

Indexation de bout-en-bout dans les bibliothèques numériques scientifiques

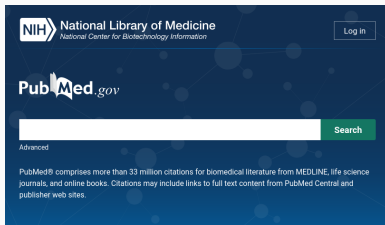
Ygor GALLINA

28/03/2022

Jury

Rapporteurs :	Josiane MOTHE Patrick PAROUBEK	Professeure, Université de Toulouse Ingénieur de recherche, Université de Paris-Saclay
Examineurs :	Lorraine GOEURLOT Richard DUFOUR	Maître de conférences, Université Grenoble Alpes Professeur, Nantes Université
Directrice :	Béatrice DAILLE	Professeure, Nantes Université
Encadrant :	Florian BOUDIN	Maître de conférences, Nantes Université

Bibliothèques numériques scientifiques



NIH National Library of Medicine
National Center for Biotechnology Information

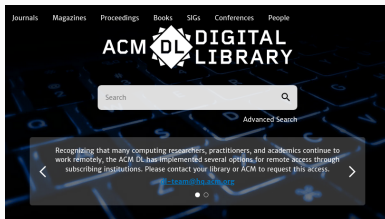
Log in

PubMed.gov

Advanced

Search

PubMed® comprises more than 33 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.



Journals Magazines Proceedings Books SIGs Conferences People

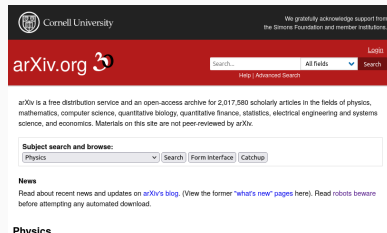
ACM DIGITAL LIBRARY

Search

Advanced Search

Recognizing that many computing researchers, practitioners, and academics continue to work remotely, the ACM DL has implemented several options for remote access through subscribing institutions. Please contact your library or ACM to request this access.

dl-team@hl.acm.org



Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv.org

Search... All Fields Search

Help / Advanced Search

arXiv is a free distribution service and an open-access archive for 2,017,580 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv.

Subject search and browse:

Physics Search Form Interface Catchup

News

Read about recent news and updates on [arXiv's blog](#). (View the former "what's new" pages here). Read robots beware before attempting any automated download.

Physics



ACL Anthology FAQ Corrections Submissions Search...

Welcome to the ACL Anthology!

The ACL Anthology currently hosts 74395 papers on the study of computational linguistics and natural language processing.

[Subscribe to the mailing list](#) to receive announcements and updates to the Anthology.

ACL Events

Venue	2021 - 2020	2019 - 2010	2009 - 2000	1999 - 1991
AAACL	20			
ACL	21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00			99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81 80
ANLP			00	97 94
CL	21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00			99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81 80

Bibliothèques numériques scientifiques

NIH National Library of Medicine
National Center for Biotechnology Information

PubMed

Bio-médical

24 566 348 doc.

Search

Advanced

PubMed® comprises more than 35 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

Cornell University

arXiv.org

Search

All Fields

Search

Help | Advanced Search

arXiv is a free distribution service and an open-access archive for 2,017,580 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, technical engineering and systems science, and economics.

Sciences formelles

1 999 642 doc.

Subject search and browse:

Physics

News

Read about recent news and updates on arXiv's blog. (View the former "what's new" pages here). Read robots beware before attempting any automated download.

Physics

Journals Magazines Proceedings Books SIGs Conferences People

ACM DIGITAL LIBRARY

ACMDL

Informatique

654 532 doc.

Recognizing that many academic libraries continue to work without the ability to access research through subscribing institutions, please contact your library or ACM to request this access.

acm@acm.org

ACL Anthology

FAQ Corrections Submissions

Search

Welcome to the ACL Anthology!

The ACL Anthology currently hosts 74395 papers in the field of computational linguistics and natural language processing.

Subscribe to the mailing list to receive announcements and updates to the Anthology.

ACL

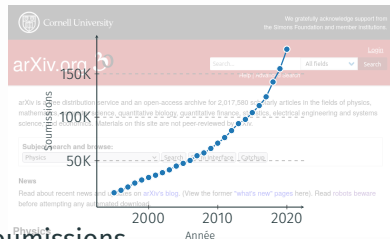
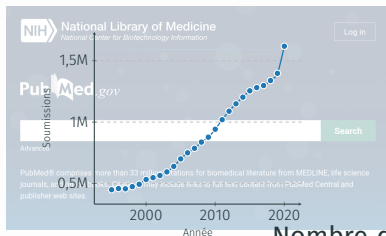
TALN

74 432 doc.

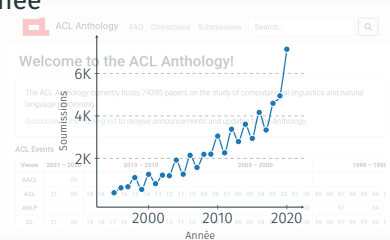
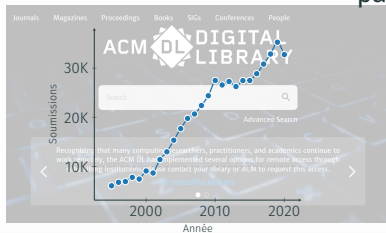
ACL Events

Venue	2021 - 2020	2019	2018	1999 - 1991
ACL	20			
ACL	21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81 80		
ANLP				00 97 94
CL	21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81 80		

Bibliothèques numériques scientifiques



Nombre de soumissions
par année



keyphrase generation using convolution network

Scholar Environ 17 500 résultats (0,4) 17 500 ré ANNÉE ▾ ☰

Deep **keyphrase generation** with a convolutional sequence to sequence model
Y Zhang, Y Fang, X Weidong - 2017 4th International ..., 2017 - ieeexplore.ieee.org
... model to **generate keyphrases from** vocabulary could ... **network**(RNN) suffers **from** low efficiency problem. We propose an architecture based entirely on **convolutional neural networks**. ...
☆ Enregistrer ⓘ Citer Cité 24 fois Autres articles

-Guided Encoding for **Keyphrase Generation** [PDF] aaai.org
[W Chen](#), [Y Gao](#), [J Zhang](#), [I King](#), [MR Lyu](#) - Proceedings of the AAAI ..., 2019 - ojs.aaai.org
... the generative setting **using** deep neural **networks**. However, ... -Guided **Network** (TG-Net) for automatic **keyphrase generation** ... Language modeling with gated **convolutional networks**. In ...
☆ Enregistrer ⓘ Citer Cité 60 fois Autres articles Les 9 versions ⓘ

GCN-based document representation for **keyphrase generation** enhanced by maximizing mutual information
P Yang, Y Ge, Y Yao, Y Yang - Knowledge-Based Systems, 2022 - Elsevier
... users quickly obtain valuable information **from** a large number of ... Neural **Network** (RNN) based **keyphrase generation** ... text, we apply Graph **Convolutional Network** (GCN) on document-...
☆ Enregistrer ⓘ Citer

Représenter un document pour qu'il soit facilement recherchable.

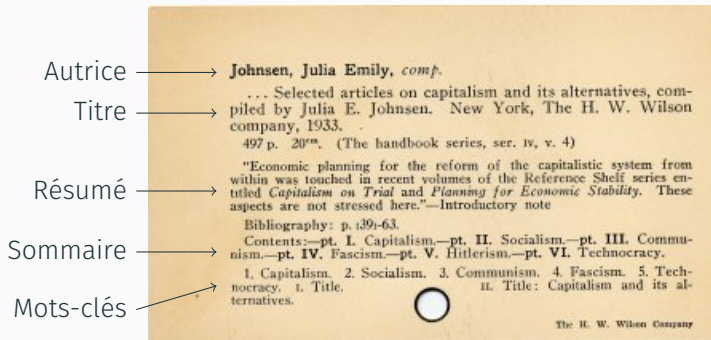


Figure – Notice scientifique. Source : libraryhistorybuff.org/catalog-cards.htm

Représenter un document pour qu'il soit facilement recherachable.

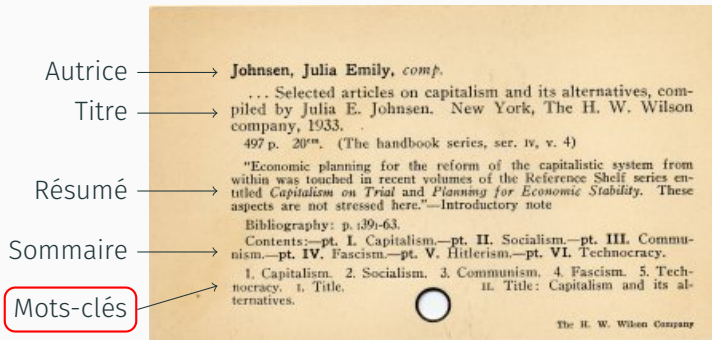


Figure – Notice scientifique. Source : libraryhistorybuff.org/catalog-cards.htm

Indexation par mots-clés

Les mots-clés sont généralement des **syntagmes nominaux** qui représentent les **concepts les plus importants** d'un document et servent de **condensateur textuel**. (Amar, 1997)

Types d'annotation

- **indexeur professionnel** (bibliothèque)
- **auteur** (conférences / journaux)
- **lecteur** (logiciel gestion bibliographique / étudiant·es)

Coût de l'annotation par des indexeurs

$\simeq 10\$/\text{doc}$ dans PubMed

Indexation par mots-clés

Les mots-clés sont généralement des **syntagmes nominaux** qui représentent les **concepts les plus importants** d'un document et servent de **condensateur textuel**. (Amar, 1997)

Types d'annotation

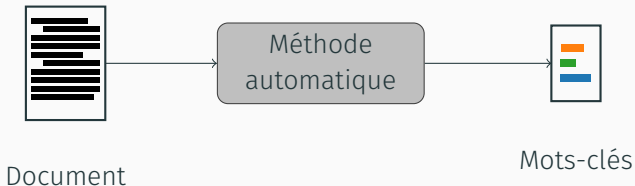
- **indexeur professionnel** (bibliothèque)
- **auteur** (conférences / journaux)
- **lecteur** (logiciel gestion bibliographique / étudiant-es)

Coût de l'annotation par des indexeurs

≈ 10\$/doc dans PubMed

=> 15 M\$ en 2020 et croissance exponentielle!

