Unsupervised Keyphrase Extraction with Multipartite Graphs

Florian Boudin

Unsupervised Keyphrase Extraction with Multipartite Graphs

Florian Boudin
LS2N, Université de Nantes, France
florian.boudin@univ-nantes.fr

Abstract

We propose an unsupervised keyphrase extraction model that encodes topical information within a multipartite graph structure. Our model represents keyphrase candidates and topics in a single graph and exploits their mutually reinforcing relationship to improve candidate ranking. We further introduce a novel mechanism to incorporate keyphrase selection preferences into the model. Experiments conducted on three widely used datasets show significant improvements over state-of-the-art graph-based models.

1 Introduction

Recent years have witnessed a resurgence of interest in automatic keyphrase extraction, and a number of diverse approaches were explored in the literature (Kim et al., 2010; Hasan and Ng, 2014; Gollapalli et al., 2015; Augenstein et al., 2017). Among them, graph-based approaches are appealing in that they offer strong performance while remaining completely unsupervised. These approaches typically involve two steps: 1) building a graph representation of the document where nodes are lexical units (usually words) and edges are semantic relations between them; 2) ranking nodes using a graph-theoretic measure, from which the top-ranked ones are used to form keyphrases.

Since the seminal work of Mihalcea and Tarau (2004), researchers have devoted a substantial amount of effort to develop better ways of modelling documents as graphs. Most if not all previous work, however, focus on either measuring the semantic relatedness between nodes (Wan and Xiao, 2008; Tsatsaronis et al., 2010) or devising node ranking functions (Tixier et al., 2016; Florescu and Caragea, 2017). So far, little attention has been paid to the use of different types of graphs. Yet, a key challenge in keyphrase extraction is to ensure topical coverage and diversity, which are not naturally handled by graph-of-words representations (Hasan and Ng, 2014).

Most attempts at using topic information in graph-based approaches involve biasing the ranking function towards topic distributions (Liu et al., 2010; Zhao et al., 2011; Zhang et al., 2013). Unfortunately, these models suffer from several limitations: they aggregate multiple topic-biased rankings which makes their time complexity prohibitive for long documents[1], they require a large dataset to estimate word-topic distributions that is not always available or easy to obtain, and they assume that topics are independent of one another, making it hard to ensure topic diversity. For the latter case, supervised approaches were proposed to optimize the broad coverage of topics (Bougouin et al., 2016; Zhang et al., 2017).

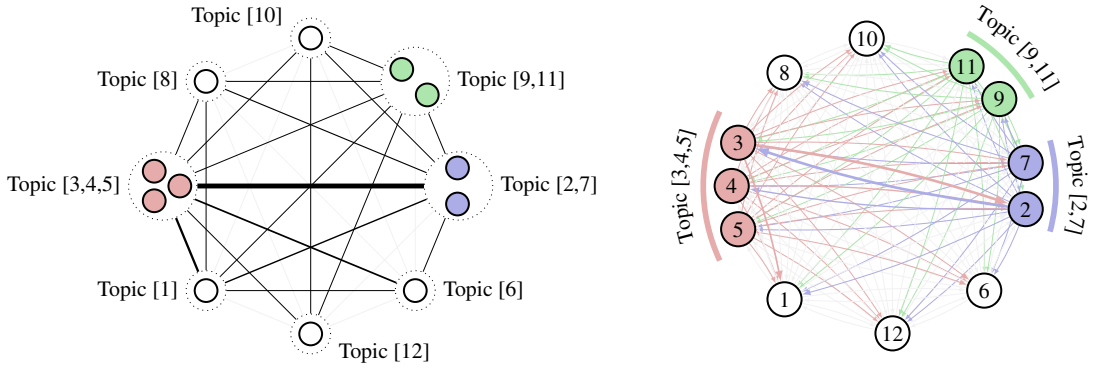
Another strand of work models documents as graphs of topics and selects keyphrases from the top-ranked ones (Bougouin et al., 2013). This higher level representation (see Figure 1a), in which topic relations are measured as the semantic relatedness between the keyphrase candidates they instantiate, was shown to improve the overall ranking and maximize topic coverage. The downside is that candidates belonging to a single topic are viewed as equally important, so that post-ranking heuristics are required to select the most representative keyphrase from each topic. Also, errors in forming topics propagate throughout the model severely impacting its performance.

