information

MDPI

Article

BibRank: Automatic Keyphrase Extraction Platform Using Metadata

Abdelrhman Eldallal and Eduard Barbu *

Institute of Computer Science, University of Tartu, 50090 Tartu, Estonia; abdelrhman.d@aucegypt.edu
* Correspondence: eduard.barbul@ut.ee

Abstract: Automatic Keyphrase Extraction involves identifying essential phrases in a document. These keyphrases are crucial in various tasks, such as document classification, clustering, recommendation, indexing, searching, summarization, and text simplification. This paper introduces a platform that integrates keyphrase datasets and facilitates the evaluation of keyphrase extraction algorithms. The platform includes BibRank, an automatic keyphrase extraction algorithm that leverages a rich dataset obtained by parsing bibliographic data in BibTeX format. BibRank combines innovative weighting techniques with positional, statistical, and word co-occurrence information to extract keyphrases from documents. The platform proves valuable for researchers and developers seeking to enhance their keyphrase extraction algorithms and advance the field of natural language processing.

Keywords: keyphrase extraction; graph algorithms; software platform; BibTeX datasets; context

check for updates

Citation: Eldallal, A.; Barbu, E. BibRank: Automatic Keyphrase Extraction Platform Using Metadata. *Information* **2023**, *14*, 549. https://doi.org/10.3390/info14100549

Academic Editor: Andrea Giovanni Nuzzolese

Received: 23 August 2023
Revised: 29 September 2023
Accepted: 29 September 2023
Published: 7 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The internet hosts an extensive collection of scientific documents, numbering in the tens of millions. Google Scholar, a web-based search engine dedicated to academic research, strives to provide comprehensive access to scholarly literature across various disciplines. One study [1] reported that by the end of 2018, Google Scholar had indexed approximately 400 million articles. Keyphrases considered concise summaries of documents, aid information retrieval, indexing, and collection browsing. Automatic keyphrase extraction is the process of automatically identifying essential phrases within a document. Keyphrases find application in document clustering, classification, summarization, recommendation systems, and question answering. Automatic keyphrase extraction methods have been developed in domains such as social media, medicine, law, and agriculture, where they support specialized systems for organizing and retrieving information [2,3].

Automatic keyphrase extraction methods can be categorized into unsupervised, supervised, and semi-supervised. Unsupervised techniques, which are domain-dependent, do not require labeled training data. On the other hand, supervised methods rely on manually annotated data, while semi-supervised ones strike a balance by requiring less annotated data compared to supervised methods.

This paper introduces a downloadable platform that integrates keyphrase datasets in BibTeX format and facilitates the evaluation of keyphrase extraction algorithms. The platform currently encompasses 19 algorithms for automatic keyphrase extraction and methods for evaluating their performance against a diverse gold standard dataset. Among the 19 algorithms is a keyphrase extraction method called BibRank. BibRank exploits an information-rich dataset created by parsing bibliographic data in BibTeX format. It combines a new weighting technique applied to the bibliographic data with positional, statistical, and word co-occurrence information.

The main contributions of this paper are as follows:

1. BibRank dataset: Construction of an information-rich dataset by parsing publicly available bibliographic data, which includes manually assigned keywords.

2. BibRank algorithm: Introduction of the BibRank algorithm, a novel method for keyphrase extraction that utilizes the bibliographic information within the BibRank dataset and statistical information.
3. BibRank platform: Provision of a downloadable platform that integrates the BibRank dataset, BibRank algorithm, and other state-of-the-art keyphrase extraction algorithms. The platform includes evaluation metrics and allows for the integration of keyphrase extraction algorithms and datasets.
4. Manual evaluation of keyphrases: Keyphrase extraction algorithms are evaluated using gold standard datasets as a benchmark. In our evaluation process, we rely on expert human evaluators to assess the quality and effectiveness of these gold-standard algorithms.

The remaining sections of the paper closely align with the contributions presented earlier. The next section briefly overviews notable keyphrase extraction algorithms and datasets. Section 3 introduces the heterogeneous BibRank dataset and presents the BibRank algorithm. Section 4 concentrates on the automatic evaluation of the BibRank algorithm and other state-of-the-art algorithms. Moreover, this section includes assessing the gold standard algorithms' quality, guided by expert human evaluators. The paper concludes by summarizing our findings.

2. Related Work

This section provides an overview of the essential stages in the automatic keyword extraction algorithms pipeline, highlighting the algorithms that influenced BibRank.

