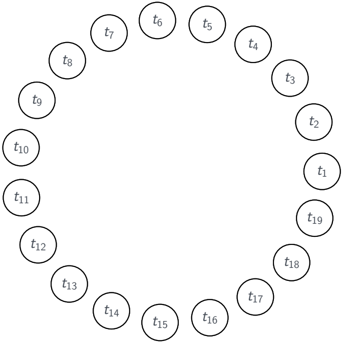
Graph-based approach detecting documents most important topics and extracting keyphrases from these topics

1	Keyphrase candidate selection	`/(NOUN | ADJ)+/`
2	Keyphrase candidates topical clustering	*lexical clustering*
3	Topic graph construction	*complete graph*
4	Graph-based topic ranking	*Google's PageRank*
5	Keyphrase extraction from the important topics	*one per topic*

Related Work 〉 Keyphrase extraction (TopicRank)

Graph-based approach detecting documents most important topics and extracting keyphrases from these topics

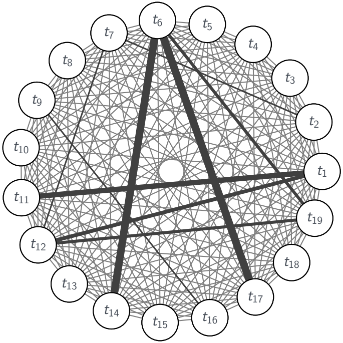
1. Keyphrase candidate selection `/(NOUN | ADJ)+/`
2. Keyphrase candidates topical clustering *lexical clustering*
3. Topic graph construction *complete graph*
4. Graph-based topic ranking *Google's PageRank*
5. Keyphrase extraction from the important topics *one per topic*



Related Work 〉 Keyphrase extraction (TopicRank)

Graph-based approach detecting documents most important topics and extracting keyphrases from these topics

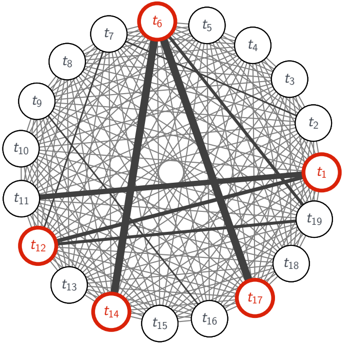
1	Keyphrase candidate selection	`/(NOUN | ADJ)+/`
2	Keyphrase candidates topical clustering	*lexical clustering*
3	Topic graph construction	*complete graph*
4	Graph-based topic ranking	*Google's PageRank*
5	Keyphrase extraction from the important topics	*one per topic*



Related Work 〉 Keyphrase extraction (TopicRank)

Graph-based approach detecting documents most important topics and extracting keyphrases from these topics

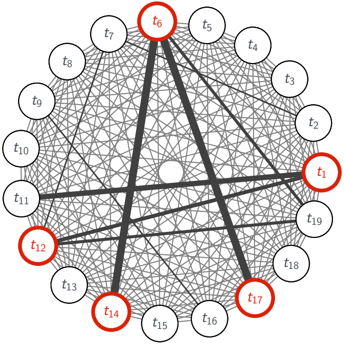
1	Keyphrase candidate selection	`/(NOUN | ADJ)+/`
2	Keyphrase candidates topical clustering	*lexical clustering*
3	Topic graph construction	*complete graph*
4	Graph-based topic ranking	*Google's PageRank*
5	Keyphrase extraction from the important topics	*one per topic*



Related Work 〉 Keyphrase extraction (TopicRank)

Graph-based approach detecting documents most important topics and extracting keyphrases from these topics

1 Keyphrase candidate selection	`/(NOUN | ADJ)+/`
2 Keyphrase candidates topical clustering	*lexical clustering*
3 Topic graph construction	*complete graph*
4 Graph-based topic ranking	*Google's PageRank*
5 Keyphrase extraction from the important topics	*one per topic*



Related Work ⟩ Keyphrase extraction (TopicRank – Example)

Toucher : le tango des sens. Problèmes de ***sémantique lexicale***

À partir d'une hypothèse sur la sémantique de l'unité lexicale 'toucher' formulée en termes de forme schématique, cette étude vise à rendre compte de la ***variation sémantique*** *manifestée par les emplois de ce* ***verbe*** *dans la construction* ***transitive*** *directe 'C0 toucher C1'. Notre étude cherche donc à articuler variation sémantique et invariance fonctionnelle. Cet article concerne essentiellement le mode de variation co-textuelle : en conséquence, elle ne constitue qu'une première étape dans la compréhension de la construction des valeurs référentielles que permet 'toucher'. Une étude minutieuse de nombreux exemples nous a permis de dégager des constantes impératives sous la forme des 4 notions suivantes : sous-détermination sémantique, contact, anormalité, et contingence. Nous avons tenté de montrer comment ces notions interprétatives sont directement dérivables de la forme schématique proposée.*