Production automatique de mots-clés



- 1972 : travaux pionniers ($TF \times IDF$ (Jones, 1972))
- 2000 : essor des méthodes extractives
- 2017 : introduction des méthodes génératives

Jeux de données annotés en mots-clés

	Corpus	Lang.	Ann.	#Entr.	#Test	#mots
Articles	CSTR (Witten et al., 1999)	en	A	130	500	11501
	NUS (Nguyen and Kan, 2007)	en	A ∪ L	-	211	8398
	PubMed (Schutz, 2008)	en	A	-	1320	5323
	ACM (Krapivin et al., 2009)	en	A	-	2304	9198
	Citeulike-180 (Medelyan et al., 2009)	en	L	-	182	8590
	SemEval-2010 (Kim et al., 2010)	en	A ∪ L	144	100	7961
Notices	Inspec (Hulth, 2003)	en	I	1000	500	135
	KDD (Caragea et al., 2014)	en	A	-	755	191
	WWW (Caragea et al., 2014)	en	A	-	1 330	164
	TermITH-Eval (Bougouin et al., 2016)	fr	I	-	400	165
	KP20k (Meng et al., 2017)	en	A	530 K	20 K	176
Journalistique	DUC-2001 (Wan and Xiao, 2008)	en	L	-	308	847
	500N-KPCrowd (Marujo et al., 2012)	en	L	450	50	465
	Wikinews (Bougouin et al., 2013)	fr	L	-	100	314

- Majorité de documents scientifiques
- Articles pleins peu accessibles (*paywall*)

Jeux de données annotés en mots-clés

		Corpus	Lang.	Ann.	#Entr.	#Test	#mots
Articles	CSTR (Witten et al., 1999)	en	A		130	500	11501
	NUS (Nguyen and Kan, 2007)	en	A ∪ L		-	211	8398
	PubMed (Schutz, 2008)	en	A		-	1320	5323
	ACM (Krapivin et al., 2009)	en	A		-	2304	9198
	Citeulike-180 (Medelyan et al., 2009)	en	L		-	182	8590
	SemEval-2010 (Kim et al., 2010)	en	A ∪ L		144	100	7961
Notices	Inspec (Hulth, 2003)	en	I		1000	500	135
	KDD (Caragea et al., 2014)	en	A		-	755	191
	WWW (Caragea et al., 2014)	en	A		-	1330	164
	TermITH-Eval (Bougouin et al., 2016)	fr	I		-	400	165
	KP20k (Meng et al., 2017)	en	A		530 K	20 K	176
Journalistique	DUC-2001 (Wan and Xiao, 2008)	en	L		-	308	847
	500N-KPCrowd (Marujo et al., 2012)	en	L		450	50	465
	Wikinews (Bougouin et al., 2013)	fr	L		-	100	314

- Majorité de documents scientifiques
- Articles pleins peu accessibles (*paywall*)

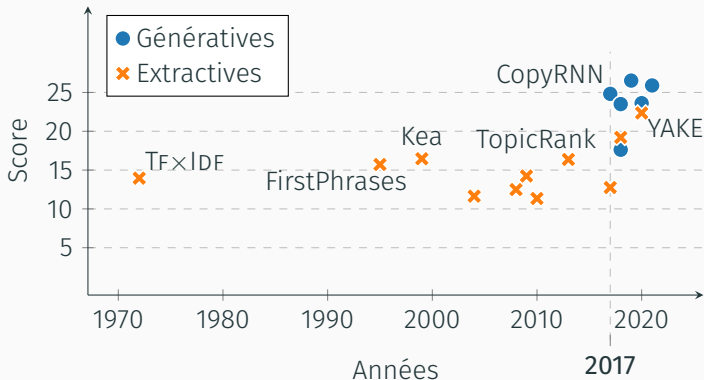
Jeux de données annotés en mots-clés

	Corpus	Lang.	Ann.	#Entr.	#Test	#mots
Articles	CSTR (Witten et al., 1999)	en	A	130	500	11501
	NUS (Nguyen and Kan, 2007)	en	A ∪ L	-	211	8398
	PubMed (Schutz, 2008)	en	A	-	1320	5323
	ACM (Krapivin et al., 2009)	en	A	-	2304	9198
	Citeulike-180 (Medelyan et al., 2009)	en	L	-	182	8590
	SemEval-2010 (Kim et al., 2010)	en	A ∪ L	144	100	7961
Notices	Inspec (Hulth, 2003)	en	I	1 000	500	135
	KDD (Caragea et al., 2014)	en	A	-	755	191
	WWW (Caragea et al., 2014)	en	A	-	1 330	164
	TermITH-Eval (Bougouin et al., 2016)	fr	I	-	400	165
	KP20k (Meng et al., 2017)	en	A	530 K	20 K	176
Journalistique	DUC-2001 (Wan and Xiao, 2008)	en	L	-	308	847
	500N-KPCrowd (Marujo et al., 2012)	en	L	450	50	465
	Wikinews (Bougouin et al., 2013)	fr	L	-	100	314

- Majorité de documents scientifiques
- Articles pleins peu accessibles (*paywall*)

1. Démontrer la validité des méthodes génératives.
 - Entraînement sur plusieurs jeux de données
 - Généralisation à d'autres genres de documents
2. Comparer les performances des méthodes état de l'art.
3. Évaluer la qualité des mots-clés au travers d'une tâche applicative.

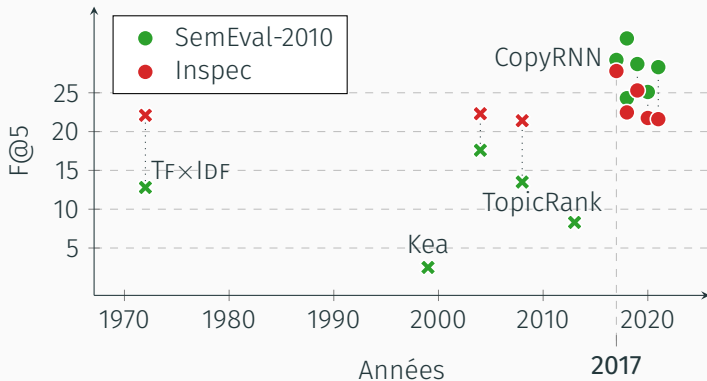
Comparaison des performances



Grande diversité de métrique et jeux de données utilisés.

Moyenne des scores rapportés de **toutes métriques et jeux de données confondus**.

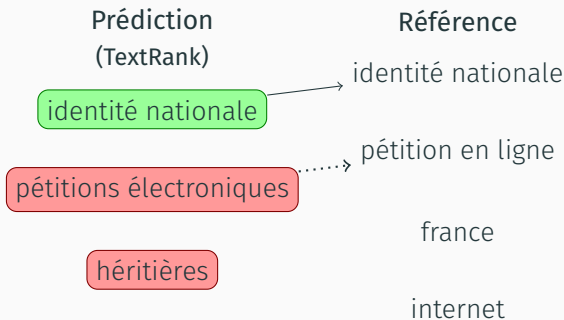
Comparaison des performances



Score de $F@5$ pour les **mots-clés présents** sur les jeux de données SemEval-2010 et Inspec.

1. Démontrer la validité des méthodes génératives.
2. Comparer les performances des méthodes état de l'art.
 - Cadre expérimental strict et unifié
 - Influence du type d'annotation sur l'évaluation
3. Évaluer la qualité des mots-clés au travers d'une tâche applicative.

Évaluation automatique



Basée sur l'**appariement strict** contre une **référence unique** subjective.

Métriques

Précision ; Rappel ; F-mesure ; MAP

Évaluation de la **correspondance** des mots-clés à une référence.

Impact dans les tâches applicatives ?

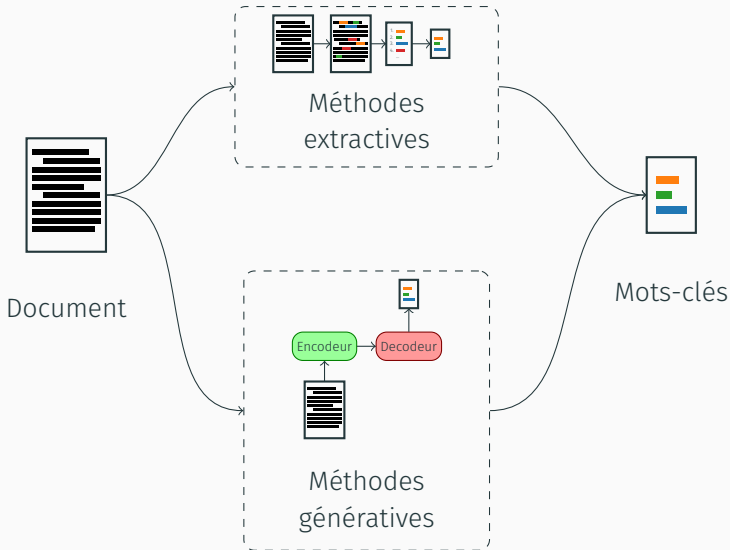
- indexation de documents
- détection d'opinion
- catégorisation de texte
- résumé automatique
- facilitation de la lecture

1. Démontrer la validité des méthodes génératives.
2. Comparer les performances des méthodes état de l'art.
3. Évaluer la qualité des mots-clés au travers d'une tâche applicative.
 - Nouveau cadre d'évaluation extrinsèque
 - Évaluation des méthodes état de l'art

