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.

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; Mihalcea and Tarau, 2004). Although they work diferently, 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. Therefore, we believe that the extraction of candidate keyphrases plays a direct role in automatic keyphrase extraction.

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 are actual 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 and we propose to use other methods developed for automatic term detection (Castellví et al., 2001; Evans and Zhai, 1996) to show that such methods provide solid keyphrase candidates. Then, we evaluate the impact of the candidate extraction methods over 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 analyzed document.

Results show that...

2 Definition of Candidate Keyphrases

Candidate keyphrases are textual units which can be selected as keyphrases of a document. 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 evaluation datasets, for keyphrase extraction, and by providing statistics about their reference keyphrases (ground truth

[1] In this work, we do not consider methods which use a manually defined controlled vocabulary.

keyphrases).

2.1 Keyphrase Extraction Datasets

Keyphrase extraction datasets are used to train or evaluate keyphrase extraction methods. Hence, they are collections of documents paired with reference keyphrases given by authors, readers or both. Unlike the methods to automatically extract keyphrases, human annotators do not only provide keyphrases contained into the documents. This problem of missing keyphrases leads to a bias of the training or evaluation of keyphrase extraction methods. In this work, we use three standard datasets which differ in terms of document size, type and language. The problem of missing keyphrases is partially bypassed using stemmed forms when comparison between reference and candidate keyphrases is needed.

The **DUC** dataset (Over, 2001) is a collection of 308 English news articles covering about 30 topics (e.g. tornadoes, gun control, etc.). This collection is the test dataset of the DUC-2001 summarization evaluation campaign. This part of DUC-2001 is the only one that contains keyphrases, annotated by Wan and Xiao (2008). We split the collection into two sets: a training set containing 208 documents and a test set containing 100 documents.

The **SemEval** dataset (Kim et al., 2010) contains 284 English papers collected from the ACM Digital Libraries (conference and workshop papers). The papers are divided into three sets: a trial set containing 40 documents (unused in this work), a training set containing 144 documents and a test set containing 100 documents. As for the associated keyphrases, these are provided by both authors and readers.

The **DEFT** dataset (Paroubek et al., 2012) is a collection of 234 French scientific papers belonging to the Humanities and Social Sciences domain. DEFT is divided into two sets: a training set containing 141 documents and a test set containing 93 documents. The only available reference keyphrases are the ones given by authors.

Table 1 shows the statistics about the three datasets. As these statistics are used to guide our work, we restrain them to the training sets. As said before, the datasets differ in terms of size, type and language. Moreover, it is worth noticing that the number of keyphrases, the ratio of missing ones and the average number of tokens per keyphrase differ too. This observation is due to the fact that there is not a unique methodology (guideline) to associate keyphrases to a document. To better fit the requirements, such guidelines should not only be used by human annotators, but also by automatic keyphrase extraction methods.

[CAPTION] Others are mainly foreign words and coordinating conjonctions.

2.2 Reference Keyphrases Analysis

Despite the fact that the data are not homogeneous, this section aims to determine the syntactic properties of most keyphrases, for English (intersecting information from DUC and SemEval training sets) and for French (using DEFT training set).

- About 80% of reference keyphrases contain only one or two words.
- Toutes les keyphrases de référence ont étées POS tagguées automatiquement, puis vérifiées manuellement, afin d'obtenir les stats du tableau 1.
- Almost every keyphrases contain nouns or proper nouns.
- Adjective modifiers are almost the only other compounds of keyphrases, except for French where prepositionnal phrases are used.

DONNER DES EXEMPLES

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

English: NOUN NOUN (huricane expert – AP880409-0015); ADJ NOUN (turbulent summer – AP880409-0015); NOUN (storms – AP880409-0015); ADJ NOUN NOUN (annual huricane forecast – AP880409-0015); NOUN NOUN NOUN (huricane reconnaissance – AP890529-0030).

French: NOUN (patrimoine – as_2002_007048ar); NOUN ADJ (tradition orale – as_2002_007048ar); PROPER NOUN (Indonésie – as_2001_000235ar); NOUN PREP DET NOUN (conservation de la nature – as_2005_011742ar); NOUN PREP NOUN (changement de terrain – as_2001_000260ar).

3 Candidate Extraction

Objectif + pré-requis.

	Statistics	Corpora		
		DUC	SemEval	DEFT
Documents	Language	English	English	French
	Type	News	Papers	Papers
	Documents	208	144	141
	Tokens/document	912.0	5134.6	7276.7
	Keyphrases/document	8.1	15.4	5.4
	Missings keyphrases	3.9%	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	90.8%	95.9%	79.3%
	Containing proper nouns	18.7%	5.8%	16.8%
	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: Training dataset statistics. As a matter of consistency regarding the training and the evaluation of keyphrase extraction methods, the percentage of missing keyphrases is determined based on the stemmed form of the reference keyphrases.

3.1 N-Gram Extraction

3.2 NP-Chunk Extraction

English: (PROPER NOUN+) | (ADJ+ NOUN) | (NOUN+)

French: (PROPER NOUN+) | (ADJ? NOUN ADJ+) | (ADJ NOUN) | (NOUN+)

3.3 Pattern Matching

Longest NP: (NOUN | ADJ)+

English: (NOUN{1, 3}) | (ADJ NOUN{1, 2}) | ((NOUN | ADJ) ADJ NOUN) | (PROPER NOUN (PROPER NOUN | NOUN)?)

French: (NOUN (PREP DET? NOUN)? ADJ?) | (PROPER NOUN+)

3.4 Term Extraction

4 Keyphrase Extraction

Fonctionnement général.