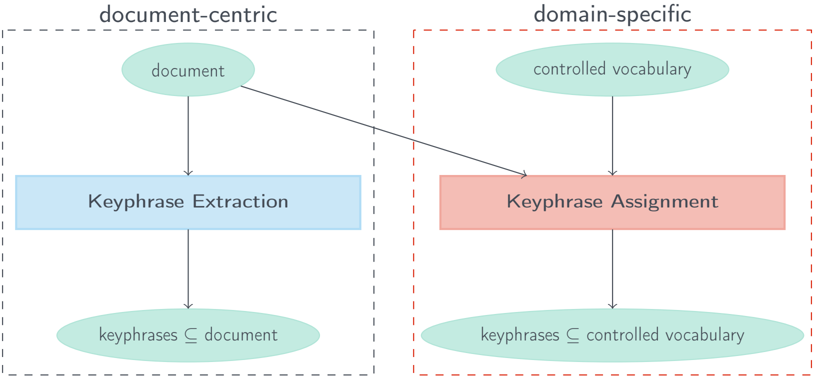
Reference keyphrases (French):

Français; *modélisation*; *analyse distributionnelle*; *interprétation sémantique*; ***variation sémantique***; ***transitif***; ***verbe***; *syntaxe*; ***sémantique lexicale***.

TopicRank 10 keyphrases (French):

Sémantique lexicale; *variation sémantique*; *problèmes*; *étude*; *forme schématique*; *sens*; *tango*; *invariance fonctionnelle*; *construction transitive directe*; *article*;

Related Work



Keyphrase assignment

- Relying on a domain-specific controlled vocabulary (thesaurus)
- Aiming for consistency across domain

Handful of attempts

- Relying on a domain-specific controlled vocabulary (thesaurus)
- Aiming for consistency across domain

Handful of attempts

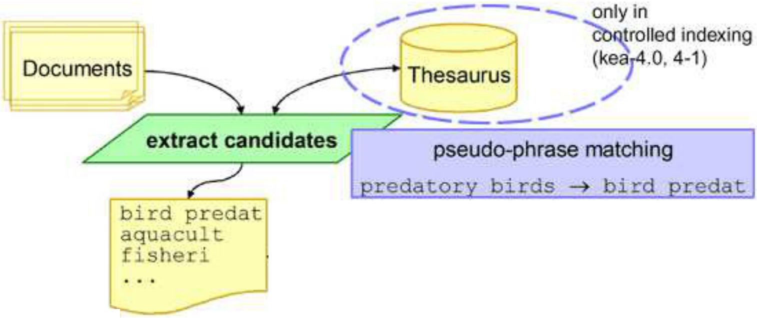


Figure from Medelyan and Witten (2006)

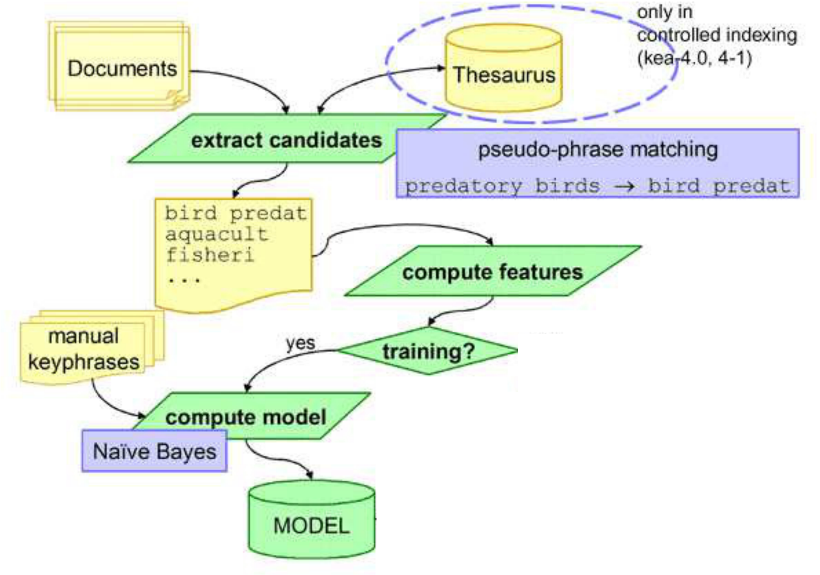


Figure from Medelyan and Witten (2006)

Related Work 〉 Keyphrase assignment (KEA++)