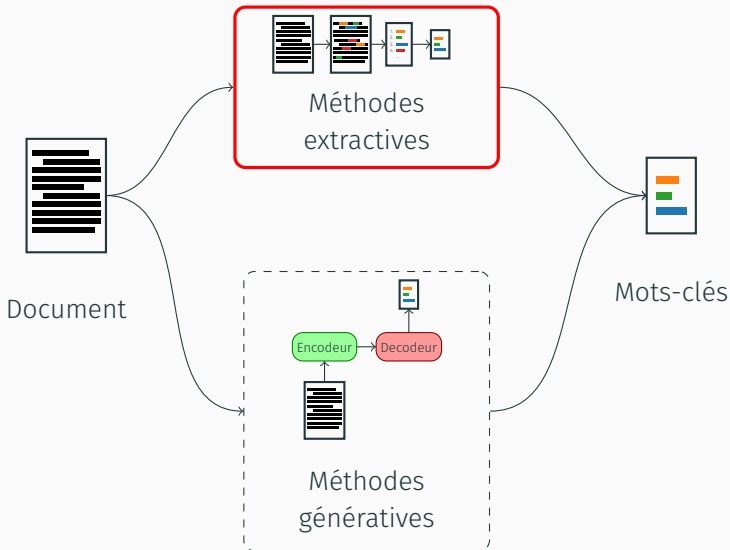
1. État de l'art de la production automatique de mots-clés
2. Contribution : Validation des méthodes génératives
3. Contribution : Évaluation comparative stricte
4. Contribution : Évaluation fondée sur la recherche documentaire

État de l'art de la production automatique de mots-clés

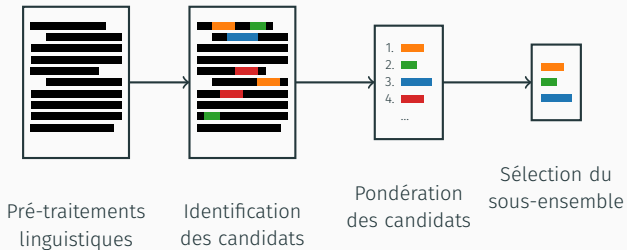
Méthodes de production automatique de mots-clés

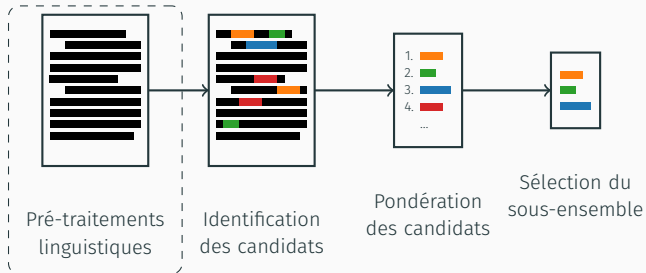


Méthodes de production automatique de mots-clés



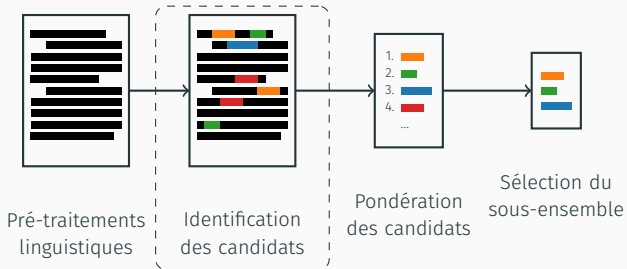
Méthodes extractives





Pré-traitements linguistiques

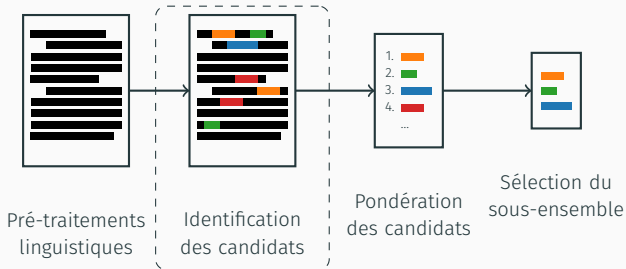
- segmentation en mots
- étiquetage morpho-syntaxique
- ...



Identification des candidats

- 3-grammes + filtrage
- noms et adjectifs
- ...

Méthodes extractives



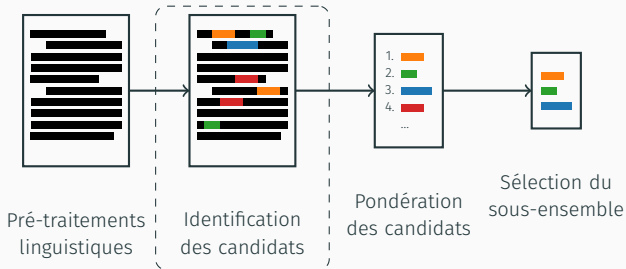
Texte : Nous présentons une méthode multilingue de catégorisation en mot vide [...]

Identification des candidats

- 3-grammes + filtrage
- noms et adjectifs
- ...

Candidats (11) :

- présentons
- présentons une méthode
- méthode multilingue
- méthode multilingue de catégorisation
- ...



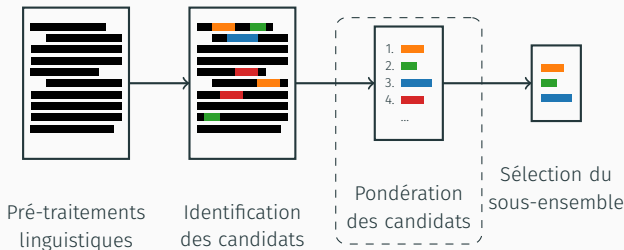
Identification des candidats

- 3-grammes + filtrage
- **noms et adjectifs**
- ...

Texte : Nous présentons une méthode multilingue de catégorisation en mot vide [...]

Candidats (3) :

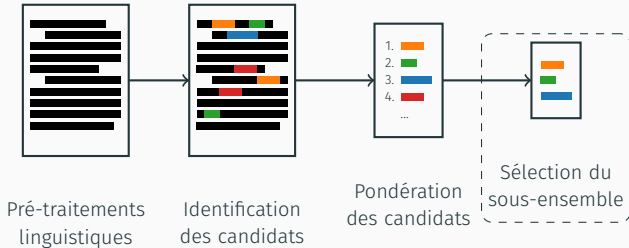
- méthode multilingue
- catégorisation
- mot vide



Pondération des candidats

- Méthodes **statistiques** : $TF \times IDF$ (Jones, 1972), YAKE (Campos et al., 2020)
- Méthodes fondées sur les **graphes** : TextRank (Mihalcea and Tarau, 2004), TopicalPageRank (Liu et al., 2010)
- Méthodes **supervisées** : Kea (Witten et al., 1999), CeKE (Caragea et al., 2014)

$$f(\text{candidat}) \\ = \textit{score}$$

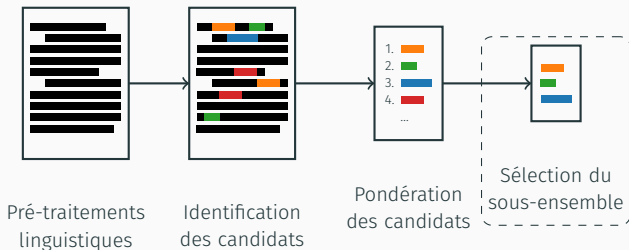


Sélection d'un sous-ensemble de mots-clés

- choix des n meilleurs
- suppression de la redondance

1. grammaires factorisées
2. dialectes apparentés
3. dialectes
4. description commune
5. grammaire
6. formalisation
7. couches

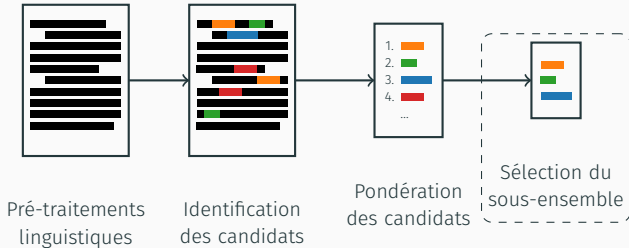
Méthodes extractives



Sélection d'un sous-ensemble de mots-clés

- choix des n meilleurs
- suppression de la redondance

1. grammaires factorisées
2. dialectes apparentés
3. dialectes
4. description commune
5. grammaire
6. formalisation
7. couches

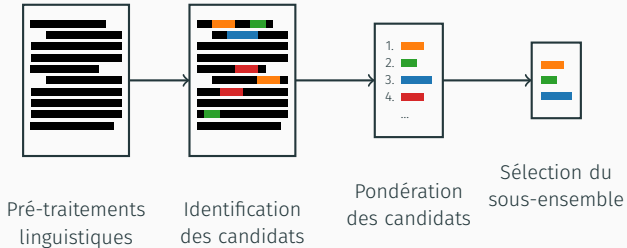


Sélection d'un sous-ensemble de mots-clés

- choix des n meilleurs
- suppression de la redondance

1. grammaires factorisées
2. dialectes apparentés
dialectes
3. description commune
grammaire
4. formalisation
5. couches

Méthodes extractives



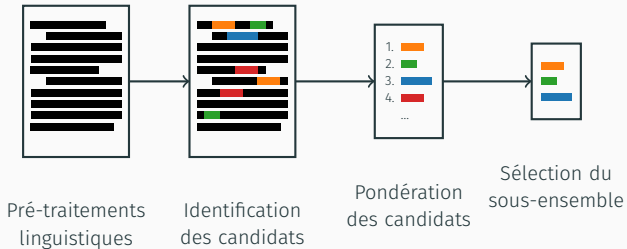
Avantages

- Rapide
- Interprétable

Inconvénients

- Propagation des erreurs
- Définition manuelle des traits
- Limité aux unités du document