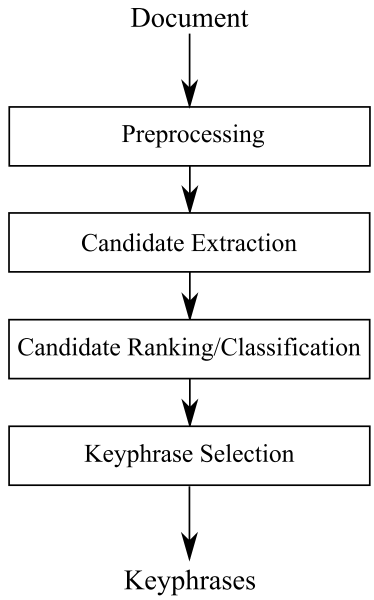

Figure 1: Processing steps of automatic keyphrase extraction methods.

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.

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.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 task 5: Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, 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.

Paul Over. 2001. Introduction to DUC-2001: an Intrinsic Evaluation of Generic News Text Summarization Systems. In *Proceedings of DUC 2001 Document Understanding Conference*.

Patrick Paroubek, Pierre Zweigenbaum, Dominic Forest, and Cyril Grouin. 2012. Indexation libre et contrôlée d'articles scientifiques. Présentation et résultats du défi fouille de textes DEFT2012 (Controlled and Free Indexing of Scientific Papers. Presentation and Results of the DEFT2012 Text-Mining Challenge) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Workshop DEFT 2012: DÉfi Fouille de Textes (DEFT 2012 Workshop: Text Mining Challenge)*, pages 1–13, Grenoble, France, June. ATALA/AFCP.

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.

Statistics	**Corpora**		
	DUC	SemEval	DEFT
Language	English	English	French
Type	News	Papers	Papers
Documents	100	100	93
Tokens/document	877.3	5177.7	6839.4
Keyphrases/document	7.94	14.7	5.2
Tokens/keyphrase	2.1	2.1	1.6
Missings keyphrases	2.8%	22.1%	21.1%

Table 2: Test dataset statistics. As a matter of consistency regarding the training and the evaluation of keyphrase extraction methods, the percentage of missing keyphrases is determined based on the stemmed form of the reference keyphrases.

Xiaojun Wan and Jianguo Xiao. 2008. Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, pages 855–860. AAAI Press.

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		DEFT	
	Candidates	Rmax	Candidates	Rmax	Candidates	Rmax
{1..2}-grams	96730	78.8	393237	61.4	524827	67.3
{1..3}-grams	160622	94.0	751424	73.1	990612	74.1
{1..4}-grams	222644	95.8	1118125	75.3	1471489	78.2
{1..5}-grams	281962	96.3	1476654	75.9	1946856	78.5
NP chunks	14734	75.6	53480	56.3	74278	63.0
Longest NPs	15559	88.7	64649	62.4	85047	61.1
Best patterns	10130	38.0	36786	39.0	5331	11.8
Subcompounds	17181	90.6	71224	64.4	86866	61.1
TermSuite	16253	46.1	50636	32.4	82884	53.4

Table 3: Candidate extraction statistics. Rmax stands for maximum recall, i.e. it is the percentage of candidates that match with reference keyphrases.

Methods	DUC			SemEval			DEFT		
	P	R	F	P	R	F	P	R	F
{1..2}-grams	9.6	13.2	11.0	16.2	11.4	13.3	11.6	21.8	15.0
{1..3}-grams	9.0	12.4	10.3	16.4	11.5	13.4	11.6	22.0	15.0
{1..4}-grams	9.0	12.4	10.3	16.4	11.5	13.4	11.6	21.9	15.0
{1..5}-grams	8.6	11.9	9.8	16.4	11.5	13.4	11.6	22.0	15.0
NP chunks	12.3	16.9	14.1	19.4	13.7	15.9	12.9	24.0	16.6
Longest NPs	13.3	18.2	15.2	18.3	12.9	15.0	12.8	23.6	16.4
Best patterns	11.3	16.0	13.0	17.8	12.7	14.7	4.9	8.4	6.1
Subcompounds	13.1	17.8	14.9	18.1	12.8	14.9	12.8	23.6	16.4
TermSuite	11.5	16.0	13.2	12.9	9.3	10.7	12.5	23.5	16.1

Table 4: Comparison of candidate extraction methods, when extracting 10 keyphrases with the **TF-IDF** method. Results are expressed as a percentage of precision (P), recall (R) and f-score (F).

Methods	DUC			SemEval			DEFT		
	P	R	F	P	R	F	P	R	F
{1..2}-grams	3.9	5.4	4.5						
{1..3}-grams	2.0	2.7	2.3						
{1..4}-grams	2.5	3.7	2.9						
{1..5}-grams									
NP chunks	16.1	21.1	18.0	15.7	10.8	12.7			
Longest NPs	17.7	23.2	19.8						
Best patterns	10.5	14.6	12.0	12.6	8.9	10.3			
Subcompounds	18.3	24.0	20.5						
TermSuite	10.2	13.7	11.5	9.0	6.6	7.5			

Table 5: Comparison of candidate extraction methods, when extracting 10 keyphrases with **TopicRank**. Results are expressed as a percentage of precision (P), recall (R) and f-score (F).

Methods	DUC			SemEval			DEFT		
	P	R	F	P	R	F	P	R	F
{1..2}-grams									
{1..3}-grams									
{1..4}-grams									
{1..5}-grams									
NP chunks									
Longest NPs									
Best patterns									
Subcompounds									
TermSuite									

Table 6: Comparison of candidate extraction methods, when extracting 10 keyphrases with **KEA**. Results are expressed as a percentage of precision (P), recall (R) and f-score (F).