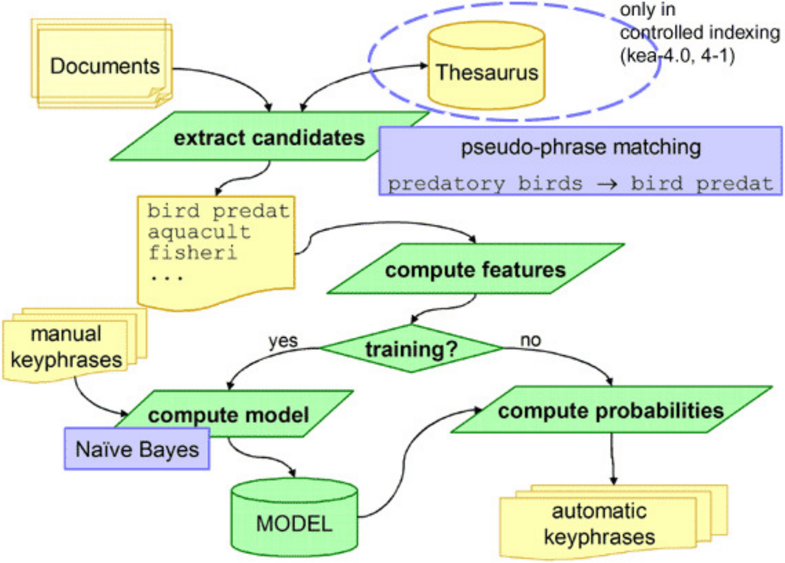


Figure from Medelyan and Witten (2006)

Related Work 〉 Keyphrase extraction (KEA++ – Example)

*Toucher : le tango des sens. Problèmes de **sémantique lexicale***

*À partir d'une hypothèse sur la sémantique de l'unité lexicale 'toucher' formulée en termes de forme schématique, cette étude vise à rendre compte de la **variation sémantique** manifestée par les emplois de ce **verbe** dans la construction **transitive** directe 'C0 toucher C1'. Notre étude cherche donc à articuler variation sémantique et invariance fonctionnelle. Cet article concerne essentiellement le mode de variation co-textuelle : en conséquence, elle ne constitue qu'une première étape dans la compréhension de la construction des valeurs référentielles que permet 'toucher'. Une étude minutieuse de nombreux exemples nous a permis de dégager des constantes impératives sous la forme des 4 notions suivantes : sous-détermination sémantique, contact, anormalité, et contingence. Nous avons tenté de montrer comment ces notions interprétatives sont directement dérivables de la forme schématique proposée.*

Reference keyphrases (French):

Français; *modélisation*; *analyse distributionnelle*; *interprétation sémantique*; ***variation sémantique***; ***transitif***; ***verbe***; *syntaxe*; ***sémantique lexicale***.

Kea++ 10 keyphrases (French):

Toucher; *variation sémantique*; *sémantique*; *tangoa*; *direction*; *formant*; *sémantique lexicale*; *impératif*; *invariant sémantique*; *transitif*;

Related Work

Extraction	**Vs.**	**Assignment**
Leveraging document content	Vs.	Ignoring inner-document relations
Ignoring domain vocabulary	Vs.	Mapping document to its domain
Limited to document content	Vs.	Limited to vocabulary coverage

Professional indexer point of view

Annotated keyphrases should respect the vocabulary of the domain as much as possible, but should not be restricted to it in order to be exhaustive.

Related Work

Extraction	**Vs.**	**Assignment**
Leveraging document content	Vs.	Ignoring inner-document relations
Ignoring domain vocabulary	Vs.	Mapping document to its domain
Limited to document content	Vs.	Limited to vocabulary coverage

Professional indexer point of view

Annotated keyphrases should respect the vocabulary of the domain as much as possible, but should not be restricted to it in order to be exhaustive.

Outline

1. Related Work

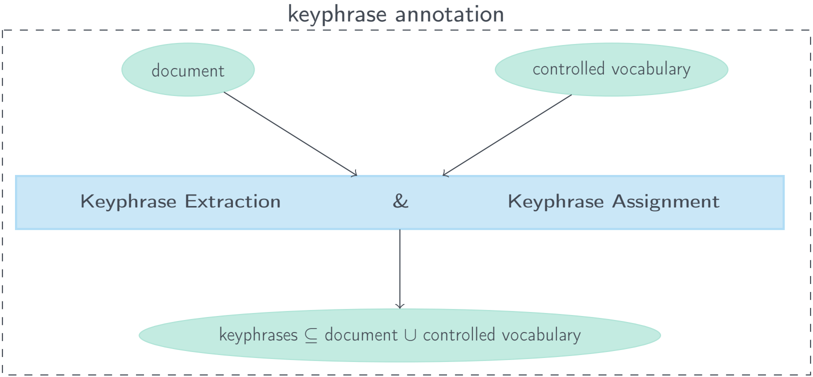
2. Proposal

3. TopicCoRank

4. Evaluation

5. Conclusion

Proposal


keyphrase annotation
document
controlled vocabulary
Keyphrase Extraction
&
Keyphrase Assignment
keyphrases ⊆ document ∪ controlled vocabulary