Méthodes extractives



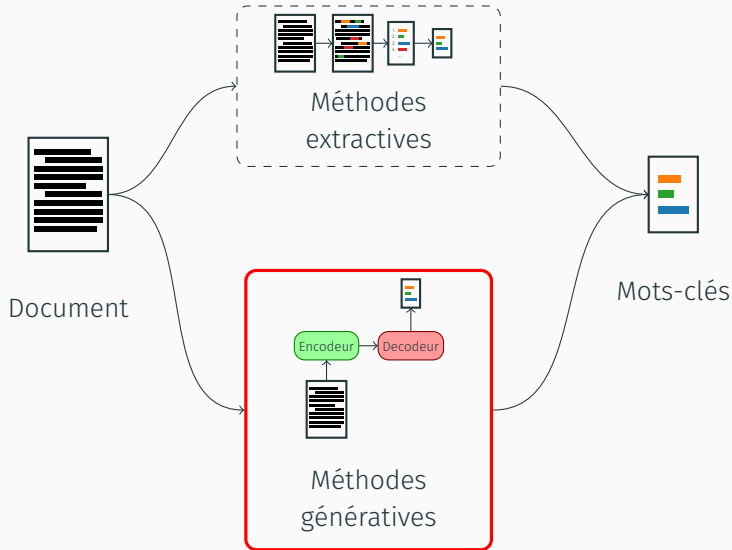
Avantages

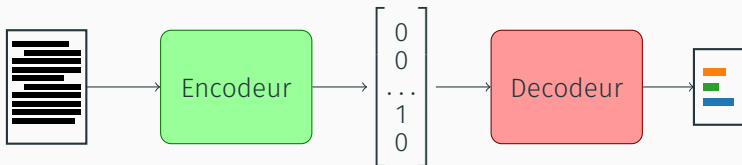
- Rapide
- Interprétable

Inconvénients

- Propagation des erreurs
- Définition manuelle des traits
- Limité aux unités du document
=> 50% des mots-clés de référence sont absents

Méthodes de production automatique de mots-clés





- Fondées sur le paradigme **encodeur-décodeur**
- Génération de mots à partir d'un vocabulaire différent du document
- Utilisation de **réseaux récurrents** (Meng et al., 2017), **transformers** (Diao et al., 2020) ou à **convolution** (Zhang et al., 2017).

Configurations d'entraînement



Configuration **One2One**

Configuration **One2Many**



...



...

CopyRNN (Meng et al., 2017)

- encodeur et décodeur récurrent bidirectionnel
- entraînement **one2one**
- mécanisme d'attention
- mécanisme de copie

CorrRNN (Chen et al., 2018)

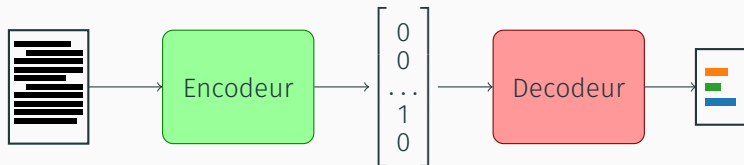
- entraînement **one2many**
- + mécanisme augmentant la diversité des mots-clés générés

CopyRNN (Meng et al., 2017)

- encodeur et décodeur récurrent bidirectionnel
- entraînement **one2one**
- mécanisme d'attention
- mécanisme de copie

Autres méthodes

- amélioration de la **diversité** des mots-clés (Chen et al., 2018; Chan et al., 2019; Yuan et al., 2020; Chen et al., 2020; Zhao and Zhang, 2019)
- amélioration de la **représentation du document** (Chen et al., 2019b,a)



Avantages

- Mots-clés absents
- De bout-en-bout

Inconvénients

- Boîte noire
- Nécessite de grandes quantités de données annotées

Contribution : Validation des méthodes génératives

- Actuellement **un seul** jeu de données de grande taille.
 - Insuffisant pour obtenir des conclusions fiables.
 - Résultats **transposables** à d'autres jeux de données?
- Nécessité de construire un nouveau jeu de données avec :
 - documents **annotés en mots-clés**;
 - **suffisamment** de documents pour entraîner des méthodes neuronales;
 - **annotation différente** de l'annotation auteur.

- Actuellement **un seul** jeu de données de grande taille.
 - Insuffisant pour obtenir des conclusions fiables.
 - Résultats **transposables** à d'autres jeux de données?
- Nécessité de construire un nouveau jeu de données avec :
 - documents **annotés en mots-clés**;
 - **suffisamment** de documents pour entraîner des méthodes neuronales;
 - **annotation différente** de l'annotation auteur.
- Les **articles journalistiques** sont disponibles en **grande quantité** sur internet
- Sont souvent **annotés en mots-clés** pour le référencement

NewYork Times

- Annotation éditeur
- \Rightarrow 296 974 articles

Japan Times

- Évaluer la généralisation
- \Rightarrow 11 057 articles

- Filtrage des documents trop longs, trop courts et redondants.

Corpus	Ann.	Lang.	Corpus			Document		
			#Entr.	#Val.	#Test	#mots	#mc	%abs
KPTimes	<i>E</i>	en	260 K	20 K	20 K	738	5,0	38,4
+ NYTimes	<i>E</i>	en	260 K	20 K	10 K	905	5,0	52,5
+ JPTimes	<i>E</i>	en	-	-	10 K	570	5,0	24,2

NewYork Times

- Annotation éditeur
- \Rightarrow 296 974 articles

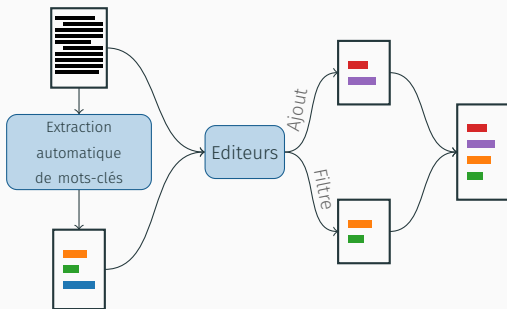
Japan Times

- Évaluer la généralisation
- \Rightarrow 11 057 articles

- Filtrage des documents trop longs, trop courts et redondants.

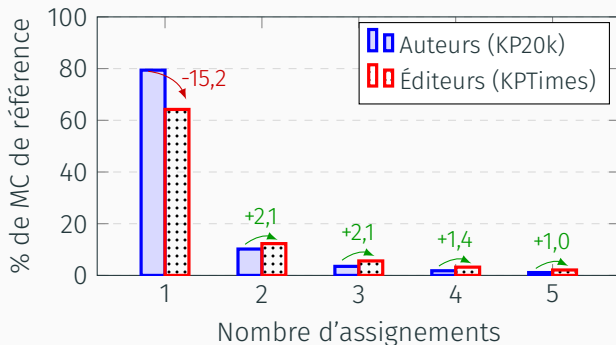
Corpus	Ann.	Lang.	Corpus			Document		
			#Entr.	#Val.	#Test	#mots	#mc	%abs
KPTimes	<i>E</i>	en	260 K	20 K	20 K	738	5,0	38,4
+ NYTimes	<i>E</i>	en	260 K	20 K	10 K	905	5,0	52,5
+ JPTimes	<i>E</i>	en	-	-	10 K	570	5,0	24,2

Processus d'annotation éditeur



- Annotation **semi-automatique** basée sur un vocabulaire contrôlé
- Les éditeurs **valident** et **complètent** les mots-clés proposés
- Annotation **cohérente** (vocabulaire contrôlé) et **exhaustive** (ajout de mots-clés)

Cohérence de l'annotation



- 80% de mots-clés associés à un seul document pour l'annotation auteur

Hypothèses :

- Évaluation plus fiable
- Apprentissage plus efficace des méthodes génératives

Méthodes extractives

- $TF \times IDF$ (Jones, 1972) : spécificité des mots
- MultiPartiteRank (Boudin, 2018) : centralité des mots
- Kea (Witten et al., 1999) : classifieur bayésien

Méthodes génératives

- CopyRNN
 - CopyNews : Entraîné sur KPTimes (articles journalistiques)
 - CopySci : Entraîné sur KP20k (notices scientifiques)

Utilisation des paramètres recommandés par les auteurs.

Métrique : F-mesure sur les 10 meilleurs mots-clés

Cadre Expérimental

Corpus	Ann.	Corpus			Document		
		#Entr.	#Val.	#Test	#mots	#mc	%abs
Journalistique							
KPTimes	<i>E</i>	-	-	20 K	738	5,0	38,4
JPTimes	<i>E</i>	-	-	10 K	570	5,0	24,2
NYTimes	<i>E</i>	260 K	20 K	10 K	905	5,0	52,5
DUC-2001	<i>L</i>	-	-	308	847	8,1	3,1
Scientifique							
KP20k	<i>A</i>	530 K	20 K	20 K	176	5,3	42,4

- Annotation lecteur de DUC-2001 : plus de mots-clés que les autres jeux de données et majoritairement présents.

Cadre Expérimental

Corpus	Ann.	Corpus			Document		
		#Entr.	#Val.	#Test	#mots	#mc	%abs
Journalistique							
KPTimes	E	-	-	20 K	738	5,0	38,4
JPTimes	E	-	-	10 K	570	5,0	24,2
NYTimes	E	260 K	20 K	10 K	905	5,0	52,5
DUC-2001	L	-	-	308	847	8,1	3,1
Scientifique							
KP20k	A	530 K	20 K	20 K	176	5,3	42,4

- Annotation lecteur de DUC-2001 : plus de mots-clés que les autres jeux de données et majoritairement présents.

La supériorité de CopyRNN est-elle toujours présente avec NYTimes ?

F@10	NYTimes
TF×IDF	9,6
MPRank	11,2
Kea	11,0
CopyNews	39,3

- Comme sur KP20k, CopyRNN obtient toujours de meilleurs résultats que les méthodes extractives.

Document similaire, type d'annotation différent

Les performances de CopyRNN sont-elles généralisables à un type d'annotation différent ?

F@10	Éditeur	Éditeur	Lecteur
	NYTimes	JPTimes	DUC-2001
TF×IDF	9,6	15,1	23,0
MPRank	11,2	16,8	25,3
Kea	11,0	16,6	26,2
CopyNews	39,3	24,6	10,5

- CopyNews connaît une première baisse de performances lors de l'évaluation sur JPTimes.
- CopyNews généralise mal à un type d'annotation différent.

Document similaire, type d'annotation différent

Les performances de CopyRNN sont-elles généralisables à un type d'annotation différent ?

	Éditeur	Éditeur	Lecteur
F@10	NYTimes	JPTimes	DUC-2001
TF×IDF	9,6	15,1	23,0
MPRank	11,2	16,8	25,3
Kea	11,0	16,6	26,2
CopyNews	39,3	24,6	10,5

- CopyNews connaît une première baisse de performances lors de l'évaluation sur JPTimes.
- CopyNews généralise mal à un type d'annotation différent.