Here, we build upon this latter line of work and propose a model that implicitly enforces topical diversity while ranking keyphrase candidates in a

[1] Recent work showed that comparable results can be achieved by computing a single topic specificity weight value for each word (Sterckx et al., 2015; Teneva and Cheng, 2017).

Proceedings of NAACL-HLT 2018, pages 667–672
New Orleans, Louisiana, June 1 - 6, 2018.

Inverse problems [1] for a **mathematical model** [2] of **ion exchange** [3] in a **compressible ion exchanger** [4]
A **mathematical model** [2] of **ion exchange** [3] is considered, allowing for **ion exchanger compression** [5] in the **process** [6] of **ion exchange** [3]. Two **inverse problems** [1] are investigated for this **model** [7], **unique solvability** [8] is proved, and **numerical solution methods** [9] are proposed. The **efficiency** [10] of the proposed **methods** [11] is demonstrated by a **numerical experiment** [12].



(a) TopicRank graph.

(b) Multipartite graph.

Figure 1: Comparison between TopicRank (Bougouin et al., 2013) and our multipartite graph representation for document 2040.abstr from the Hulth-2003 dataset. Nodes are topics (left) or keyphrase candidates (right), and edges represent co-occurrence relations.

single operation. To do this, we use a particular graph structure, called multipartite graph, to represent documents as tightly connected sets of topic related candidates (see Figure 1b). This representation allows for the seamless integration of any topic decomposition, and enables the ranking algorithm to make full use of the mutually reinforcing relation between topics and candidates.

Another contribution of this work is a mechanism to incorporate intra-topic keyphrase selection preferences into the model. It allows the ranking algorithm to go beyond semantic relatedness by leveraging information from additional salience features. Technically, keyphrase candidates that exhibit certain properties, e.g. that match a thesaurus entry or occur in specific parts of the document, are promoted in ranking through edge weight adjustments. Here, we show the effectiveness of this mechanism by introducing a bias towards keyphrase candidates occurring first in the document.

2 Proposed Model

Similar to previous work, our model operates in two steps. We first build a graph representation of the document (§2.1), on which we then apply a ranking algorithm to assign a relevance score to each keyphrase (§2.3). We further introduce an in-between step where edge weights are adjusted to capture position information (§2.2).

For direct comparability with Bougouin et al. (2013), which served as the starting point for the work reported here, we follow their setup for identifying keyphrase candidates and topics. Keyphrase candidates are selected from the sequences of adjacent nouns with one or more preceding adjectives (`/Adj*Noun+/`). They are then grouped into topics based on the stem forms of the words they share using hierarchical agglomerative clustering with average linkage. Although simple, this method gives reasonably good results. There are many other approaches to find topics, including the use of knowledge bases or unsupervised probabilistic topic models. Here, we made the choice not to use them as they are not without their share of issues (e.g. limited coverage, parameter tuning), and leave this for future work.

2.1 Multipartite graph representation

A complete directed multipartite graph is built, in which nodes are keyphrase candidates that are connected only if they belong to different topics. Again, we follow (Bougouin et al., 2013) and weight edges according to the distance between two candidates in the document. More formally, the weight w_{ij} from node i to node j is computed as the sum of the inverse distances between the occurrences of candidates c_i and c_j:

$$w_{ij} = \sum_{p_i \in \mathcal{P}(c_i)} \sum_{p_j \in \mathcal{P}(c_j)} \frac{1}{|p_i - p_j|} \quad (1)$$

where $\mathcal{P}(c_i)$ is the set of the word offset positions of candidate c_i. This weighting scheme achieves comparable results to window-based co-occurrence counts without any parameter tuning.

The resulting graph is a complete k-partite graph, whose nodes are partitioned into k different independent sets, k being the number of topics. As exemplified in Figure 1, our graph representation differs from the one of (Bougouin et al., 2013) in two significant ways. First, topics are encoded by partitioning candidates into sets of unconnected nodes instead of being subsumed in single nodes. Second, edges are directed which, as we will see in §2.2, allows to further control the incidence of individual candidates on the overall ranking.