The keyword extraction pipeline comprises linguistic preprocessing, candidate phrase selection, keyphrase feature selection, and keyphrase ranking and selection. The text is segmented into sentences and tokenized into words during linguistic preprocessing. Several language processing techniques are applied, including lemmatization, stemming, POS tagging, stop word removal, and Named Entity Recognition (NER) [2]. Sometimes, POS tagging is followed by syntactic parsing, and NER is particularly valuable in languages with reliable NER systems. Candidate phrases are selected from the processed text using n-gram sequencing and noun-phrase chunking (NP chunking) [4]. Rules based on acceptable sequences of POS tags, such as selecting sequences starting with adjectives and ending with a noun in English, are employed [5] to reduce the number of candidate phrases.

The subsequent step in the pipeline is feature selection for candidate phrases. Two types of features are calculated: in-document features and external features [2]. In-document features can be statistical [6], positional [3], linguistic [7], or context-based [8]. Statistical features such as the TF-IDF score are commonly used, while positional features indicate the candidate phrase's location in the title, abstract, or main text. Context features, such as sentence embeddings computed by deep neural networks, are also utilized. External features require resources such as Wikipedia [9] to quantify the association strengths between keyphrases. An example of a supervised keyphrase extraction algorithm that utilizes external features is CeKE [8]. CeKE employs citation-based features created from the references used in a publication.

The assignment of weights to each candidate phrase is based on the calculated features in the keyphrase ranking and selection step. Subsequently, the candidate phrases are sorted, and the most relevant ones are selected using an experimental threshold.

In the context of unsupervised methods, graph-based ranking algorithms such as TextRank [4] deserve to be mentioned. These algorithms draw inspiration from the Google PageRank algorithm [10] and have demonstrated success in text summarization and keyword extraction. The text document is represented as a graph, where candidate phrases are nodes, and their relationships are edges. These relationships can be co-occurrence relations [11], syntactic dependencies [4], or semantic relations [9].

In the keyphrase ranking step, an adapted PageRank algorithm is employed, which iterates until convergence on the graph representation of the text, ultimately selecting the top-ranked candidate phrases. Another algorithm in this family is PositionRank [12].

Building upon the principles of TextRank, PositionRank introduces a bias towards frequently occurring candidate phrases that appear early in the document. It operates at the word level, transforming the text into a graph, applying a position-based PageRank algorithm, and extracting candidate phrases.

Other initiatives that share a connection with our work encompass the creation and visualization of bibliometric networks. VosViewer stands out as a notable tool in these endeavors [13]. While VosViewer is not specifically a tool for keyphrase extraction, it is a relevant software used for creating and visualizing bibliometric networks. These networks can encompass journals, researchers, or single publications, helping to analyze and visualize trends and patterns in scientific literature. VosViewer provides multiple avenues to build, visualize, and investigate bibliometric networks, simplifying the process for users to gain insights from bibliometric data.

3. BibRank

3.1. BibRank Dataset

Keyphrase datasets serve as the standard for evaluating automatic keyphrase extraction methods, encompassing texts and lists of associated keyphrases. These gold standards are widely available across scientific publications, news articles, and web posts [7].

We utilize BibTeX entries from the web to construct a new and information-rich keyphrase extraction dataset. Unlike existing datasets that often include only the abstract, full article text, title, and keywords of a document, our dataset incorporates additional metadata, such as the publication year, journal title, and author name. An example of a BibTeX record for a publication is illustrated in Figure 1, where the entry type (e.g., "Article") is indicated after the "@", followed by various attributes (e.g., author, title, journal, and paper keywords) and their respective values.

Publicly available BibTeX records can be found in online archives such as the TUG bibliography archive. TUG's archive contains a vast collection of over 1.6 million categorized BibTeX records from various journals. The archive supports search capabilities using SQL commands [14].

To create the BibRank dataset, we processed more than 30,000 BibTeX records extracted from the TUG bibliography archive. Currently, the dataset consists of 18,193 unique records with 22 attributes. These attributes represent the distinct values in all the bib records, including publication year, journal of publication, and bib archive. The dataset includes publications from 1974 to 2019. Table 1 provides statistics on authors, journals, topics, and bib files covered by the dataset.

The bib files, referring to the archives or databases from which the papers were imported, were categorized into one of the following 12 topics: science history journals, computer science journals and topics, ACM Transactions, cryptography, fonts and typography, IEEE journals, computational/quantum chemistry/physics, numerical analysis, probability and statistics, SIAM journals, mathematics, and mathematical and computational biology. Expanding the dataset by processing additional bibliography files in BibTeX format is possible.