Document similaire, type d'annotation différent

Les performances de CopyRNN sont-elles généralisables à un type d'annotation différent ?

	Éditeur	Éditeur	Lecteur
F ₀ 10	NYTimes	JPTimes	DUC-2001
TF×IDF	9,6	15,1	23,0
MPRank	11,2	16,8	25,3
Kea	11,0	16,6	26,2
CopyNews	39,3	24,6	10,5

- CopyNews connaît une première baisse de performances lors de l'évaluation sur JPTimes.
- CopyNews généralise mal à un type d'annotation différent.

Document similaire, type d'annotation différent

Les performances de CopyRNN sont-elles généralisables à un type d'annotation différent ?

	Éditeur NYTimes	Éditeur JPTimes	Lecteur DUC-2001
F@10			
Tf×Idf	9,6	15,1	23,0
MPRank	11,2	16,8	25,3
Kea	11,0	16,6	26,2
CopyNews	39,3	24,6	10,5

Exemple de mots-clés d'un article de DUC-2001 (AP890511-0126) :

CopyNews : tuberculosis – us – prisons – new jersey – medicine and health

M.-c. lecteur : tuberculosis rate – u.s. prisons – aids-virus infections –
tuberculosis cases – airborne transmission – cdc

Document différent, type d'annotation différent

Les performances de CopyRNN sont-elles généralisables à d'autres genres de documents ?

F@10	Éditeur	Auteur
	KPTimes	KP20k
CopyNews	31,9	6,6
CopySci	14,9	25,5

- CopyNews obtient de meilleures performances grâce à son annotation plus cohérente.
- Faible généralisation à un type d'annotation et un genre différent.

Document différent, type d'annotation différent

Les performances de CopyRNN sont-elles généralisables à d'autres genres de documents ?

F@10	Éditeur	Auteur
	KPTimes	KP20k
CopyNews	31,9	6,6
CopySci	14,9	25,5

Exemple de mots-clés d'un article de KP20k (011355) :

CopyNews : research – science and technology – medicine and health – diagnostic problem-solving – science journal

M.-c. auteur : diagnosis – multiple disorders – competition – neural networks – learning

Problématique

Entraînement des méthodes génératives sur **un seul** jeu de données : **KP20k**.

Introduction de **KPTimes**, le seul jeu de données de **grande taille** d'**articles journalistiques** annotés en mots-clés par des **éditeurs**.

Conclusion

1. Résultats transposables à KPTimes.
2. Faible généralisation à des documents de **genres différents** et à un **type d'annotation différent**.
3. Faibles performances en partie liées à l'évaluation.

Contribution : Évaluation comparative stricte

Les résultats rapportés dans les articles ne sont **pas directement comparables**.

Incomparabilité causé par

1. Jeux de données différents
2. Métriques différentes
3. Pré-traitements différents

Trois articles publiés à ACL 2017 ne partagent aucun jeu de données et aucune métrique : (Meng et al., 2017), (Florescu and Caragea, 2017), (Teneva and Cheng, 2017)

Les résultats rapportés dans les articles ne sont **pas directement comparables**.

Incomparabilité causé par

1. Jeux de données différents
2. Métriques différentes
3. Pré-traitements différents

Trois articles publiés à ACL 2017 ne partagent aucun jeu de données et aucune métrique : (Meng et al., 2017), (Florescu and Caragea, 2017), (Teneva and Cheng, 2017)

⇒ Évaluation à l'aide d'un cadre expérimental **strict** et **unifié**.

Méthodes évaluées

Méthodes de base

- FirstPhrases
- TextRank (Mihalcea and Tarau, 2004)
- $TF \times IDF$ (Jones, 1972)

Méthodes non supervisées

- PositionRank (Florescu and Caragea, 2017)
- MultiPartiteRank (Boudin, 2018)
- EmbedRank (Bennani-Smires et al., 2018)

Méthodes supervisées

- Kea (Witten et al., 1999)
- CopyRNN (Meng et al., 2017)
- CorrRNN (Chen et al., 2018)

	Corpus	Lang.	Ann.	#Entr.	#Test	#mots	#mc	%abs
Journalistiques	Articles							
	PubMed (Schutz, 2008)	en	A	-	1320	5323	5	17
	ACM (Krapivin et al., 2009)	en	A	-	2304	9198	5	16
	SemEval-2010 (Kim et al., 2010)	en	A ∪ L	144	100	7961	15	20
	Notices							
	Inspec (Hulth, 2003)	en	I	1 000	500	135	10	22
	WWW (Caragea et al., 2014)	en	A	-	1 330	164	5	52
	KP20k (Meng et al., 2017)	en	A	530 K	20 K	176	5	43
	Journalistiques							
	DUC-2001 (Wan and Xiao, 2008)	en	L	-	308	847	8	4
	500N-KPCrowd (Marujo et al., 2012)	en	L	450	50	465	46	11
	KPTimes (Gallina et al., 2019)	en	E	260 K	20 K	784	5	41

- Représentatifs des jeux de données utilisés
- Différents types d'annotation (Auteur, Lecteur, Indexeur, Éditeur)

Paramètres expérimentaux unifiés

- **Prétraitements** : réalisé avec Stanford CoreNLP.
- **Sélection des candidats** : syntagmes nominaux ($A*N+$) + filtrage.
- **Métrique** : F@10
- **Entraînement** :
 - Kea : en validation croisée si pas de documents d'entraînement.
 - Méthodes génératives : sur KP20k et KPTime en fonction du genre de document.
- Utilisation des paramètres recommandés par les auteurs dans les articles originaux.

Cadre expérimental strict

Réimplémentation

Est-ce que nos réimplémentations obtiennent des résultats comparables aux méthodes originales ?

Méthode	Jeu de données	Métrique	Orig.	Réimp.	Diff.
PositionRank	WWW	F@8	12,3	11,7	-0,6
MPRank	SemEval-2010	F@10	14,5	14,3	-0,2
EmbedRank	Inspec	F@10	37,1	35,6	-1,5
CopyRNN	KP20k	F@10 (prs.)	26,2	28,2	+2
CorrRNN	ACM	F@10 (prs.)	27,8	24,7	-3,1

- Résultats comparables
- Différences liées aux paramètres peu explicités

Analyse des résultats

Résultats généraux

F@10	Articles scientifiques			Notices scientifiques			Articles journalistiques		
	PubMed	ACM	SemEval	Inspec	WWW	KP20k	DUC-2001	KPCrowd	KPTimes
FirstPhrases	15,4	13,6	13,8	29,3	10,2	13,5	24,6	17,1	9,2
TextRank	1,8	2,5	3,5	35,8	8,4	10,2	21,5	7,1	2,7
Tf×Idf	16,7	12,1	17,7	36,5	9,3	11,6	23,3	16,9	9,6
PositionRank	4,9	5,7	6,8	34,2	11,6 [†]	14,1 [†]	28,6 [†]	13,4	8,5
MPRank	15,8	11,6	14,3	30,5	10,8 [†]	13,6 [†]	25,6	18,2	11,2 [†]
EmbedRank	3,7	2,1	2,5	35,6	10,7 [†]	12,4	29,5 [†]	12,4	4,0
Kea	18,6 [†]	14,2 [†]	19,5 [†]	34,5	11,0 [†]	14,0 [†]	26,5 [†]	17,3	11,0 [†]
CopyRNN	24,2 [†]	24,4 [†]	20,3 [†]	28,2	22,2 [†]	25,4 [†]	10,5	8,4	39,3 [†]
CorrRNN	20,8 [†]	21,1 [†]	19,4	27,9	19,9 [†]	21,8 [†]	10,5	7,8	20,5 [†]

- Les méthodes génératives obtiennent les meilleures performances.
- Tf×Idf et FirstPhrases sont compétitives.
- Inspec (annotation indexeur) obtient les meilleures performances en général.

Résultats généraux

F@10	Articles scientifiques			Notices scientifiques			Articles journalistiques		
	PubMed	ACM	SemEval	Inspec	WWW	KP20k	DUC-2001	KPCrowd	KPTimes
FirstPhrases	15,4	13,6	13,8	29,3	10,2	13,5	24,6	17,1	9,2
TextRank	1,8	2,5	3,5	35,8	8,4	10,2	21,5	7,1	2,7
Tf×Idf	16,7	12,1	17,7	36,5	9,3	11,6	23,3	16,9	9,6
PositionRank	4,9	5,7	6,8	34,2	11,6 [†]	14,1 [†]	28,6 [†]	13,4	8,5
MPRank	15,8	11,6	14,3	30,5	10,8 [†]	13,6 [†]	25,6	18,2	11,2 [†]
EmbedRank	3,7	2,1	2,5	35,6	10,7 [†]	12,4	29,5[†]	12,4	4,0
Kea	18,6 [†]	14,2 [†]	19,5 [†]	34,5	11,0 [†]	14,0 [†]	26,5 [†]	17,3	11,0 [†]
CopyRNN	24,2[†]	24,4[†]	20,3[†]	28,2	22,2[†]	25,4[†]	10,5	8,4	39,3[†]
CorrRNN	20,8 [†]	21,1 [†]	19,4	27,9	19,9 [†]	21,8 [†]	10,5	7,8	20,5 [†]

- Les méthodes génératives obtiennent les meilleures performances.
- Tf×Idf et FirstPhrases sont compétitives.
- Inspec (annotation indexeur) obtient les meilleures performances en général.

