Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Doctoral Consortium Track

KRAKEN: A Novel Semantic-Based Approach for Keyphrases Extraction

Simone D'Amico[1]
[1]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy
s.damico4@campus.unimib.it

Abstract

We propose *KRAKEN*, a novel approach for the extraction of keyphrases from texts. To this aim, *KRAKEN* makes use of distributional semantics to identify, as completely as possible, representative portions of documents, i.e. keyphrases. In addition, we define novel metrics to assess a weighted significance to the keyphrases extracted from a document, identifying the most important ones by assessing their semantic similarity with the text of the document they belong to.

1 Introduction

Keyphrases are a set of relevant terms that provide a high-level description of a textual document. The Keyphrases Extraction task (KPE) defines a series of approaches and techniques aimed to identify keyphrases from a text. These tasks range from *keyphrases Assignment*, in which the most relevant phrases are identified starting from a dictionary of words or expressions, to *keyphrases Extraction* in which the various phrases are identified and extracted directly from the examined corpus. The work described in this document focuses on defining a new approach for extracting keyphrases directly from the text. The typical keyphrases extraction pipeline consists of two steps: (i) in the keyphrases extraction step, a set of candidate keyphrases are identified and extracted; (ii) in keyphrases ranking step a rank is applied to identify the best keyphrases among those extracted, according to criteria of semantic similarity both among keyphrase's terms and with the rest of the keyphrases.

2 Motivation and Contribution

The state-of-the-art approaches to identify keyphrases consider only the syntactic aspect of sentences, ignoring the semantics of words in the text. The aims of this study is to propose a novel approach, namely `KRAKEN` (**K**eyphrease ext**RA**ction ma**K**ing use of **E**mbeddi**N**gs), that performs keyphrases extraction using distributional semantics techniques. Such techniques allow obtaining vector representations maintaining syntactic and semantic word relationships. The expected contributions are the following: (i) the study, design, and benchmark evaluation of a new method that exploits word embedding for identifying keyphrases, accounting for the correlation among identified words and the rest of the text and (ii) the definition of a metric to rank the keyphrases considering both the correlation of words within the phrase and with the text from which it was extracted.

3 Related Works

There are three types of Keyphrase extraction techniques: *Deep Learning Techniques*, *Supervised Techniques* and *Unsupervised Techniques* keyphrase extraction. In particular, in the latter category, we can distinguish between *text construction-based* and *relationship-based*. An example of the former is YAKE, proposed by [Campos *et al.*, 2018]. In YAKE, for each candidate keyphrase, different features are considered, such as the position or frequency of the terms. These features are then combined into a heuristic score to identify the best keyphrases. In the case of *relationship-based* techniques, we distinguish between *graph-based* and *topic-based* approaches, in these algorithms, it is assumed that the words that co-occur within a certain window in the text have some relationship. In topic-based approaches, it is assumed that words in the same topic are related. Words are assigned to topics using techniques such as LDA or clustering. In graph-based approaches, graphs are constructed with words as nodes, connected to each other if the associated words co-occur in a window of fixed size. Different algorithms, such as TextRank or PageRank, are used to identify the sequences of words with the highest score to construct the keyphrases. In this stream, [Chi and Hu, 2021] propose a PageRank-like method called ISKE to weigh relationships between sentences based on the assumption that there must be strong causality between adjacent sentences. [Liu *et al.*, 2010] proposed TopicPageRank (TPR), in which a graph is constructed using the co-occurrence statistics of the words, then PageRank is applied to weigh words according to their importance in the topics in which they appear. The approach by [Boudin, 2018] uses a multipartite graph structure (MUL) to encode keyphrases as graph nodes and their co-occurrence statistics as edges. Then they use TextRank to rank the nodes.

Cited approaches are all relevant, nevertheless, none of them used embeddings techniques, as a result, semantic information is lost (e.g. synonymous or different words in similar context).

4 KRAKEN: A Novel Approach to KPE

KRAKEN is composed by two steps: the identification and extraction of keyphrases and the selection of the best keyphrases. In the first step, after a text preprocessing phase, KRAKEN take as candidate phrases nouns and adjectives. The approach operates iteratively by building a window around these words, incorporating the previous and next words into this window, therefore, the concept of keyphrases corresponds to the window obtained. The construction of the window takes place thanks to the use of a model of word embeddings: iterates on the part of text preceding the window, the Pearson correlation index is calculated between the embeddings vector of the word p preceding the window and the vector of the window itself. If this correlation is greater than the correlation calculated at the previous iteration, namely without p, then p is added to the window itself and the process is iterated. The stop criterion of the algorithm is reached if the new correlation is lower than that calculated on the previous iteration. The word embeddings model used has a fastText architecture with CBOW algorithm, a learning rate of 0.1 and vectors of 300 components were generated.

