Candidate Extraction Impact on Automatic Keyphrase Extraction

Adrien Bougouin and **Florian Boudin** and **Béatrice Daille**
Université de Nantes, LINA, France
{adrien.bougouin,florian.boudin,beatrice.daille}@univ-nantes.fr

Abstract

8+2 pages maximum...

1 Introduction

Keyphrases are single or multi-word expressions that represent the main topics of a document. Keyphrases are useful in many tasks such as information retrieval (Medelyan and Witten, 2008), document summarization (Litvak and Last, 2008) or document clustering (Han et al., 2007). Although scientific articles usually provide them, most of the documents have no associated keyphrases. Therefore, the problem of automatically assigning keyphrases to documents is an active field of research.

Introduire free indexing et controlled indexing

Automatic keyphrase extraction methods are divided into two categories: supervised and unsupervised methods. Supervised methods typically recast keyphrase extraction as a binary classification task (Witten et al., 1999; Sujian et al., 2003; Eichler and Neumann, 2010). For unsupervised methods, keyphrase extraction is often considered as a ranking task and many approaches are used (Barker and Cornacchia, 2000; Tomokiyo and Hurst, 2003; Mihalcea and Tarau, 2004). As distinct as they are, both supervised and unsupervised methods rely on a preliminary candidate extraction step which identifies single and multi-word expressions that have the same syntactic properties than a keyphrase. These expressions are the only textual units that can be extracted as keyphrases.

In this paper, we focus on the candidate extraction step and show its impact on the performance of automatic keyphrase extraction. Various methods are commonly employed to extract keyphrase candidates[1]. Usually, a set of either single words, n-grams filtered by stop words, NP-chunks or sequences of words matching given patterns is extracted (Hulth, 2003). According to the chosen method, the extracted set contains more or less candidates, and the amount of these that match with the ground truth keyphrases may vary. Hence, a few questions arise. How the different sets influence the keyphrase extraction? Do large candidate sets introduce noise that affects the performance of some keyphrase extraction methods?

We seek to better understand the impact of candidate extraction methods on keyphrase extraction by studying the aforementioned questions. We first quantify the differences between the candidate sets obtained by the commonly used methods. Also, we propose to use another method developed to extract noun-phrases for document indexing (Evans and Zhai, 1996) and we argue that such term detection method (Castellví et al., 2001) provides solid keyphrase candidates. Then, we evaluate the impact of the candidate extraction methods on three dissimilar keyphrase extraction methods. We select KEA (Witten et al., 1999) to represent supervised methods, TF-IDF (Spärck Jones, 1972) to represent unsupervised methods that require a collection of documents and TopicRank (Bougouin et al., 2013) to represent unsupervised methods that only make use of the document to analyse.

Results show that...

2 Definition of candidate Keyphrases

Candidate keyphrases are textual units which can be selected as keyphrases for a document they are extracted from. Hence, they must have the same syntactic and linguistic properties than ground truth keyphrases. This section aims to determine those properties by analysing three standard

[1] In this work, we do not consider methods using a manually defined controlled vocabulary (dependent of the domain of the processed data).

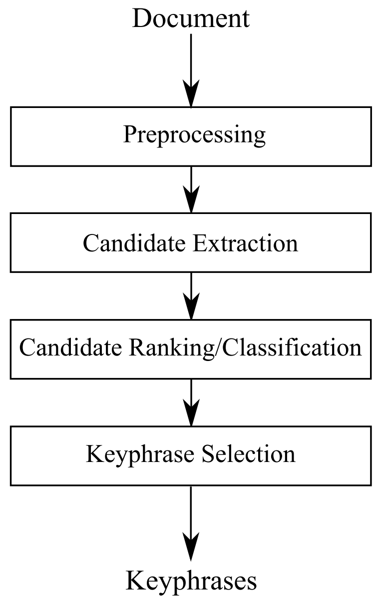

Figure 1: Processing steps of automatic keyphrase extraction methods.

evaluation datasets, for keyphrase extraction, and by providing statistics about reference keyphrases (ground truth keyphrases).

2.1 Keyphrase extraction datasets

Keyphrase extraction datasets are used to evaluate or train keyphrase extraction methods. Hence, the datasets are collections of documents paired with reference keyphrases, given by authors, readers or both.

Présentation générale des corpus pour l'extraction de termes-clés.

Présentation des corpus qui seront utilisés

2.2 Keyphrase analysis

Donner les séquences de POS les plus fréquentes dans le gold standard.

3 Candidate Extraction

Objectif + pré-requis.

3.1 N-Gram Extraction

3.2 NP-Chunk Extraction

3.3 Pattern Matching

3.4 Term Extraction

4 Keyphrase Extraction

Fonctionnement général.

4.1 TF-IDF

4.2 TopicRank

4.3 KEA

5 Evaluation

Expliquer les deux évaluations: intrinsèque et extrinsèque.

5.1 Experimental Setting

5.2 Candidate Extraction

Donner le rappel max et comparer avec la taille des différents ensemble.

Methods	DUC		SemEval	
	Candidates	Rmax	Candidates	Rmax
1-grams				
2-grams				
3-grams				
4-grams				
5-grams				
Chunks				
Patterns1				
Patterns2				
Terms				

Table 2: Candidate extraction statistics.