Résultats généraux

F@10	Articles scientifiques			Notices scientifiques			Articles journalistiques		
	PubMed	ACM	SemEval	Inspec	WWW	KP20k	DUC-2001	KPCrowd	KPTimes
FirstPhrases	15,4	13,6	13,8	29,3	10,2	13,5	24,6	17,1	9,2
TextRank	1,8	2,5	3,5	35,8	8,4	10,2	21,5	7,1	2,7
Tf×Idf	16,7	12,1	17,7	36,5	9,3	11,6	23,3	16,9	9,6
PositionRank	4,9	5,7	6,8	34,2	11,6 [†]	14,1 [†]	28,6 [†]	13,4	8,5
MPRank	15,8	11,6	14,3	30,5	10,8 [†]	13,6 [†]	25,6	18,2	11,2 [†]
EmbedRank	3,7	2,1	2,5	35,6	10,7 [†]	12,4	29,5 [†]	12,4	4,0
Kea	18,6 [†]	14,2 [†]	19,5 [†]	34,5	11,0 [†]	14,0 [†]	26,5 [†]	17,3	11,0 [†]
CopyRNN	24,2 [†]	24,4 [†]	20,3 [†]	28,2	22,2 [†]	25,4 [†]	10,5	8,4	39,3 [†]
CorrRNN	20,8 [†]	21,1 [†]	19,4	27,9	19,9 [†]	21,8 [†]	10,5	7,8	20,5 [†]

- Les méthodes génératives obtiennent les meilleures performances.
- Tf×Idf et FirstPhrases sont compétitives.
- Inspec (annotation indexeur) obtient les meilleures performances en général.

Résultats généraux

F@10	Articles scientifiques			Notices scientifiques			Articles journalistiques		
	PubMed	ACM	SemEval	Inspec	WWW	KP20k	DUC-2001	KPCrowd	KPTimes
FirstPhrases	15,4	13,6	13,8	29,3	10,2	13,5	24,6	17,1	9,2
TextRank	1,8	2,5	3,5	35,8	8,4	10,2	21,5	7,1	2,7
Tf×Idf	16,7	12,1	17,7	36,5	9,3	11,6	23,3	16,9	9,6
PositionRank	4,9	5,7	6,8	34,2	11,6 [†]	14,1 [†]	28,6 [†]	13,4	8,5
MPRank	15,8	11,6	14,3	30,5	10,8 [†]	13,6 [†]	25,6	18,2	11,2 [†]
EmbedRank	3,7	2,1	2,5	35,6	10,7 [†]	12,4	29,5 [†]	12,4	4,0
Kea	18,6 [†]	14,2 [†]	19,5 [†]	34,5	11,0 [†]	14,0 [†]	26,5 [†]	17,3	11,0 [†]
CopyRNN	24,2 [†]	24,4 [†]	20,3 [†]	28,2	22,2 [†]	25,4 [†]	10,5	8,4	39,3 [†]
CorrRNN	20,8 [†]	21,1 [†]	19,4	27,9	19,9 [†]	21,8 [†]	10,5	7,8	20,5 [†]

- Les méthodes génératives obtiennent les meilleures performances.
- Tf×Idf et FirstPhrases sont compétitives.
- Inspec (annotation indexeur) obtient les meilleures performances en général.

Impact de l'annotation sur l'évaluation

Annotation **indexeur** et **auteur** de 64 documents communs à Inspec et KP20k.

Comparaison de l'annotation indexeur et auteur (id Inspec : 2107)

Indexeur (13) : deindividuation – personal identifiability – group identity – asynchronous computer-mediated group interaction – group processes – group cohesion – e-mail discussions – social identity theory – geographically dispersed computer users – group polarization – social issues – psychology – internet

Tf×Idf : deindividuation – personal identifiability – group identity – asynchronous computer-mediated group interaction – group processes

Auteur (4) : deindividuation – social identity – computer-mediated communication – e-mail

Impact de l'annotation sur l'évaluation

Annotation **indexeur** et **auteur** de 64 documents communs à Inspec et KP20k.

Comparaison de l'annotation indexeur et auteur (id Inspec : 2107)

Indexeur (13) : deindividuation – personal identifiability – group identity – asynchronous computer-mediated group interaction – group processes – group cohesion – e-mail discussions – social identity theory – geographically dispersed computer users – group polarization – social issues – psychology – internet

Tf×Idf : deindividuation – personal identifiability – group identity – asynchronous computer-mediated group interaction – group processes

Auteur (4) : deindividuation – social identity – computer-mediated communication – e-mail

Impact de l'annotation sur l'évaluation

Annotation **indexeur** et **auteur** de 64 documents communs à Inspec et KP20k.

Comparaison de l'annotation indexeur et auteur (id Inspec : 2107)

Indexeur (13) : deindividuation – personal identifiability – group identity – asynchronous computer-mediated group interaction – group processes – group cohesion – e-mail discussions – social identity theory – geographically dispersed computer users – group polarization – social issues – psychology – internet

TF×IDF : deindividuation – personal identifiability – group identity – *asynchronous computer-mediated group interaction* – group processes

Auteur (4) : deindividuation – social identity – computer-mediated communication – e-mail

Impact de l'annotation sur l'évaluation

Méthode	F@10	Index.	Auteur
FirstPhrases	26,9	13,4	
TextRank	34,5	12,0	
TF×IDF	35,0	14,6	
PositionRank	33,2	15,3	
MPRank	27,9	13,7	
EmbedRank	35,3	15,1	
Kea	32,9	15,4	
CopyRNN	33,8	27,9[‡]	
CorrRNN	28,7	25,0	
Moy.	32,0	17,0	

Impact de l'annotation sur l'évaluation

Méthode	F@10	Index.	Auteur
FirstPhrases	26,9	13,4	
TextRank	34,5	12,0	
TF×IDF	35,0	14,6	
PositionRank	33,2	15,3	
MPRank	27,9	13,7	
EmbedRank	35,3	15,1	
Kea	32,9	15,4	
CopyRNN	33,8	27,9[‡]	
CorrRNN	28,7	25,0	
Moy.	32,0	17,0	

- Performances sur la référence indexeur **plus haute** que sur la référence auteur.

Impact de l'annotation sur l'évaluation

Méthode	F@10	Index.	Auteur
FirstPhrases	26,9	13,4	
TextRank	34,5	12,0	
TF×IDF	35,0	14,6	
PositionRank	33,2	15,3	
MPRank	27,9	13,7	
EmbedRank	35,3	15,1	
Kea	32,9	15,4	
CopyRNN	33,8	27,9 [‡]	
CorrRNN	28,7	25,0	
Moy.	32,0	17,0	

- Performances sur la référence indexeur **plus haute** que sur la référence auteur.
- Contraste extractives / génératives inexistant avec l'annotation indexeur.

Impact de l'annotation sur l'évaluation

Méthode	F@10	Index.	Auteur
FirstPhrases	26,9	13,4	
TextRank	34,5	12,0	
TF×IDF	35,0	14,6	
PositionRank	33,2	15,3	
MPRank	27,9	13,7	
EmbedRank	35,3	15,1	
Kea	32,9	15,4	
CopyRNN	33,8	27,9[‡]	
CorrRNN	28,7	25,0	
Moy.	32,0	17,0	

- Performances sur la référence indexeur **plus haute** que sur la référence auteur.
- Contraste extractives / génératives inexistant avec l'annotation indexeur.
- ⇒ Évaluation peu fiable.

Comparaison des performances des méthodes état de l'art

Problématique

Comparaison directe des performances impossible à cause de la variabilité dans les jeux de données et les métriques utilisées.

Évaluation des méthodes à l'aide d'un cadre expérimental strict.

Conclusion

- Méthodes de base ($TF \times IDF$) toujours compétitives sans données d'apprentissage.
- Les méthode génératives (CopyRNN) représentent l'état de l'art.
- Annotation auteur sous-évalue les méthodes.
- Conclusions tirées de l'évaluation peu fiables car changeantes en fonction du type d'annotation.

Contribution : Évaluation fondée sur la recherche documentaire

Évaluation intrinsèque

- peu fiable et pessimiste.
- correspondance à l'annotation sans tenir compte de leur qualité.

Évaluation intrinsèque

- peu fiable et pessimiste.
- correspondance à l'annotation sans tenir compte de leur qualité.

⇒ **Évaluation extrinsèque** grâce à une tâche de recherche documentaire.

- Retourner une liste de documents **ordonnés** par **pertinence** par rapport à une **requête**.
- Calcul de score grâce à des jugements de pertinence.

Requête : Architecture of the DNA computer

1. An Investigation of Education of Drafting by CAD-System in the Field of Building Equipment and Machine.
2. DNA Computing and Related Fields
3. Design of a Processor Core for Massively Parallel Computers
4. A study on the computational accuracy in DNA computation
5. Studies in Brain-Structured Supercomputers

- Retourner une liste de documents **ordonnés** par **pertinence** par rapport à une **requête**.
- Calcul de score grâce à des **jugements de pertinence**.

Requête : Architecture of the DNA computer

1. An Investigation of Education of Drafting by CAD-System in the Field of Building Equipment and Machine.
2. DNA Computing and Related Fields
3. Design of a Processor Core for Massively Parallel Computers
4. A study on the computational accuracy in DNA computation
5. Studies in Brain-Structured Supercomputers

Systèmes de recherche d'information

1. Okapi-BM25 (Robertson et al., 1999)
2. + RM3 (Abdul-Jaleel et al., 2004) : expansion de requête
 - BM25 est toujours compétitive (Thakur et al., 2021)
 - Implémentations de *anserini* (Yang et al., 2017)

Collection de test

- NTCIR-2 (Kando, 2001) : compétition de recherche ad-hoc de notices scientifiques en anglais
- Mots-clés auteurs pour 98% des documents

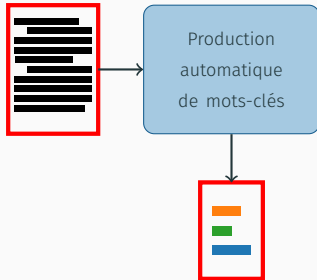
Collection	#Doc.	#Dmots	#Req.	#Rmots	#pert.	#mc	%abs
NTCIR-2	322 058	156,8	49	11,3	28,8	4,8	38,1

Configurations d'indexation

1. Titre et Résumé ($T+R$)

Est-ce que les mots-clés automatiques aident la recherche documentaire ?