Outline

1. Related Work

2. Proposal

3. TopicCoRank

4. Evaluation

5. Conclusion

TopicCoRank

Supervised extension of TopicRank to add assignment capabilities alongside extraction

Hypothesis

- The domain supplements the document

⇒ improves keyphrase extraction

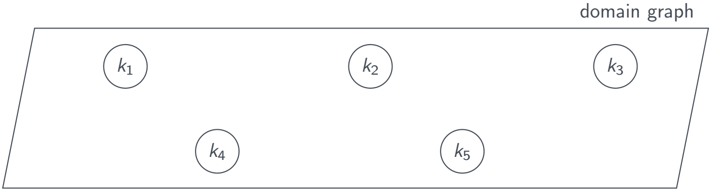
- The training documents represent the domain

⇒ keyphrases circumvents use of controlled vocabulary

TopicCoRank 〉 Domain representation

Undirected graph of keyphrases annotated to training documents

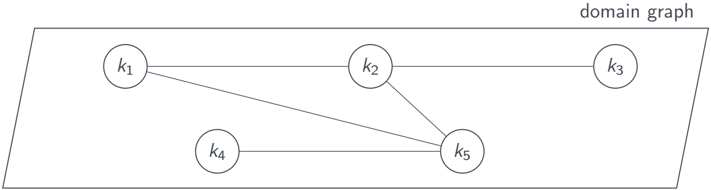
- Each keyphrase is represented by a vertex
- Keyphrases of the same documents are connected
- Edges are weighted by the number of documents containing both keyphrases



Domain representation

Undirected graph of keyphrases annotated to training documents

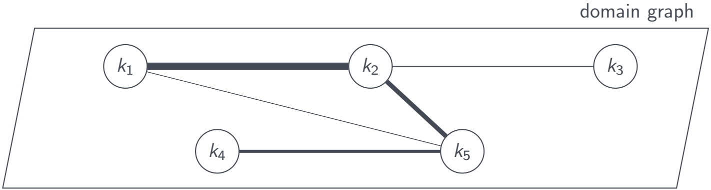
- Each keyphrase is represented by a vertex
- Keyphrases of the same documents are connected
- Edges are weighted by the number of documents containing both keyphrases



Domain representation

Undirected graph of keyphrases annotated to training documents

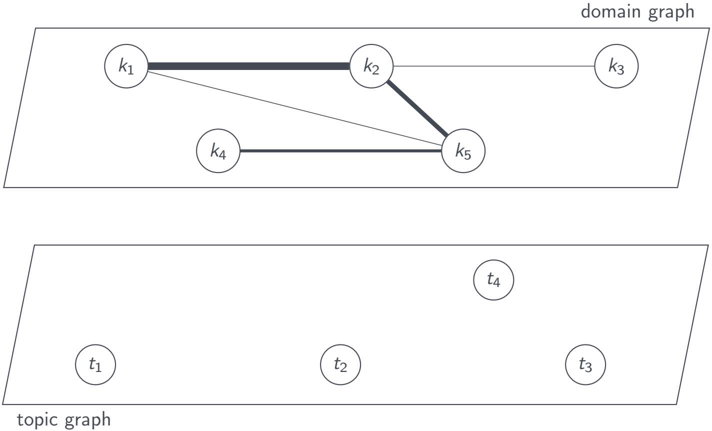
- Each keyphrase is represented by a vertex
- Keyphrases of the same documents are connected
- Edges are weighted by the number of documents containing both keyphrases



TopicCoRank 〉 Integration with TopicRank

Domain graph unified to TopicRank's topic graph

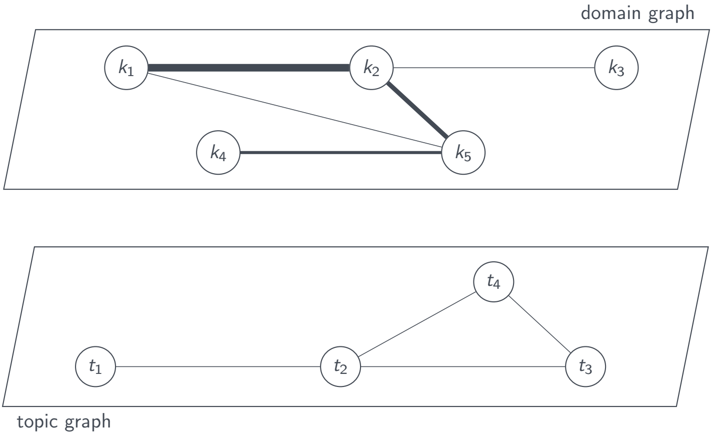
- Each topic is represented by a vertex
- Topics referred to within the same sentences are connected
- Edges are weighted by the number of sentences containing both topics