The result of the previous step is a certain number of candidate keyphrases for each text. In this evaluation step, each keyphrases is evaluated with the aim of assigning a score to be used for selecting the best keyphases to use for the evaluation with the baseline. KRAKEN also defines two measures to rank the various keyphrases: (i) the intra-window correlation measures the internal cohesion of the phrases in terms of the coherence of their words. The average of the correlations between all the pairs of words in the window is calculated, if this average is higher than a certain threshold then the keyphrases is selected, otherwise it is discarded; (ii) the inter-windows correlation expresses the capacity of a keyphrase to have a meaning similar to all the others and therefore, the higher it is, the more this phrases is able to represent the text. For each of these candidate keyphrases, the mean correlation with all the windows extracted from the same text is calculated.

If the keyphrase is composed only by a single word, it is used only in the inter-windows correlation, otherwise for multiple words keyphrase, we first use the intra-window correlation for a first filtering of the windows and then the inter-window correlation. For both metrics the Pearson correlation index was used, obtained between the embedding vectors of the keyphrases.

Figure 1 shows an example of KP extraction with KRAKEN, the identified KPs are highlighted in red and in the next box there are all the baseline KPs for that text.

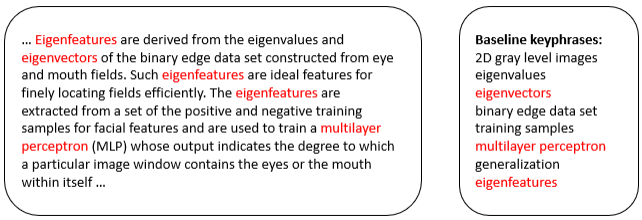

Figure 1: Example of KP extracted from a text.

Dataset	Eval	MUL	TPR	ISKE	Yake	KRAKEN
Inspec	P@k	3.9	2.7	4	1.4	**11.2**
	R@k	5.1	3.7	5.3	2	**12.9**
	F_1@k	4.1	2.9	4.2	1.5	**11**
KDD	P@k	10.1	8.1	**12**	3.1	10.7
	R@k	12.2	9.7	14.3	4	**17.4**
	F_1@k	10.7	8.5	12.3	3.4	**13.3**

Table 1: Performance@5 on datatsets.

5 First Baseline Evaluation

KRAKEN's results will be compared with the results obtained with state of the art methods and applied to benchmark datasets evaluated on *precision@k*, *recall@k*, F_1-*measure@k*.

Table 1 shows the first results obtained selecting the first best 5 keyphrases with the highest score using the *KDD* [Das Gollapalli and Caragea, 2014] datasets, a collection of 757 documents and *Inspec* [Hulth, 2003] dataset that consists of 2000 documents. The first results show an improvement over the other approaches.

6 Conclusion

Although the research activity on KRAKEN is still in progress, the results obtained so far are promising. There is an improvement in performance compared to other methods and a new metric for the scoring of keyphrases has been defined that considers semantic information.

Acknowledgements

The work described in this paper is supported by the University of Milano-Bicocca, in particular by Dr. Malandri, Prof. Mercorio and Prof. Mezzanzanica

References

[Boudin, 2018] Florian Boudin. Unsupervised keyphrase extraction with multipartite graphs. In *Conference of NAACL-HLT, Vol. 2*, pages 667–672, 2018.

[Campos *et al.*, 2018] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. Yake! collection-independent automatic keyword extractor. In *AIR*, pages 806–810, 2018.

[Chi and Hu, 2021] Ling Chi and Liang Hu. Iske: An unsupervised automatic keyphrase extraction approach using the iterated sentences based on graph method. *KBS*, 223:107014, 2021.

[Das Gollapalli and Caragea, 2014] Sujatha Das Gollapalli and Cornelia Caragea. Extracting keyphrases from research papers using citation networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), 2014.

[Hulth, 2003] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *EMNLP*, pages 216–223, 2003.

[Liu *et al.*, 2010] Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. Automatic keyphrase extraction via topic decomposition. In *EMNLP*, pages 366–376, 2010.