Document (T+R)



Ajout des 5 meilleurs mots-clés des méthodes :

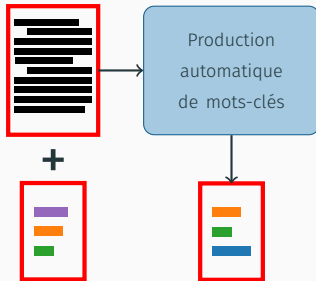
- MPRank
- Kea
- CorrRNN
- CopyRNN

Configurations d'indexation

2. Titre, Résumé et Mots-clés de référence ($T+R+M$)

Est-ce que les mots-clés automatiques sont complémentaires des mots-clés auteur ?

Document (T+R)



Mots-clés
auteur (M)

Ajout des 5 meilleurs mots-clés
des méthodes :

- MPRank
- Kea
- CorrRNN
- CopyRNN

Impact des mots-clés de référence

Indexation	BM25	+RM3	F@5
T+R	29,6	32,8	-
T+R+M	31,9	35,5	-

1. Les mots-clés auteurs sont utiles!

TABLEAU – Scores de MAP sur la collection NTCIR-2.

Impact des mots-clés de référence

Indexation	BM25	+RM3	F@5
T+R	29,6	32,8	-
+ MPRank	29,7 0,1	33,0 0,2	17,1
+ Kea (KP20k)	0,3 0,7	33,9 1,1	18,5
+ CorrRNN	31,6 [†] 2,1	35,0[†] 2,2	22,0
+ CopyRNN	31,4 [†] 1,9	34,8 [†] 2,0	23,9
T+R+M	31,9	35,5	-

1. Les mots-clés auteurs sont utiles!

TABLEAU – Scores de MAP sur la collection NTCIR-2.

Impact des mots-clés de référence

Indexation	BM25		+RM3		F@5
T+R	29,6		32,8		-
+ MPRank	29,7	0,1	33,0	0,2	17,1
+ Kea (KP20k)	0,3	0,7	33,9	1,1	18,5
+ CorrRNN	31,6	2,1	35,0 [†]	2,2	22,0
+ CopyRNN	31,4 [†]	1,9	34,8 [†]	2,0	23,9
T+R+M	31,9		35,5		-

1. Les mots-clés auteurs sont utiles!
2. Les mots-clés produits par les méthodes génératives sont une **alternative** aux mots-clés auteurs

TABLEAU – Scores de MAP sur la collection NTCIR-2.

Impact des mots-clés de référence

Indexation	BM25		+RM3		F@5
T+R	29,6		32,8		-
+ MPRank	29,7	0,1	33,0	0,2	17,1
+ Kea (KP20k)	0,3	0,7	33,9	1,1	18,5
+ CorrRNN	31,6 [†]	2,1	35,0[†]	2,2	22,0
+ CopyRNN	31,4 [†]	1,9	34,8 [†]	2,0	23,9
T+R+M	31,9		35,5		-
+ MPRank	32,0	0,1	35,8	0,3	17,1
+ Kea (KP20k)	32,1	0,2	36,0	0,5	18,5
+ CorrRNN	32,4	0,5	36,9 [†]	1,4	22,0
+ CopyRNN	32,5	0,5	37,1[†]	1,6	23,9

1. Les mots-clés auteurs sont utiles!
2. Les mots-clés produits par les méthodes génératives sont une alternative aux mots-clés auteurs

TABLEAU – Scores de MAP sur la collection NTCIR-2.

Impact des mots-clés de référence

Indexation	BM25		+RM3		F@5
T+R	29,6		32,8		-
+ MPRank	29,7	0,1	33,0	0,2	17,1
+ Kea (KP20k)	0,3	0,7	33,9	1,1	18,5
+ CorrRNN	31,6 [†]	2,1	35,0	2,2	22,0
+ CopyRNN	31,4 [†]	1,9	34,8 [†]	2,0	23,9
T+R+M	31,9		35,5		-
+ MPRank	32,0	0,1	35,8	0,3	17,1
+ Kea (KP20k)	32,1	0,2	36,0	0,5	18,5
+ CorrRNN	32,4	0,5	36,9	1,4	22,0
+ CopyRNN	32,5	0,5	37,1[†]	1,6	23,9

1. Les mots-clés auteurs sont utiles!
2. Les mots-clés produits par les méthodes génératives sont une alternative aux mots-clés auteurs
3. Les mots-clés de **référence** et les mots-clés **automatiques** sont **complémentaires**!

TABLEAU – Scores de MAP sur la collection NTCIR-2.

Évaluation de la qualité des mots-clés par une tâche applicative

Problématique

Évaluation intrinsèque peu fiable.

Introduction d'un nouveau cadre d'**évaluation extrinsèque** fondé sur une tâche de recherche documentaire.

Conclusion

- Mots-clés produits automatiquement sont **utiles** même **en complément** de mots-clés annotés manuellement.
- Seules les méthodes **génératives** sont **assez performantes** pour impacter significativement la recherche documentaire.

Conclusion

- **Évaluation stricte** des méthodes de l'état de l'art réalisée grâce à la création du jeu de données **KPTimes**.
- Comparaison directe des méthodes impossible avant cette étude.
- Méthodes génératives peu généralisables à d'autres genres de documents et annotation.
- Sans données d'entraînement les **méthodes de base** sont toujours compétitives.
- Avec données d'entraînement les **méthodes génératives** sont l'état de l'art.
- Évaluation par **appariement exact** à une référence peu fiable. Confirme l'évaluation manuelle réalisée par Bougouin (2015).

- **Évaluation extrinsèque** pour étudier la qualité des mots-clés dans un **cadre applicatif**.
- Mots-clés produits par les récentes méthodes génératives sont **assez qualitatifs** pour améliorer une tâche de recherche documentaire contrairement aux méthodes extractives.

Perspectives

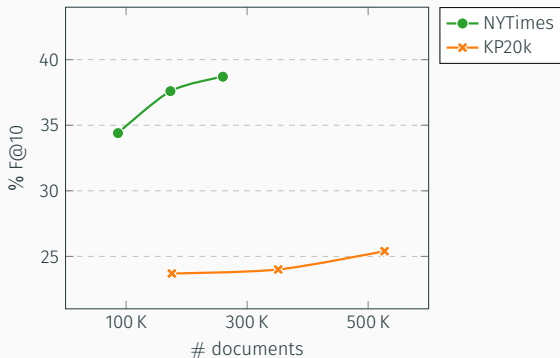
- **Évaluer les méthodes génératives plus récentes** pour mesurer l'impact de leurs améliorations incrémentales sur la recherche documentaire.
- **Jeux de données en français** pour transposer nos expériences à cette langue. Peu pertinent pour l'informatique mais pertinent pour les sciences sociales par exemple.
- **Cohérence des mots-clés produits** pour aider à la navigation des bibliothèques.

- Les mots-clés sont considérés comme une **fin en soi**.
- **Redéfinition de la tâche** : mots-clés pour la navigation, pour la RI, pour la catégorisation, pour la création de thésaurus, ...
- Les mots-clés sont assez qualitatifs, comment les intégrer aux bibliothèques, systèmes de RI ?

Merci pour votre attention

Questions?

Impact de l'annotation sur l'entraînement



- Entraînement de CopyRNN avec 33%, 66% et 100% des jeux de données KP20k et KPTimes.
- L'ajout de document a KP20k ne permet pas d'augmenter significativement les performances car l'annotation est peu cohérente; contrairement aux mots-clés de KPTimes.

Document similaire, type d'annotation différent

Les performances de CopyRNN sont-elles généralisables à un type d'annotation différent ?

	Éditeur NYTimes	Éditeur JPTimes	Lecteur DUC-2001
F ₀ 10			
TF×IDF	9,6	15,1	23,0
MPRank	11,2	16,8	25,3
Kea	11,0	16,6	26,2
CopyNews	39,3	24,6	10,5
%abs.	52,5	24,2	3,1

- CopyNews connaît une première baisse de performances lors de l'évaluation sur JPTimes.
- CopyNews généralise mal à un type d'annotation différent.
- Méthodes extractives désavantagées par le taux de mots-clés absent

Document différent, annotation différente, absents

%	KPTimes		KP20k	
	Prs	Abs	Prs	Abs
CopyNews	61,2	38,8	51,1	48,9
CopySci	94,0	5,1	92,0	8,0

TABLEAU – Pourcentage de mots-clés présents et absents.

- CopySci produit presque exclusivement des mots-clés présents.
- Il est plus **risqué** de produire des **mots-clés absents** en situation de **généralisation**.

Validation du choix des méthodes

Système	MAP	P@10
BM25+RM3	35,5	38,9
QL+RM3	34,4	36,1
1 ^{er} (Fujita and Corporation, 2001)	31,9	37,4
BM25	31,9	37,1
2 nd (Murata et al., 2001)	31,3	36,1
QL	31,2	35,1
3 ^{èm} (Chen et al., 2001)	26,2	33,9

Scores des meilleurs systèmes de la compétition NTCIR-2.

Impact des mots-clés seuls

	BM25	+RM3
T+R	-	-
+ Kea (KP20k)	13,66	11,59
+ MPRank	13,68	13,48
+ CopyRNN	16,61	17,02
+ CorrRNN	15,54	15,63
T+R+M	15,51	16,00
+ Kea (KP20k)	21,97	25,68
+ MPRank	22,42	25,78
+ CopyRNN	21,91	27,32
+ CorrRNN	21,61	27,39

Impact des mots-clés seuls (cont'd)

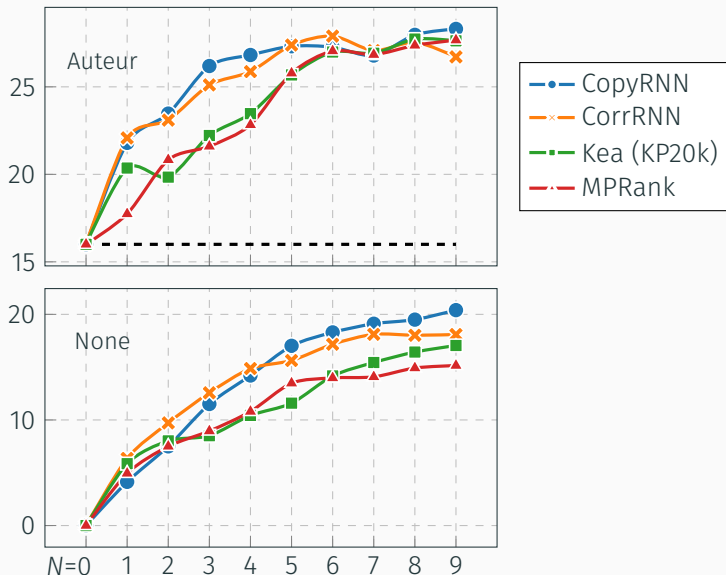


Figure – Scores de MAP pour BM25+RM3 sur NTCIR-2 en fonction du nombre

Catégorisation PRMN

Study on the Structure of Index Data for Metasearch System (id : gakkai-e-0001384947)

This paper proposes a new technique for Metasearch system, which is based on the grouping of both keywords and URLs. This technique enables metasearch systems to share information and to reflect the estimation of users' preference. With this system, users can search not only by their own keywords but by similarity of HTML documents. In this paper, we describe the principle of the grouping technique as well as the summary of the existing search systems.