Integration with TopicRank

Domain graph unified to TopicRank's topic graph

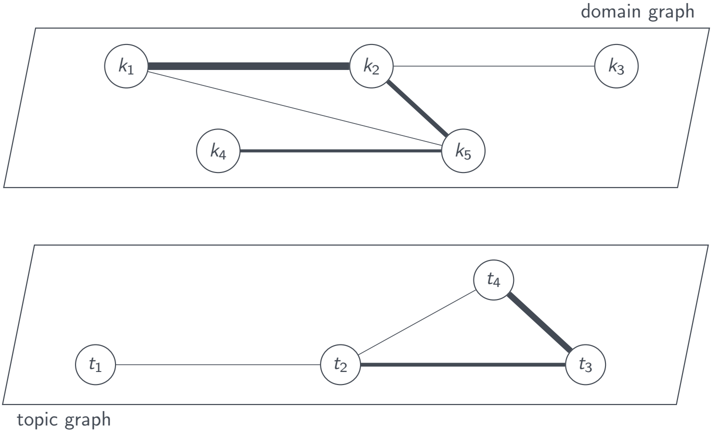
- Each topic is represented by a vertex
- Topics referred to within the same sentences are connected
- Edges are weighted by the number of sentences containing both topics



TopicCoRank 〉 Integration with TopicRank

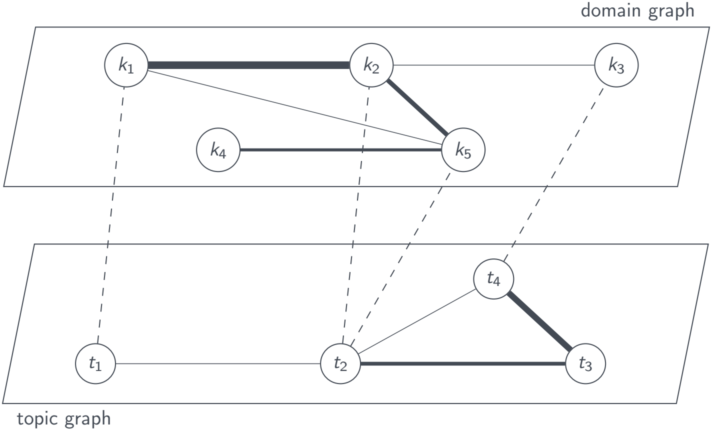
Domain graph unified to TopicRank's topic graph

- Each topic is represented by a vertex
- Topics referred to within the same sentences are connected
- Edges are weighted by the number of sentences containing both topics



Domain graph unified to TopicRank's topic graph

- Each topic is represented by a vertex
- Topics referred to within the same sentences are connected
- Edges are weighted by the number of sentences containing both topics



- Domain keyphrases k_i are as much important as they are strongly connected to as much other important keyphrases k_j as possible

$$R_{in}(k_i) = \sum_{k_j \in E_{\text{in}}(k_i)} \frac{w_{ij} S(k_j)}{\sum_{k_k \in E_{\text{in}}(k_j)} w_{jk}}$$

- Document topics t_i are as much important as they are strongly connected to as much other important topics t_j as possible

$$R_{in}(t_i) = \sum_{t_j \in E_{\text{in}}(t_i)} \frac{w_{ij} S(t_j)}{\sum_{t_k \in E_{\text{in}}(t_j)} w_{jk}}$$

Graph-based co-ranking I

- Domain keyphrases k_i are as much important as they are strongly connected to as much other important keyphrases k_j as possible

$$R_{in}(k_i) = \sum_{k_j \in E_{\text{in}}(k_i)} \frac{w_{ij} S(k_j)}{\sum_{k_k \in E_{\text{in}}(k_j)} w_{jk}}$$

- Document topics t_i are as much important as they are strongly connected to as much other important topics t_j as possible

$$R_{in}(t_i) = \sum_{t_j \in E_{\text{in}}(t_i)} \frac{w_{ij} S(t_j)}{\sum_{t_k \in E_{\text{in}}(t_j)} w_{jk}}$$

- Domain keyphrases and document topics v_i gain importance from each other

$$R_{out}(v_i) = \sum_{v_j \in E_{\text{out}}(v_i)} \frac{S(v_j)}{|E_{out}(v_j)|}$$