The proposed representation makes no assumptions about how topics are obtained, and thus allows direct use of any topic decomposition. It implicitly promotes the number of topics covered in the selected keyphrases by dampening intra-topic recommendation, and captures the mutually reinforcing relationship between topics and keyphrase candidates. In other words, removing edges between candidates belonging to a single topic ensures that the overall recommendation of each topic is distributed throughout the entire graph. Also, a benefit of encoding topic related candidates differentially is that the ones that best underpin each topic are directly given by the model.

2.2 Graph weight adjustment mechanism

Selecting the most representative keyphrase candidates for each topic is a difficult task, and relying only on their importance in the document is not sufficient (Hasan and Ng, 2014). Among the features proposed to address this problem in the literature, the position of the candidate within the document is most reliable. In order to capture this in our model, we adjust the incoming edge weights of the nodes corresponding to the first occurring candidate of each topic.

More formally, candidates that occur at the beginning of the document are promoted according to the other candidates belonging to the same topic. Figure 2 gives an example of applying graph weight adjustment for promoting a given candidate. Note that the choice of the candidates to promote, i.e. the selection heuristic, can be adapted to fit other needs such as prioritising candidates from a thesaurus.

Incoming edge weights for the first occurring candidate of each topic are modified by the following equation:

$$w_{ij} = w_{ij} + \alpha \cdot e^{\left(\frac{1}{p_i}\right)} \cdot \sum_{c_k \in \mathcal{T}(c_j) \setminus \{c_j\}} w_{ki} \quad (2)$$

where w_{ij} is the edge weight between nodes c_i and c_j, $\mathcal{T}(c_j)$ is the set of candidates belonging to the same topic as c_j, p_i is the offset position of the first occurrence of candidate c_i, and α is a hyper-parameter that controls the strength of the weight adjustment.

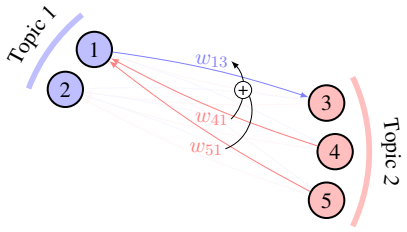


Figure 2: Illustration of the graph weight adjustment mechanism. Here, node 3 is promoted by increasing the weight of its incoming edge according to the outgoing edge weights of nodes 4 and 5.

2.3 Ranking and extraction

After the graph is built, keyphrase candidates are ordered by a graph-based ranking algorithm, and the top N are selected as keyphrases. Here, we adopt the widely used TextRank algorithm (Mihalcea and Tarau, 2004) in the form in which it leverages edge weights:

$$S(c_i) = (1 - \lambda) + \lambda \cdot \sum_{c_j \in \mathcal{I}(c_i)} \frac{w_{ij} \cdot S(c_j)}{\sum_{c_k \in \mathcal{O}(c_j)} w_{jk}} \quad (3)$$

where $\mathcal{I}(c_i)$ is the set of predecessors of c_i, $\mathcal{O}(c_j)$ is the set of successors of c_j, and λ is a damping factor set to 0.85 as in (Mihalcea and Tarau, 2004). Note that other ranking algorithms can be applied. We use TextRank because it was shown to perform consistently well (Boudin, 2013).

3 Experiments

3.1 Datasets and evaluation measures

We carry out our experiments on three datasets:

SemEval-2010 (Kim et al., 2010), which is composed of scientific articles collected from the ACM Digital Library. We use the set of combined author- and reader-assigned keyphrases as reference keyphrases.

Model	SemEval-2010			Hulth-2003			Marujo-2012		
	F_1@5	F_1@10	MAP	F_1@5	F_1@10	MAP	F_1@5	F_1@10	MAP
(Bougouin et al., 2013)	9.7	12.3	7.3	25.3	29.3	24.3	12.1	17.6	14.6
(Sterckx et al., 2015)	9.3	10.5	7.4	21.9	30.2	25.3	11.7	16.4	16.1
(Florescu and Caragea, 2017)	10.6	12.2	8.9	23.5	30.3	26.0	10.9	17.2	16.1
Proposed model	**12.2**†	**14.5**†	**11.8**†	**25.9**†	**30.6**	**29.2**†	**12.5**	**18.2**	**17.2**†
w/o weight adjustment	8.8	12.4	9.4	21.1	26.8	25.2	12.2	17.8	16.9

Table 1: F_1-scores computed at the top 5, 10 extracted keyphrases and Mean Average Precision (MAP) scores. † indicate significance at the 0.05 level using Student's t-test.