Mots-clés présent : Metasearch – Search System

Mots-clés absent :

Information	Sharing	–	Information	Retrieval	–	User's	Behavior	–	Retrieval Support
<u>R</u> éordonné				<u>M</u> ixte			<u>M</u> ixte		<u>N</u> on-vu

- La définition actuelle des mots-clés **présents** et **absents** n'est pas pertinente pour les systèmes de RI
- Pour distinguer les mots-clés qui vont modifier la **pondération** de ceux qui vont l'**étendre**.

PRMN : Impact des mots-clés de référence

MAP	Bm25	+RM3	#mc
T+R	29,6	32,8	-
Pond. (P+R)	30,6 [†] 1,1	33,8 1,0	3,3
Exp. (M+N)	30,8[†] 1,3	34,3 1,5	1,5
+ P+R+M+N	31,9 ^{†‡} 2,3	35,5 ^{†‡} 2,7	4,7

- Toutes les catégories de mots-clés augmentent les résultats
- Les mots-clés qui **étendent** le document sont à l'origine de la majorité des gains de score.

PRMN : Impact des mots-clés produits automatiquement

MAP BM25+RM3	T+R		T+R+M		F@5
	MAP	#mc	MAP	#mc	
-	32,8	0,0	35,5	4,8	
CorrRNN	35,0[†] 2,2	5,0	36,9[†] 1,4	9,7	22,1
Pond. (P+R)	34,6 [†] 1,8	5,0	36,7 1,2	9,7	25,5
Exp. (M+U)	33,4 0,6	1,9	35,8 0,3	5,2	1,7
CopyRNN	34,8 [†] 2,0	5,0	37,1 [†] 1,6	9,7	24,0
Pond. (P+R)	35,0[†] 2,2	5,0	37,5[†] 2,0	9,7	27,6
Exp. (M+U)	32,2 -0,6	3,2	34,5 -1,0	6,8	0,7

- Les mots-clés absents ne sont pas assez qualitatifs pour améliorer les scores de MAP (cf. F@5).
- Ils dégradent même les scores pour CopyRNN

- Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Smucker, M. D., and Wade, C. (2004). UMass at TREC 2004 : Novelty and HARD :. Technical report, Defense Technical Information Center, Fort Belvoir, VA.
- Amar, M. (1997). *Les fondements théoriques de l'indexation : une approche linguistique*. Thèse de doctorat, Université Lumière, Lyon, France.
- Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., and Jaggi, M. (2018). Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 221–229, Brussels, Belgium. Association for Computational Linguistics.

- Boudin, F. (2018). Unsupervised Keyphrase Extraction with Multipartite Graphs. In *Proceedings of NAACL-HLT 2018*. Association for Computational Linguistics.
- Bougouin, A. (2015). *Indexation automatique par termes-clés en domaines de spécialité*. These de doctorat, Nantes.
- Bougouin, A., Barreaux, S., Romary, L., Boudin, F., and Daille, B. (2016). TermITH-Eval : a French Standard-Based Resource for Keyphrase Extraction Evaluation. In *LREC - Language Resources and Evaluation Conference*, Potoroz, Slovenia.
- Bougouin, A., Boudin, F., and Daille, B. (2013). TopicRank : Graph-Based Topic Ranking for Keyphrase Extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 543–551, Nagoya, Japan.

- Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509 :257–289.
- Caragea, C., Bulgarov, F. A., Godea, A., and Das Gollapalli, S. (2014). Citation-Enhanced Keyphrase Extraction from Research Papers : A Supervised Approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, Doha, Qatar. Association for Computational Linguistics.
- Chan, H. P., Chen, W., Wang, L., and King, I. (2019). Neural Keyphrase Generation via Reinforcement Learning with Adaptive Rewards. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2163–2174, Florence, Italy. Association for Computational Linguistics.

- Chen, A., Gey, F. C., and Jiang, H. (2001). Berkeley at NTCIR-2 : Chinese, Japanese, and English IR Experiments. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, page 9.
- Chen, J., Zhang, X., Wu, Y., Yan, Z., and Li, Z. (2018). Keyphrase Generation with Correlation Constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.
- Chen, W., Chan, H. P., Li, P., Bing, L., and King, I. (2019a). An Integrated Approach for Keyphrase Generation via Exploring the Power of Retrieval and Extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and*

- Short Papers*), pages 2846–2856, Minneapolis, Minnesota. Association for Computational Linguistics.
- Chen, W., Chan, H. P., Li, P., and King, I. (2020). Exclusive Hierarchical Decoding for Deep Keyphrase Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1095–1105, Online. Association for Computational Linguistics.
- Chen, W., Gao, Y., Zhang, J., King, I., and Lyu, M. R. (2019b). Title-Guided Encoding for Keyphrase Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :6268–6275.
- Diao, S., Song, Y., and Zhang, T. (2020). Keyphrase Generation with Cross-Document Attention. *arXiv :2004.09800 [cs]*. arXiv : 2004.09800.

- Florescu, C. and Caragea, C. (2017). PositionRank : An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1105–1115, Vancouver, Canada. Association for Computational Linguistics.
- Fujita, S. and Corporation, J. (2001). Notes on the Limits of CLIR Effectiveness NTCIR-2 Evaluation Experiments at Justsystem. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, page 8.

- Gallina, Y., Boudin, F., and Daille, B. (2019). KPTimes : A Large-Scale Dataset for Keyphrase Generation on News Documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 130–135, Tokyo, Japan. Association for Computational Linguistics.
- Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing -*, volume 10, pages 216–223, Not Known. Association for Computational Linguistics.
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1) :11–21. Publisher : MCB UP Ltd.

- Kando, N. (2001). Overview of the Second NTCIR Workshop. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*.
- Kim, S. N., Medelyan, O., Kan, M.-Y., and Baldwin, T. (2010). SemEval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, pages 21–26, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Krapivin, M., Autaeu, A., and Marchese, M. (2009). Large Dataset for Keyphrases Extraction. Departmental Technical Report, University of Trento.

- Liu, Z., Huang, W., Zheng, Y., and Sun, M. (2010). Automatic Keyphrase Extraction via Topic Decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*, pages 366–376, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Marujo, L., Gershman, A., Carbonell, J., Frederking, R., and Neto, J. P. (2012). Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 399–403, Istanbul, Turkey. European Language Resources Association (ELRA).

- Medelyan, O., Frank, E., and Witten, I. H. (2009). Human-competitive Tagging Using Automatic Keyphrase Extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3 - Volume 3*, EMNLP '09, pages 1318–1327, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., and Chi, Y. (2017). Deep keyphrase generation. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, pages 582–592.
- Mihalcea, R. and Tarau, P. (2004). TextRank : Bringing Order into Texts. In *Proceedings of {EMNLP-04}and the 2004 Conference on Empirical Methods in Natural Language Processing*, page 8, Barcelona, Spain.

- Murata, M., Utiyama, M., Ma, Q., Ozaku, H., and Isahara, H. (2001). CRL at NTCIR2. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, page 11.
- Nguyen, T. D. and Kan, M.-Y. (2007). Keyphrase Extraction in Scientific Publications. In Goh, D. H.-L., Cao, T. H., Sølvsberg, I. T., and Rasmussen, E., editors, *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, volume 4822, pages 317–326. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title : Lecture Notes in Computer Science.
- Robertson, S. E., Walker, S., and Beaulieu, M. (1999). Okapi at TREC 7 : automatic ad hoc, ltering, VLC and interactive track. *Proceedings of the Seventh Text REetrieval Conference (TREC-7)*, 1999, page 12.

- Schutz, A. T. (2008). *Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods*. PhD thesis, National University of Ireland, Galway.
- Teneva, N. and Cheng, W. (2017). Saliency Rank : Efficient Keyphrase Extraction with Topic Modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 530–535, Vancouver, Canada. Association for Computational Linguistics.
- Thakur, N., Reimers, N., Rüchlé, A., Srivastava, A., and Gurevych, I. (2021). BEIR : A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models.

- Wan, X. and Xiao, J. (2008). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2, AAAI'08*, pages 855–860, Chicago, Illinois. AAAI Press.
- Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. (1999). KEA : practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries - DL '99*, pages 254–255, Berkeley, California, United States. ACM Press.
- Yang, P., Fang, H., and Lin, J. (2017). Anserini : Enabling the Use of Lucene for Information Retrieval Research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17*, pages 1253–1256, New York, NY, USA. Association for Computing Machinery.

- Yuan, X., Wang, T., Meng, R., Thaker, K., Brusilovsky, P., He, D., and Trischler, A. (2020). One Size Does Not Fit All : Generating and Evaluating Variable Number of Keyphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7961–7975, Online. Association for Computational Linguistics.
- Zhang, Y., Fang, Y., and Weidong, X. (2017). Deep keyphrase generation with a convolutional sequence to sequence model. In *2017 4th International Conference on Systems and Informatics (ICSAI)*, pages 1477–1485.
- Zhao, J. and Zhang, Y. (2019). Incorporating Linguistic Constraints into Keyphrase Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5224–5233, Florence, Italy. Association for Computational Linguistics.