- Both inner- and outer- recommendation are combined with empirically tuned damping factors

$$S(k_i) = (1 - \lambda_k)\ R_{out}(k_i) + \lambda_k\ R_{in}(k_i)$$

$$S(t_i) = (1 - \lambda_t)\ R_{out}(t_i) + \lambda_t\ R_{in}(t_i)$$

Graph-based co-ranking II

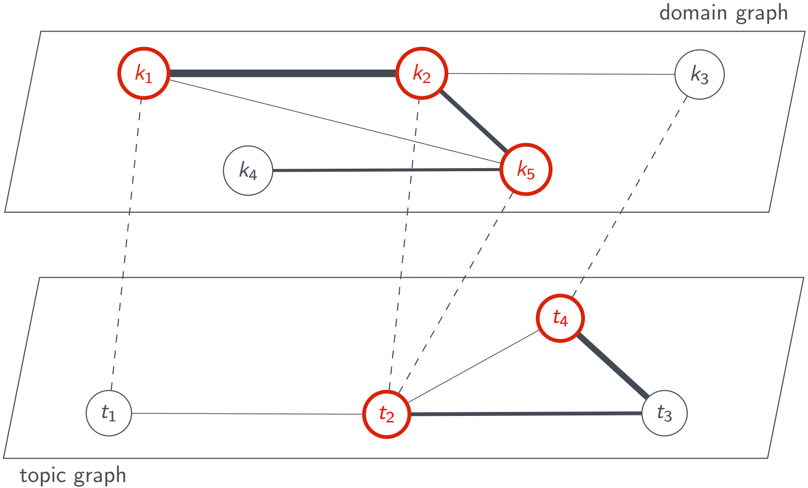
- Domain keyphrases and document topics v_i gain importance from each other

$$R_{out}(v_i) = \sum_{v_j \in E_{\text{out}}(v_i)} \frac{S(v_j)}{|E_{out}(v_j)|}$$

- Both inner- and outer- recommendation are combined with empirically tuned damping factors

$$S(k_i) = (1 - \lambda_k)\ R_{out}(k_i) + \lambda_k\ R_{in}(k_i)$$

$$S(t_i) = (1 - \lambda_t)\ R_{out}(t_i) + \lambda_t\ R_{in}(t_i)$$


domain graph
k_1
k_2
k_3
k_4
k_5
t_4
t_1
t_2
t_3
topic graph

Toucher : le tango des sens. Problèmes de ***sémantique lexicale***

À partir d'une hypothèse sur la sémantique de l'unité lexicale 'toucher' formulée en termes de forme schématique, cette étude vise à rendre compte de la ***variation sémantique*** *manifestée par les emplois de ce* ***verbe*** *dans la construction* ***transitive*** *directe 'C0 toucher C1'. Notre étude cherche donc à articuler variation sémantique et invariance fonctionnelle. Cet article concerne essentiellement le mode de variation co-textuelle : en conséquence, elle ne constitue qu'une première étape dans la compréhension de la construction des valeurs référentielles que permet 'toucher'. Une étude minutieuse de nombreux exemples nous a permis de dégager des constantes impératives sous la forme des 4 notions suivantes : sous-détermination sémantique, contact, anormalité, et contingence. Nous avons tenté de montrer comment ces notions interprétatives sont directement dérivables de la forme schématique proposée.*

Reference keyphrases (French):

Français; *modélisation*; *analyse distributionnelle*; *interprétation sémantique*; ***variation sémantique***; ***transitif***; ***verbe***; *syntaxe*; ***sémantique lexicale***.

TopicCoRank 10 keyphrases (French):

Sémantique lexicale; *sémantique*; *verbe*; *variation sémantique*; *français*; *hypothèse*; *syntaxe*; *pragmatique*; *interprétation sémantique*; *analyse distributionnelle*;

TopicCoRank

Benefits

- + Combines extraction and assignment: ↗ recall & ↗ precision
- + Circumvents the need of a controlled vocabulary

Outline

1. Related Work
2. Proposal
3. TopicCoRank
4. Evaluation
5. Conclusion

Evaluation 〉 Datasets

3 French corpora covering 3 domains of Humanities and Social Sciences

- Manually annotated by professional indexers
- Provided with controlled vocabularies
- Annotated based on both controlled vocabulary and content

Corpus	**Linguistics**			**Information Science**			**Archaeology**		
	train	⊃ dev	test	train	⊃ dev	test	train	⊃ dev	test
Doc.	515	100	200	506	100	200	518	100	200
Tokens/Doc.	161	151	147	105	152	157	221	201	214
Keyphrases	8.6	8.8	8.9	7.8	10.0	10.2	16.9	16.4	15.6
Silence (%)	60.6	63.2	62.8	67.9	63.1	66.9	37.0	48.4	37.4