The file for the dataset, together with the essential tools for altering and producing new datasets, is available in the BibRank project's GitHub repository. This repository grants users access to the original data and equips them with the requisite resources for customizing the data to their particular requirements or generating entirely new datasets.

```
@Article{Wang:2009:EKF,
  author =       "Zidong Wang and Xiaohui Liu and Yurong Liu and Jinling
                 Liang and Veronica Vinciotti",
  title =        "An Extended {Kalman} Filtering Approach to Modeling
                 Nonlinear Dynamic Gene Regulatory Networks via Short
                 Gene Expression Time Series",
  journal =      j-TCBB,
  volume =       "6",
  number =       "3",
  pages =        "410-419",
  month =        jul,
  year =         "2009",
  CODEN =        "ITCBCY",
  DOI =          "https://doi.org/10.1109/TCBB.2009.5",
  ISSN =         "1545-5963 (print), 1557-9964 (electronic)",
  ISSN-L =       "1545-5963",
  bibdate =      "Tue Aug 11 18:13:22 MDT 2009",
  bibsource =    "http://portal.acm.org/;
                 http://www.math.utah.edu/pub/tex/bib/tcbb.bib",
  abstract =     "In this paper, the extended Kalman filter (EKF)
                 algorithm is applied to model the gene regulatory
                 network from gene time series data. The gene regulatory
                 network is considered as a nonlinear dynamic stochastic
                 model that consists of the gene measurement equation
                 and the gene regulation equation. After specifying the
                 model structure, we apply the EKF algorithm for
                 identifying both the model parameters and the actual
                 value of gene expression levels. It is shown that the
                 EKF algorithm is an online estimation algorithm that
                 can identify a large number of parameters (including
                 parameters of nonlinear functions) through iterative
                 procedure by using a small number of observations. Four
                 real-world gene expression data sets are employed to
                 demonstrate the effectiveness of the EKF algorithm, and
                 the obtained models are evaluated from the viewpoint of
                 bioinformatics.",
  acknowledgement = ack-nhfb,
  fjournal =     "IEEE/ACM Transactions on Computational Biology and
                 Bioinformatics",
  journal-URL =  "http://portal.acm.org/browse_dl.cfm?idx=J954",
  keywords =     "clustering; DNA microarray technology; extended Kalman
                 filtering; gene expression; Modeling; time series
                 data.",
}
```

Figure 1. BibTeX record example.

Table 1. BibRank dataset.

Data	Count
Records (abstracts)	18,193
Authors	16,883
Journals	693
Bib Files	285
Topics	12
Avg Words	121
Avg Keyphrases	9

3.2. BibRank Algorithm

The BibRank algorithm, comprising five steps, presents an innovative method for weighting candidate phrases, emphasizing the abstracts of scientific publications and based on the concept of a context for a group of BibTeX records.

1. Candidate Selection. The candidate phrases in the document are noun chunks. To identify the noun chunks, we apply rules based on sequences of POS tags. In our workflow, we use the Stanford CoreNLP Natural Language Processing Toolkit [15], but other noun chunkers can be easily integrated into the platform.
2. PositionRank Weight Calculation. The PositionRank algorithm [12] assigns position weights to candidate phrases. Higher weights are given to the words appearing earlier in the document. For example, if a phrase consists of positions 3, 6, and 8, its weight is calculated as follows: $\frac{1}{3} + \frac{1}{5} + \frac{1}{8} = \frac{5}{8} = 0.625$. The final weight of each candidate phrase is determined by summing and normalizing the position weights of each word in the phrase. Additionally, the scores of each word are recursively computed using the PageRank algorithm, as described by Equation (1) [4,12]:

$$S(v_i) = (1-d) \cdot \hat{p}_i + d \cdot \sum_{v_j \in \mathrm{In}(v_i)} \frac{w_{ji}}{Out(v_j)} S(v_j) \quad (1)$$

   where $S(v_i)$ represents the weight of each word i in a candidate phrase p, represented by the vertex v_i, the damping factor d reflects the probability of jumping to a random vertex in the graph, $\hat{p}$ is the position weight of the word i, the set $\mathrm{In}(v_i)$ contains the adjacent vertices pointing to vertex i, w_{ji} is the edge weight between v_i and v_j, and finally, $Out(v_j)$ is the set of adjacent vertices pointed to by vertex i, and is computed as $\sum_{V_k \in Out(V_j)} w_{jk}$.
3. Context Formulation. The computation of the context for a publication involves selecting a set of BibTeX records according to specific criteria. For instance, if we consider a computer science article published in 2012, the context could be formed by including all computer science papers published within the same year. With the original BibRank dataset containing 22 attributes, each attribute can potentially define a distinct context.
4. Bib Weight Calculation. The bib weights aim to capture the occurrence frequency of candidate phrases within the context. Each record includes a list of keyphrases, allowing for the calculation of weights for candidate phrases based on Equation (2):