Hulth-2003 (Hulth, 2003), which is made of paper abstracts about computer science and information technology. Reference keyphrases were assigned by professional indexers.

Marujo-2012 (Marujo et al., 2012) that contains news articles distributed over 10 categories (e.g. Politics, Sports). Reference keyphrases were assigned by readers via crowdsourcing.

We follow the common practice and evaluate the performance of our model in terms of f-measure (F_1) at the top N keyphrases, and apply stemming to reduce the number of mismatches. We also report the Mean Average Precision (MAP) scores of the ranked lists of keyphrases.

3.2 Baselines and parameter settings

We compare the performance of our model against that of three baselines. The first baseline is TopicRank (Bougouin et al., 2013) which is the model that is closest to ours. The second baseline is Single Topical PageRank (Sterckx et al., 2015), an improved version of Liu et al. (2010) that biases the ranking function towards topic distributions inferred by Latent Dirichlet Allocation (LDA). The third baseline is PositionRank (Florescu and Caragea, 2017), a model that, like ours, leverages additional features (word's position and its frequency) to improve ranking accuracy.

Over-generation errors[2] are frequent in models that rank keyphrases according to the sum of the weights of their component words (Hasan and Ng, 2014; Boudin, 2015). This is indeed the case for the second and third baselines, and we partially address this issue by normalizing candidate scores by their length, as proposed in (Boudin, 2013). We use the parameters suggested by the authors for each model, and estimate LDA topic distributions on the training set of each dataset. Our model introduces one parameter, namely α, that controls the strength of the graph weight adjustment. This parameter is tuned on the training set of the SemEval-2010 dataset, and set to $\alpha = 1.1$ for all our experiments. For a fair and meaningful comparison, we use the same candidate selection heuristic (§2) across models.

3.3 Results

Results for the baselines and the proposed model are detailed in Table 1. Overall we observe that our model achieves the best results and significantly outperforms the baselines on most metrics. Relative improvements are smaller on the Hulth-2003 and Marujo-2012 datasets because they are composed of short documents, yielding a much smaller search space (Hasan and Ng, 2014). TopicRank obtains the highest precision among the baselines, suggesting that its –one keyphrase per topic– policy succeeds in filtering out topic-redundant candidates. On the other hand, TopicRank is directly affected by topic clustering errors as indicated by the lowest MAP scores, which supports the argument in favour of enforcing topical diversity implicitly. In terms of MAP, the best performing baseline is PositionRank, highlighting the positive effect of leveraging multiple features.

Additionally, we report the performance of our model without applying the weight adjustment mechanism. Results are higher or on-par with baselines that use topic information, and show that our model makes good use of the reinforcing relations between topics and the candidates they instantiate. We note that the drop-off in performance is more severe for F1@5 on the Semeval-2010 dataset, going from best to worst performance. Although further investigation is needed, we hypoth-

[2]These errors occur when a model correctly outputs a keyphrase because it contains an important word, but at the same time erroneously predicts other keyphrases because they contain the same word.

esise that our model struggles with selecting the most representative candidate from each topic using TextRank as a unique feature.

We also computed the topic coverage of the sets of keyphrases extracted by our model. With over 92% of the top-10 keyphrases assigned to different topics, our model successfully promotes diversity without the need of hard constraints. A manual inspection of the topic-redundant keyphrases reveals that a good portion of these are in fact clustering errors, that is, they have been wrongly assigned to the same topic (e.g. 'students' and 'student attitudes'). Some exhibit a hypernym-hyponym relation while both being in the gold references (e.g. 'model' and 'bayesian hierarch model' for document H-7 from the Semeval-2010 dataset), thus indicating inconsistencies in the gold data.