Evaluation 〉 Datasets

3 French corpora covering 3 domains of Humanities and Social Sciences

- Manually annotated by professional indexers
- Provided with controlled vocabularies
- Annotated based on both controlled vocabulary and content

Corpus	Linguistics			Information Science			Archaeology		
	train	⊃ dev	test	train	⊃ dev	test	train	⊃ dev	test
Doc.	515	100	200	506	100	200	518	100	200
Tokens/Doc.	161	151	147	105	152	157	221	201	214
Keyphrases	8.6	8.8	8.9	7.8	10.0	10.2	16.9	16.4	15.6
Silence (%)	60.6	63.2	62.8	67.9	63.1	66.9	37.0	48.4	37.4

Evaluation 〉 Datasets

3 French corpora covering 3 domains of Humanities and Social Sciences

- Manually annotated by professional indexers
- Provided with controlled vocabularies
- Annotated based on both controlled vocabulary and content

Corpus	**Linguistics**			**Information Science**			**Archaeology**		
	train	⊃ dev	test	train	⊃ dev	test	train	⊃ dev	test
Doc.	515	100	200	506	100	200	518	100	200
Tokens/Doc.	161	151	147	105	152	157	221	201	214
Keyphrases	8.6	8.8	8.9	7.8	10.0	10.2	16.9	16.4	15.6
Silence (%)	60.6	63.2	62.8	67.9	63.1	66.9	37.0	48.4	37.4

3 French corpora covering 3 domains of Humanities and Social Sciences

- Manually annotated by professional indexers
- Provided with controlled vocabularies
- Annotated based on both controlled vocabulary and content

Corpus	Linguistics			Information Science			Archaeology		
	train	⊃ dev	test	train	⊃ dev	test	train	⊃ dev	test
Doc.	515	100	200	506	100	200	518	100	200
Tokens/Doc.	161	151	147	105	152	157	221	201	214
Keyphrases	8.6	8.8	8.9	7.8	10.0	10.2	16.9	16.4	15.6
Silence (%)	60.6	63.2	62.8	67.9	63.1	66.9	37.0	48.4	37.4

3 French corpora covering 3 domains of Humanities and Social Sciences

- Manually annotated by professional indexers
- Provided with controlled vocabularies
- Annotated based on both controlled vocabulary and content

Corpus	Linguistics			Information Science			Archaeology		
	train	⊃ dev	test	train	⊃ dev	test	train	⊃ dev	test
Doc.	515	100	200	506	100	200	518	100	200
Tokens/Doc.	161	151	147	105	152	157	221	201	214
Keyphrases	8.6	8.8	8.9	7.8	10.0	10.2	16.9	16.4	15.6
Silence (%)	60.6	63.2	62.8	67.9	63.1	66.9	37.0	48.4	37.4

Evaluation 〉 Baselines

Related work:

TopicRank	*Extraction*
Kea++	*Assignment*

TopicCoRank alternatives:

TopicCoRank$_{\text{extr}}$	*Extraction*
TopicCoRank$_{\text{assign}}$	*Assignment*

Evaluation 〉 Baselines

Related work:

- TopicRank *Extraction*
- Kea++ *Assignment*

TopicCoRank alternatives:

- TopicCoRank$_{\text{extr}}$ *Extraction*
- TopicCoRank$_{\text{assign}}$ *Assignment*

Evaluation 〉 Baselines

Related work:

TopicRank	*Extraction*
Kea++	*Assignment*

TopicCoRank alternatives:

TopicCoRank$_{extr}$	*Extraction*
TopicCoRank$_{assign}$	*Assignment*

Evaluation 〉 Baselines

Related work:

- TopicRank *Extraction*
- Kea++ *Assignment*

TopicCoRank alternatives:

- $\text{TopicCoRank}_{\text{extr}}$ *Extraction*
- $\text{TopicCoRank}_{\text{assign}}$ *Assignment*

Evaluation 〉 Baselines

Related work:

TopicRank	*Extraction*
Kea++	*Assignment*

TopicCoRank alternatives:

TopicCoRank$_{\text{extr}}$	*Extraction*
TopicCoRank$_{\text{assign}}$	*Assignment*