$$\lambda_p = \frac{1}{\alpha} \sum_{d \subseteq D} c_{pd} \quad (2)$$

   where λp is the bib weight, α is a factor used for normalization, D is the set of all records that belong to the chosen context, d is a record, and c is the occurrence of a candidate phrase in the record's keyphrases list. α was calculated as the maximum bib weight across all keyphrases in the context documents.
5. Candidate Phrase Ranking and Selection. The ranking of candidate phrases is determined by combining their bib weights and position scores. The scores of individual words within each candidate phrase are added to the phrase's bib weight, resulting in a sum that determines the final ranking of the candidate phrases, as illustrated in Equation (3). The document's keyphrases are then determined by selecting the top N candidate phrases:

$$S_{final}(p) = \sum_{v_i \in \mathrm{V_p}} S(v_i) + \lambda_p \quad (3)$$

   where V_p is the set of words that belongs to candidate phrase p and λ_p is the calculated bib weight for the candidate phrase p.

As illustrated in Figure 2, The BibRank algorithm begins by processing the input text, extracting nouns and noun phrases such as 'Keyword' and 'automatic identification', which are considered as selected candidates. It then infers keyphrases, including 'Keyword extraction' and 'automatic identification', assigning them scores of 0.38 and 0.30, respectively. These scores denote their relevance and significance to the document's main topic, calculated based on position weight and Bib weights.

- **Input Text**: Keyword extraction is tasked with the automatic identification of terms that best describe the subject of a document.
- **Nouns and Noun phrases**: 'extraction', 'subject', 'Keyword', 'identification', 'automatic identification', 'document', 'terms', 'Keyword extraction', 'best', 'automatic'
- **Keyphrases**: 'Keyword extraction', 'automatic identification', 'extraction', 'automatic', 'Keyword', 'identification'
- **scores**: 0.38, 0.30, 0.23, 0.17, 0.15, 0.13

Figure 2. BibRank keyphrase extraction example.

3.3. BibRank Platform

BibRank is a versatile online platform developed in Python that simplifies the integration of keyphrase extraction algorithms, encompassing three modules: Datasets, Algorithms, and Evaluation.

One of the standout attributes of the platform is its comprehensive support for keyphrase extraction datasets. It seamlessly incorporates user datasets and features multiple pre-integrated datasets, such as the BibRank dataset (see Section 3.1) and five others extensively detailed in Table 2. This table provides crucial information about the papers linked to each dataset, the number of documents contained, and the document types, distinguishing between abstracts and full papers.

Moreover, BibRank facilitates users in crafting personalized datasets with ease. The platform offers user-friendly routines tailored to process BibTeX files, simplifying the generation of new datasets that align with the user's specific needs and requirements.

Table 2. BibRank platform datasets.

Dataset	Documents	Type
ACM [16]	2304	Full papers
NUS [17]	211	Full papers
Inspec [18]	2000	Abstracts
WWW [8]	1330	Abstracts
KDD [8]	755	Abstracts
BibRank Dataset	18,193	Abstracts and Metadata

The platform offers a comprehensive range of keyphrase extraction algorithms, including the BibRank algorithm (refer to Section 3.2) and ten additional ones, all clearly specified in Table 3. It provides a user-friendly interface for effortlessly integrating the user's own keyphrase extraction algorithms. For smooth integration, the user's algorithm must extend a superclass that encompasses the blueprint for the crucial extraction operations, where the algorithm's name is designated as a class attribute. Additionally, the algorithm must incorporate a function that efficiently returns the extracted keyphrases and their corresponding weights. The platform incorporates PKE, an open-source toolkit for keyphrase [19].

To assess the accuracy of a keyphrase extraction algorithm on a given dataset, the platform provides an evaluation module in the form of a Python script. Users can select the algorithm to be evaluated and specify the metadata for the dataset, such as the year of publication or journal. The evaluation script computes the recall (R), precision (P), and F1 scores, widely recognized as standard measures of algorithm performance.

Table 3. BibRank platform models.

Method	Year	Approach Type
TFIDF [20]	1999	Statistical
KPMiner [21]	2010	Statistical
YAKE [22]	2020	Statistical
TextRank [4]	2004	Graph based
CollabRank [23]	2008	Graph based
TopicRank [24]	2013	Graph based
PositionRank [12]	2017	Graph based
SGRank [6]	2015	Hybrid Statistical-graphical
sCAKE [25]	2018	Hybrid Statistical-graphical
KeyBERT [26]	2021	Sentence Embeddings

4. Results

4.1. Evaluation Methodology

The widely accepted assumption that the gold standard serves as the reference truth for evaluating algorithms is acknowledged. However, a comprehensive two-fold evaluation process was conducted to examine this assumption critically. The first evaluation aimed to assess the algorithms against the gold standard, while the second evaluation focused on evaluating the gold standard itself.