4 Conclusion

We introduced an unsupervised keyphrase extraction model that builds on a multipartite graph structure, and demonstrated its effectiveness on three public datasets. Our code and data are available at https://github.com/boudinfl/pke. In future work, we would like to apply ranking algorithms that leverage the specific structure of our graph representation, such as the one proposed in (Becker, 2013).

Acknowledgements

We thank the anonymous reviewers for their comments. This work was supported in part by the TALIAS project (grant of CNRS PEPS INS2I 2016). We also thank the members of the TALN team for their support.

References

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 546–555. http://www.aclweb.org/anthology/S17-2091.

N Becker. 2013. *Ranking on multipartite graphs*. Ph.D. thesis, Diploma thesis, Institute of Computer Science, LMU, Munich.

Florian Boudin. 2013. A comparison of centrality measures for graph-based keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, pages 834–838. http://www.aclweb.org/anthology/I13-1102.

Florian Boudin. 2015. Reducing over-generation errors for automatic keyphrase extraction using integer linear programming. In *Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction*. Association for Computational Linguistics, Beijing, China, pages 19–24. http://www.aclweb.org/anthology/W15-3605.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, pages 543–551. http://www.aclweb.org/anthology/I13-1062.

Adrien Bougouin, Florian Boudin, and Beatrice Daille. 2016. Keyphrase annotation with graph co-ranking. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 2945–2955. http://aclweb.org/anthology/C16-1277.

Corina Florescu and Cornelia Caragea. 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)*. Association for Computational Linguistics, Vancouver, Canada, pages 1105–1115. http://aclweb.org/anthology/P17-1102.

Sujatha Das Gollapalli, Cornelia Caragea, Xiaoli Li, and C. Lee Giles, editors. 2015. *Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction*. Association for Computational Linguistics, Beijing, China. http://www.aclweb.org/anthology/W15-36.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 1262–1273. http://www.aclweb.org/anthology/P14-1119.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '03, pages 216–223. https://doi.org/10.3115/1119355.1119383.

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*. Association for Computational Linguistics, Uppsala, Sweden, pages 21–26. http://www.aclweb.org/anthology/S10-1004.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, pages 366–376. http://www.aclweb.org/anthology/D10-1036.

Lus Marujo, Anatole Gershman, Jaime Carbonell, Robert Frederking, and Joao P. Neto. 2012. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*. Association for Computational Linguistics, Barcelona, Spain, pages 404–411. http://www.aclweb.org/anthology/W04-3252.

Lucas Sterckx, Thomas Demeester, Johannes Deleu, and Chris Develder. 2015. Topical word importance for fast keyphrase extraction. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '15 Companion, pages 121–122. https://doi.org/10.1145/2740908.2742730.

Nedelina Teneva and Weiwei Cheng. 2017. Salience rank: Efficient keyphrase extraction with topic modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 530–535. http://aclweb.org/anthology/P17-2084.

Antoine Tixier, Fragkiskos Malliaros, and Michalis Vazirgiannis. 2016. A graph degeneracy-based approach to keyword extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1860–1870. https://aclweb.org/anthology/D16-1191.

George Tsatsaronis, Iraklis Varlamis, and Kjetil Nørvåg. 2010. Semanticrank: Ranking keywords and sentences using semantic graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Coling 2010 Organizing Committee, Beijing, China, pages 1074–1082. http://www.aclweb.org/anthology/C10-1121.

Xiaojun Wan and Jianguo Xiao. 2008. Collabrank: Towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Coling 2008 Organizing Committee, Manchester, UK, pages 969–976. http://www.aclweb.org/anthology/C08-1122.

Fan Zhang, Lian'en Huang, and Bo Peng. 2013. Wordtopic-multirank: A new method for automatic keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, pages 10–18. http://www.aclweb.org/anthology/I13-1002.

Yuxiang Zhang, Yaocheng Chang, Xiaoqing Liu, Sujatha Das Gollapalli, Xiaoli Li, and Chunjing Xiao. 2017. Mike: Keyphrase extraction by integrating multidimensional information. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '17, pages 1349–1358. https://doi.org/10.1145/3132847.3132956.

Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achanauparp, Ee-Peng Lim, and Xiaoming Li. 2011. Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 379–388. http://www.aclweb.org/anthology/P11-1039.