Stem-based comparisons at the top 10 outputed keyphrases

- Recall
- Precision
- F1-score

Evaluation 〉 Results

Method	Linguistics			Information Science			Archaeology		
	P	R	F	P	R	F	P	R	F
TopicRank	11.82	13.1	11.9	12.1	12.8	12.1	27.5	19.7	21.8
KEA++	11.6	13.0	12.1	9.5	10.2	9.6	23.5	16.2	18.8
TopicCoRank	**24.5**	**28.3**	**25.9**	**19.4**	**19.6**	**19.0**	**46.6**	**31.4**	**36.7**
TopicCoRank$_{extr}$	15.9	18.2	16.7	15.9	16.2	15.6	39.6	26.4	31.0
TopicCoRank$_{assign}$	25.8	29.6	27.2	19.9	20.0	19.5	49.6	33.3	39.0

- TopicCoRank outperforms baselines
- Graph-based co-ranking is succesful for extraction alone
- Graph-based co-ranking is succesful for assignment alone
- TopicCoRank$_{assign}$ performs best due to datasets specificities

Evaluation 〉 Results

Method	Linguistics			Information Science			Archaeology		
	P	R	F	P	R	F	P	R	F
TopicRank	11.82	13.1	11.9	12.1	12.8	12.1	27.5	19.7	21.8
KEA++	11.6	13.0	12.1	9.5	10.2	9.6	23.5	16.2	18.8
TopicCoRank	**24.5**	**28.3**	**25.9**	**19.4**	**19.6**	**19.0**	**46.6**	**31.4**	**36.7**
TopicCoRank$_{extr}$	15.9	18.2	16.7	15.9	16.2	15.6	39.6	26.4	31.0
TopicCoRank$_{assign}$	25.8	29.6	27.2	19.9	20.0	19.5	49.6	33.3	39.0

- TopicCoRank outperforms baselines
- Graph-based co-ranking is succesful for extraction alone
- Graph-based co-ranking is succesful for assignment alone
- TopicCoRank$_{assign}$ performs best due to datasets specificities

Evaluation 〉 Results

Method	Linguistics			Information Science			Archaeology		
	P	R	F	P	R	F	P	R	F
TopicRank	11.82	13.1	11.9	12.1	12.8	12.1	27.5	19.7	21.8
KEA++	11.6	13.0	12.1	9.5	10.2	9.6	23.5	16.2	18.8
TopicCoRank	**24.5**	**28.3**	**25.9**	**19.4**	**19.6**	**19.0**	**46.6**	**31.4**	**36.7**
TopicCoRank$_{extr}$	15.9	18.2	16.7	15.9	16.2	15.6	39.6	26.4	31.0
TopicCoRank$_{assign}$	25.8	29.6	27.2	19.9	20.0	19.5	49.6	33.3	39.0

- TopicCoRank outperforms baselines
- Graph-based co-ranking is succesful for extraction alone
- Graph-based co-ranking is succesful for assignment alone
- TopicCoRank$_{assign}$ performs best due to datasets specificities

Evaluation 〉 Results

Method	**Linguistics**			**Information Science**			**Archaeology**		
	P	R	F	P	R	F	P	R	F
TopicRank	11.82	13.1	11.9	12.1	12.8	12.1	27.5	19.7	21.8
KEA++	11.6	13.0	12.1	9.5	10.2	9.6	23.5	16.2	18.8
TopicCoRank	24.5	28.3	25.9	19.4	19.6	19.0	46.6	31.4	36.7
TopicCoRank$_{extr}$	15.9	18.2	16.7	15.9	16.2	15.6	39.6	26.4	31.0
TopicCoRank$_{assign}$	**25.8**	**29.6**	**27.2**	**19.9**	**20.0**	**19.5**	**49.6**	**33.3**	**39.0**

- TopicCoRank outperforms baselines
- Graph-based co-ranking is succesful for extraction alone
- Graph-based co-ranking is succesful for assignment alone
- TopicCoRank$_{assign}$ performs best due to datasets specificities

Outline

1. Related Work
2. Proposal
3. TopicCoRank
4. Evaluation
5. Conclusion

Conclusion

TopicCoRank:

- Supervised extension of TopicRank
- Combination of keyphrase extraction and assignment in a mutual reinforcing manner
- Good performances overall

Future work:

- Apply the supervised extension to other range of graph-based methods
- Investigate application/adaptation to non domain-specific documents
- Investigate impact on terminilogy/controlled vocabulary maintenance

Conclusion

TopicCoRank:

- Supervised extension of TopicRank
- Combination of keyphrase extraction and assignment in a mutual reinforcing manner
- Good performances overall

Future work:

- Apply the supervised extension to other range of graph-based methods
- Investigate application/adaptation to non domain-specific documents
- Investigate impact on terminilogy/controlled vocabulary maintenance

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`.

Olena Medelyan and Ian H Witten. Thesaurus Based Automatic Keyphrase Indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 296–297. ACM, 2006.