Datasets with manually assigned keywords were used as benchmarks to assess the algorithms' performance. The evaluations were carried out using the BibRank platform, where the algorithms were tested on the BibRank dataset with parameter adjustments. The default setting for the first parameter, determining the number of keywords to extract, was 10 for all algorithms. The second parameter, the tokenizer, utilized the Stanford CoreNLP toolkit, as explained in the BibRank algorithm section. The damping factor α was set to 0.85, and the window size was set to 2 based on experiments by [12]. Extracted keyphrases were compared to the manually assigned keywords in the gold standard dataset to measure the algorithms' performance, considering exact matches as successful hits. Standard evaluation metrics such as recall, precision, and F1 score were computed.

Evaluators with expertise were sought through a reputable freelancing platform to evaluate the gold standard. These evaluators were carefully selected based on specific criteria, including fluency in English and a proven track record in similar tasks. Two experts were assigned to evaluate 100 annotated documents containing keywords using seven algorithms and the gold standard. The evaluators were kept unaware of the algorithm names or the gold standard during the evaluation process to prevent potential bias. The evaluators meticulously annotated the different datasets using a five-point scale:

1. Very bad: The keywords are considered inadequate and do not meaningfully represent the text.
2. Bad: The keywords are a mix of poor and good choices, lacking consistency and not fully capturing the essence of the text.
3. Acceptable: The keywords are generally satisfactory and represent the text to a reasonable extent.
4. Good: The keywords are of good quality, although they may not fully encompass all the text's main ideas.
5. Very good: The provided keywords accurately summarize the text and effectively capture the main ideas.

Overall, our two-fold evaluation approach provides a comprehensive analysis of both the algorithm and the gold standard, allowing us to understand the strengths and weaknesses of each.

4.2. Results

The evaluation of the algorithms involved three experiments, each utilizing a different section of the BibRank dataset. The experiments focused on specific domains, namely "Computer science (compsci), "ACM", and "history, philosophy, and science", consisting of 335, 127, and 410 papers, respectively. In choosing the dataset years, we aimed for diverse temporal coverage and ran tests on various combinations to ensure validity. For Computer science (compsci), bib scores were generated using publications from the years 1980 to 1987, and the test data were sourced from publications in 1988; ACM bib scores were derived from 1990 to 1996 and tested against 1997 to 2020 publications; for "history, philosophy, and science", scores were based on 2009 to 2011, testing with publications from 2012 to 2014. For a comprehensive overview of these experiments, including the categories used, please refer to Table 4. The table displays the categories the articles belong to and seven selected algorithms for evaluation. We selected these algorithms to exemplify various keyphrase extraction approaches discussed in the Related Works section, showcasing the implementation of distinct methodologies for keyword extraction.

Upon closer inspection, the BibRank algorithm demonstrates consistent enhancements across different datasets, as can be seen in the Tables 5–7. When compared to TextRank and PositionRank, which use comparable techniques, the integration of Bib Weights in the BibRank algorithm leads to a noticeable enhancement in performance.

1. YAKE (Yet Another Keyword Extractor) is a statistical keyphrase extraction algorithm that utilizes a "maximal marginal relevance" approach to promote diversity in the selected keywords. This ensures that the extracted keyphrases cover a wide range of topics and concepts.
2. The SGRank and sCake methods are algorithms used to extract keyphrases from a document. They employ statistical analysis and graph-based techniques, blending both advantages to identify important keywords. Notably, sCake stands out for integrating domain-specific knowledge into its process when analyzing documents.
3. KeyBERT represents a user-friendly and lightweight algorithm for keyword extraction. It harnesses the power of BERT transformers' embeddings to identify important keywords in a given text. Using an unsupervised technique, KeyBERT calculates the cosine similarity between each phrase and document to determine the most relevant keyphrases.
4. The preceding sections contain in-depth discussions about graph-based techniques, including TextRank, PositionRank, and BibRank. These algorithms use graph-based approaches to analyze word relationships and extract essential keywords from a text.

Our objective in incorporating these algorithms is to comprehensively evaluate various keyphrase extraction techniques.

In addition to using standard gold keyphrases, the chosen experts manually evaluated seven keyphrase extraction approaches. To gauge the performance of each method, the experts assigned scores from 1 to 5 to the generated keywords for 100 randomly selected documents. Table 8 summarizes the average performance of each evaluated approach. These evaluations offer valuable insights into the effectiveness of the diverse keyphrase extraction methods.

Table 4. Evaluation results of selected keyphrase extraction algorithms, including BibRank.