Quels sont les termes candidats communs aux ensembles, les propriétés ?

5.3 Keyphrase Extraction

Quelles sont les performances de chaque méthode avec chaque ensemble de termes candidats ?

References

Ken Barker and Nadia Cornacchia. 2000. Using Noun Phrase Heads to Extract Document Keyphrases. In *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 40–52, London, UK. Springer-Verlag.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-Based Topic Ranking for Keyphrase Extraction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, Nagoya, Japan, October.

M. Teresa Cabré Castellví, Rosa Estopà Bagot, and Jordi Vivaldi Palatresi. 2001. Automatic Term Detection: A Review of Current Systems. *Recent Advances in Computational Terminology*, pages 53–88.

Kathrin Eichler and Günter Neumann. 2010. DFKI KeyWE: Ranking Keyphrases Extracted from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 150–153, Stroudsburg, PA, USA. Association for Computational Linguistics.

David A. Evans and Chengxiang Zhai. 1996. Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. In *Proceedings of the 34th annual meeting of the Association for Computational Linguistics*, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Juhyun Han, Taehwan Kim, and Joongmin Choi. 2007. Web Document Clustering by Using Automatic Keyphrase Extraction. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 56–59, Washington, DC, USA. IEEE Computer Society.

Anette Hulth. 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marina Litvak and Mark Last. 2008. Graph-Based Keyword Extraction for Single-Document Summarization. In *Proceedings of the Workshop on Multi-Source Multilingual Information Extraction and Summarization*, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Olena Medelyan and Ian H. Witten. 2008. Domain-Independent Automatic Keyphrase Indexing with Small Training Sets. *Journal of the American Society for Information Science and Technology*, 59(7):1026–1040, may.

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

Karen Spärck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–21.

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

Takashi Tomokiyo and Matthew Hurst. 2003. 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. Association for Computational Linguistics.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill Manning. 1999. KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of the 4th ACM Conference on Digital Libraries*, pages 254–255, New York, NY, USA. ACM.

Methods	DUC			SemEval		
	P	R	F	P	R	F
TF-IDF						
1-grams	00.0	00.0	00.0	00.0	00.0	00.0
2-grams	00.0	00.0	00.0	00.0	00.0	00.0
3-grams	00.0	00.0	00.0	00.0	00.0	00.0
4-grams	00.0	00.0	00.0	00.0	00.0	00.0
5-grams	00.0	00.0	00.0	00.0	00.0	00.0
Chunks	00.0	00.0	00.0	00.0	00.0	00.0
Patterns1	00.0	00.0	00.0	00.0	00.0	00.0
Patterns2	00.0	00.0	00.0	00.0	00.0	00.0
Terms	00.0	00.0	00.0	00.0	00.0	00.0
TopicRank						
1-grams	00.0	00.0	00.0	00.0	00.0	00.0
2-grams	00.0	00.0	00.0	00.0	00.0	00.0
3-grams	00.0	00.0	00.0	00.0	00.0	00.0
4-grams	00.0	00.0	00.0	00.0	00.0	00.0
5-grams	00.0	00.0	00.0	00.0	00.0	00.0
Chunks	00.0	00.0	00.0	00.0	00.0	00.0
Patterns1	00.0	00.0	00.0	00.0	00.0	00.0
Patterns2	00.0	00.0	00.0	00.0	00.0	00.0
Terms	00.0	00.0	00.0	00.0	00.0	00.0
KEA						
1-grams	00.0	00.0	00.0	00.0	00.0	00.0
2-grams	00.0	00.0	00.0	00.0	00.0	00.0
3-grams	00.0	00.0	00.0	00.0	00.0	00.0
4-grams	00.0	00.0	00.0	00.0	00.0	00.0
5-grams	00.0	00.0	00.0	00.0	00.0	00.0
Chunks	00.0	00.0	00.0	00.0	00.0	00.0
Patterns1	00.0	00.0	00.0	00.0	00.0	00.0
Patterns2	00.0	00.0	00.0	00.0	00.0	00.0
Terms	00.0	00.0	00.0	00.0	00.0	00.0

Table 3: Comparison of TF-IDF, TopicRank and KEA, when using various candidate extraction methods and when extracting 10 keyphrases.

	Statistics	Corpora		
		DUC	SemEval	DEFT
Documents	Language	English	English	French
	Type	News	Papers	Papers
	Documents	208	144	141
	Tokens/document		5134.6	7276.7
	Keyphrases/document	8.1	15.4	5.4
	Missings keyphrases		13.5%	18.2%
Keyphrases	Unigrams	26.2%	20.2%	66.4%
	Bigrams	54.1%	53.4%	20.7%
	Trigrams and more	19.7%	26.4%	12.9%
	Containing nouns	99.5%	98.8%	95.5%
	Containing adjectives	41.6%	40.5%	28.8%
	Containing verbs	0.9%	3.4%	0.5%
	Containing adverbs	1.3%	0.6%	0.5%
	Containing prepositions	0.2%	1.2%	12.7%
	Containing determiners	0.0%	0.0%	8.1%
	Containing others	1.3%	2.1%	5.8%

Table 1: Dataset statistics. The missing keyphrase percentage is determined based on the stemmed form of the gold standard keyphrases.