	Compsci			**History, Philosophy, and Science**			**Probstat**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
YAKE	0.0728	0.0367	0.0458	0.0606	0.0705	0.0602	0.0171	0.0366	0.0228
SGRank	0.1282	0.0730	0.0861	0.0645	0.0903	0.0690	0.0594	0.1235	0.0783
sCake	0.1213	0.0714	0.0829	0.0676	0.0949	0.0724	0.0549	0.1141	0.0725
KeyBert	0.0839	0.0564	0.0617	0.0315	0.0501	0.0368	0.0380	0.0880	0.0520

Table 4. *Cont.*

	Compsci			History, Philosophy, and Science			Probstat		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
TextRank	0.1236	0.0716	0.0835	0.0621	0.0912	0.0685	0.0562	0.1175	0.0745
PositionRank	0.1579	0.0953	0.1094	0.0740	0.1102	0.0817	0.0605	0.1347	0.0815
BibRank	0.1812	0.109	0.1249	0.0811	0.1136	0.0867	0.0659	0.1457	0.0886

Table 5. BibRank improvements: Compsci.

	Bib Wegihts Records	**P**	**R**	**F1**
TextRank	0	0.1236	0.0716	0.0835
PositionRank	0	0.1579	0.0953	0.1094
BibRank	299	0.1764	0.1065	0.1216
	976	0.1764	0.1063	0.1218
	1155	0.1809	0.1083	0.1242
	1746	0.1812	0.109	0.1249

Table 6. BibRank improvements: History, philosophy, and science.

		History, Philosophy, and Science		
	Bib Weights	**P**	**R**	**F1**
TextRank	0	0.0621	0.0912	0.0685
PositionRank	0	0.0740	0.1102	0.0817
BibRank	54	0.0780	0.1098	0.0833
	81	0.0787	0.1108	0.0842
	173	0.0811	0.1136	0.0867

Table 7. BibRank Improvements: probstat.

	Bib Wegihts Records	**P**	**R**	**F1**
TextRank	0	0.0562	0.1175	0.0745
PositionRank	0	0.0605	0.1347	0.0815
BibRank	139	0.0646	0.1422	0.0868
	237	0.0649	0.1423	0.0870
	367	0.0659	0.1457	0.0886

Table 8. Manual evaluation.

Model	**Expert 1**	**Expert 2**
Gold Standard	2.85	2.08
YAKE	2.95	1.65
SGRank	4.2	3.47
sCake	3.71	3.39
KeyBert	4.61	4.39

Table 8. *Cont.*

Model	Expert 1	Expert 2
TextRank	4.77	3.99
PositionRank	4.41	3.37
BibRank	4.4	3.77

4.3. Discussion

The YAKE algorithm and the gold standard sets of keyphrases received the lowest scores from the experts in our evaluation. This result was expected for YAKE, as it is the only statistical approach among the evaluated techniques. Prior research [5] has also indicated that models relying on statistical features exhibit lower average performance in keyphrase extraction tasks. However, the surprising finding was the performance of the gold standard keyphrases.

We conducted interviews with the experts who participated in the evaluation to gain deeper insights. One expert mentioned that the gold standard keyphrases are overly general and limited in scope. They are designed to capture the central ideas or keyphrases of the document, which may result in the omission of some important keywords. In contrast, algorithms such as BibRank, PositionRank, TextRank, and KeyBERT better understood the document's meaning, enabling them to extract more relevant and specific keyphrases.

Figure 3 presents an abstract that the experts evaluated, and the corresponding scores provided by the experts are listed in Table 9. The gold standard keywords received low scores despite including important keyphrases like "Chinese dependency parsing" and "unlabeled data". However, there were cases where essential keyphrases were missing, while some keywords not explicitly mentioned in the abstract were included in the gold standard set. For instance, the term "semi-supervised learning" was incorporated in the gold standard keyword list but did not appear in the original abstract.

YAKE achieved a low score, indicating that the algorithm lacks the contextual understanding exhibited by the other keyword extraction methods.

SGRank outperformed the gold standard, effectively highlighting essential keywords, such as "long-distance word", "unlabeled attachment score", and "supervised learning method".

SCake also demonstrated strong performance, successfully extracting detailed keywords related to different types of dependency parsers and incorporating "short dependency information".

KeyBERT showcased robust performance, extracting comprehensive keywords such as "improves parsing performance" and "parsing approach incorporating", which enhanced the understanding of the paper's content.

TextRank consistently performed well, generating similar keywords to SCake and SGRank, indicating its consistency in identifying key concepts.

PositionRank, with a score of 5, provided additional context by introducing terms such as "short dependencies".

BibRank consistently scored 5 in both evaluations, effectively extracting keywords related to various parser types, "short dependency information", and specific performance metrics such as "high performance". It also included additional contextual keywords, such as "machine translation", providing a comprehensive overview of the abstract's content.

Overall, these evaluations shed light on the strengths and weaknesses of different keyphrase extraction methods and help us understand their performance characteristics in the context of academic literature.

The detailed results of our evaluations, substantiating the findings discussed in this paper, are recorded and made available for public scrutiny and exploration. These results can be found in our GitHub repository's evaluation_results folder.

Table 9. The expert evaluation for the abstract presented in Figure 3.

Model	Expert 1	Expert 2
Gold Standard	2	1
YAKE	3	1
SGRank	4	3
sCake	4	3
KeyBert	4	4
TextRank	5	4
PositionRank	5	5
BibRank	5	5

- **Abstract:** *Dependency parsing has become increasingly popular for a surge of interest lately for applications such as machine translation and question answering. Currently, several supervised learning methods can be used for training high-performance dependency parsers if sufficient labeled data are available. However, currently used statistical dependency parsers provide poor results for words separated by long distances. In order to solve this problem, this article presents an effective dependency parsing approach of incorporating short dependency information from unlabeled data. The unlabeled data are automatically parsed by using a deterministic dependency parser, which exhibits a relatively high performance for short dependencies between words. We then train another parser that uses the information on short dependency relations extracted from the output of the first parser. The proposed approach achieves an unlabeled attachment score of 86.52%, an absolute 1.24% improvement over the baseline system on the Chinese treebank dataset. The results indicate that the proposed approach improves the parsing performance for longer distance words.*
- **gold**: 'chinese dependency parsing', 'semi-supervised learning', 'unlabeled data'
- **YAKE**: 'dependency', 'parser', 'datum', 'performance', 'word', 'unlabeled', 'short', 'approach', 'distance', 'high'
- **Sgrank**: 'long distance word', 'chinese treebank datum', 'unlabeled attachment score', 'deterministic dependency parser', 'short dependency relation', 'statistical dependency parser', 'short dependency information', 'performance dependency parser', 'supervised learning method', 'unlabeled datum'
- **sCake**: 'performance dependency parser', 'statistical dependency parser', 'deterministic dependency parser', 'short dependency information', 'dependency parsing', 'short dependency relation', 'effective dependency', 'unlabeled datum', 'chinese treebank datum', 'high performance'
- **KeyBert**: 'improves parsing performance', 'effective dependency parsing', 'performance dependency parsers', 'dependency parsing approach', 'dependency parsers provide', 'statistical dependency parsers', 'parsers provide poor', 'parsing performance', 'deterministic dependency parser', 'parsing approach incorporating'
- **TextRank**: 'performance dependency parser', 'deterministic dependency parser', 'statistical dependency parser', 'short dependency information', 'short dependency relation', 'dependency parsing', 'effective dependency', 'unlabeled datum', 'long distance word', 'chinese treebank datum'
- **PositionRank**: 'performance dependency parsers', 'statistical dependency parsers', 'deterministic dependency parser', 'short dependency information', 'short dependency relations', 'short dependencies', 'effective dependency', 'dependency', 'first parser', 'machine translation'
- **BibRank**: 'performance dependency parsers', 'statistical dependency parsers', 'deterministic dependency parser', 'short dependency information', 'short dependency relations', 'short dependencies', 'effective dependency', 'dependency', 'high performance', 'chinese treebank data'

Figure 3. Generative model-based keyphrase extraction example.

5. Conclusions

This paper introduces the BibRank platform, a versatile online platform developed in Python, which simplifies the integration of keyphrase extraction algorithms. A new keyphrase extraction dataset, the BibRank dataset, is presented to benchmark keyphrase extraction algorithms. The paper also introduces a state-of-the-art keyphrase extraction algorithm, BibRank, which utilizes the notion of context to compute keyphrases.

The main keyphrase extraction algorithms are comprehensively evaluated in the study using a two-fold approach: evaluating the algorithms against the gold standard and evaluating the gold standard itself. The evaluations are conducted on the BibRank dataset using

standard evaluation metrics. Expert evaluators assess the gold standard using a five-point scale. The results demonstrate that some algorithms, such as BibRank and PositionRank, outperform the gold standard in extracting relevant and specific keyphrases, while others, such as YAKE, achieve lower scores due to their statistical nature. This evaluation provides valuable insights into the strengths and weaknesses of different keyphrase extraction methods in the context of the academic literature.

The BibRank algorithm demonstrates state-of-the-art performance when evaluated against the gold standard. The authors encourage researchers to use the BibRank platform for evaluating their own keyphrase extraction algorithms. To ensure reproducibility, the BibRank platform, BibRank algorithm, and the BibRank dataset are publicly available (see the Data Availability Statement) for use by the research community. Platforms such as BibRank and other keyphrase extraction tools have the potential to operate alongside VosViewer. If the research community starts using BibRank, we will think about adding a plugin for integration with VosViewer.

Author Contributions: Conceptualization, E.B. and A.E.; methodology, A.E. and E.B.; software, A.E. All authors have read and agreed to the published version of the manuscript.

Funding: Eduard Barbu has been supported by the EKTB55 project "Teksti lihtsustamine eesti keeles".

Data Availability Statement: The BibRank keyphrase extraction framework is readily available on GitHub to facilitate reproducibility. The repository includes: The implementation of BibRank and 18 other keyphrase extraction methods; A detailed installation guide; Examples of evaluations; The Bib dataset used for evaluation; Comprehensive instructions for running experiments with the BibRank model; Reviewers' full evaluation results. GitHub repository: https://github.com/dallal9/Bibrank, (accessed on 3 October 2023).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Gusenbauer, M. Google Scholar to Overshadow Them All? Comparing the Sizes of 12 Academic Search Engines and Bibliographic Databases. *Scientometrics* **2019**, *118*, 177–214. [CrossRef]
2. Merrouni, Z.A.; Frikh, B.; Ouhbi, B. Automatic keyphrase extraction: An overview of the state of the art. In Proceedings of the 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, Morocco, 24–26 October 2016; pp. 306–313. [CrossRef]
3. Merrouni, Z.A.; Frikh, B.; Ouhbi, B. Automatic keyphrase extraction: A survey and trends. *J. Intell. Inf. Syst.* **2019**, *54*, 391–424. [CrossRef]
4. Mihalcea, R.; Tarau, P. Textrank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 404–411.
5. Hasan, K.S.; Ng, V. Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 23–25 June 2014; pp. 1262–1273.
6. Danesh, S.; Sumner, T.; Martin, J.H. Sgrank: Combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction. In Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, Denver, CO, USA, 4–5 June 2015; pp. 117–126.
7. Papagiannopoulou, E.; Tsoumakas, G. A review of keyphrase extraction. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1339. [CrossRef]
8. Caragea, C.; Bulgarov, F.; Godea, A.; Gollapalli, S.D. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1435–1446.
9. Li, D.; Li, S.; Li, W.; Wang, W.; Qu, W. A semi-supervised key phrase extraction approach: Learning from title phrases through a document semantic network. In Proceedings of the ACL 2010 Conference Short Papers, Uppsala, Sweden, 11–16 July 2010; pp. 296–300.
10. Page, L.; Brin, S.; Motwani, R.; Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. *Proc. ASIS* **1998**, *98*, 161–172.
11. Beliga, S.; Meštrović, A.; Martinčić-Ipšić, S. An overview of graph-based keyword extraction methods and approaches. *J. Inf. Organ. Sci.* **2015**, *39*, 1–20.

12. Florescu, C.; Caragea, C. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1105–1115.
13. van Eck, N.J.; Waltman, L.; Dekker, R.; van den Berg, J. A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2405–2416. [CrossRef]
14. Beebe, N.H. BIBTEX meets relational databases. *J. TUGboat* **2009**, *30*, 252–271.
15. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nD Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; pp. 55–60.
16. Schutz, A.T. Keyphrase Extraction from Single Documents in the Open Domain Exploiting Linguistic and Statistical Methods. Master's Thesis, National University of Ireland, Galway, Ireland, 2008.
17. Nguyen, T.D.; Kan, M.Y. Keyphrase extraction in scientific publications. In Proceedings of the International Conference on Asian Digital Libraries, Hanoi, Vietnam, 10–13 December 2007; pp. 317–326.
18. Hulth, A. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 11–12 July 2003; pp. 216–223.
19. Boudin, F. pke: An open source python-based keyphrase extraction toolkit. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan, 11–16 December 2016; pp. 69–73.
20. Frank, E. Domain-specific keyphrase extraction. In Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 31 July–6 August 1999; pp. 668–673.
21. El-Beltagy, S.R.; Rafea, A. Kp-miner: Participation in semeval-2. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010; pp. 190–193.
22. Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.; Nunes, C.; Jatowt, A. YAKE! Keyword extraction from single documents using multiple local features. *Inf. Sci.* **2020**, *509*, 257–289. [CrossRef]
23. Wan, X.; Xiao, J. CollabRank: Towards a collaborative approach to single-document keyphrase extraction. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, 18–22 August 2008; pp. 969–976.
24. Bougouin, A.; Boudin, F.; Daille, B. Topicrank: Graph-based topic ranking for keyphrase extraction. In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Nagoya, Japan, 14–18 October 2013; pp. 543–551.
25. Duari, S.; Bhatnagar, V. sCAKE: Semantic connectivity aware keyword extraction. *Inf. Sci.* **2019**, *477*, 100–117. [CrossRef]
26. Grootendorst, M. MaartenGr/KeyBERT. 2021. Available online: https://zenodo.org/record/4461265 (accessed on 28 September 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.