Keyness Analysis and Its Representation in Russian Academic Papers on Computational Linguistics: Evaluation of Algorithms

Maria Khokhlova
and Mikhail Koryshev

St Petersburg State University
Universitetskaya emb. 7-9-11,
199034 St Petersburg, Russia
m.khokhlova@spbu.ru, m.koryshev@spbu.ru

Abstract. Extraction of relevant lexis has gained significance as the amount of information is continuously growing with news, posts on social networks, reviews, academic papers, etc. piling up. Automated algorithms are needed to analyze texts to facilitate understanding of their content. The paper scrutinizes methods for keyword extraction in abstracts of Russian scientific texts on computational linguistics. Unsupervised algorithms based on statistics, graphs and machine learning principles are considered. The results are evaluated against the keywords assigned by authors themselves, followed by expert opinion. Log-likelihood produced the best results in comparison with author keywords, while KeyBERT implementation with vectorizers outperformed other algorithms according to expert assessment.

Keywords: Keyword extraction, Academic papers, Abstracts, Computational linguistics, Log-likelihood, TextRank, RAKE, YAKE, KeyBERT

1 Introduction

Keyword extraction has never been more relevant. Continuous increase in information volume makes it difficult for users to familiarize with all the emerging data. It is impossible to read all the papers on a particular topic or all the news. This challenge can be partly resolved by automated methods that allow us to grasp the main content of any texts.

Our study focuses on academic texts and tools to extract significant and meaningful phrases, in particular. We explore a number of methods and evaluate them using keywords tagged by authors or selected manually by experts.

The paper has the following structure: Section 2 provides a brief overview of related studies; Section 3 describes data and presents methods and relevant notions; Section 4 examines the results, followed by Section 4 that concludes the paper and outlines future perspectives.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 25–33, 2022.

2 Background Research

The term "key word" refers to "a word which occurs with unusual frequency in a given text" [1, p. 236], meaning that its frequency is unusual compared to a large reference corpus. Extraction of relevant words and phrases from texts is closely related to the tasks of computational linguistics dealing with automatic term and collocation extraction or named entity recognition. We can say that, to a certain extent, the methods that are used to solve them intersect (chi-squared, log-likelihood ratio [2]).

The simplest way to find significant phrases in a document is to make a list of n-grams of either lemmas or word forms ranging them according their occurrences. There is a class of measures based on the comparison between reference and focus (or domain-specific) corpora. This statistical approach was implemented in Wordsmith Tools [3] and then further applied in a number of dictionaries and software systems. Sketch Engine supports keyword and term extraction using its own score to compare frequencies of single word or multiword units in focus and reference corpora [4,5]. Depending on a specific parameter, the measure favors either low-frequency words (with high keyness and thus highly relevant for a focus corpus), or high-frequency words (with low keyness). Similarly, AntConc weights candidates in different corpora that a user can upload [6].

More elaborate statistical methods involve calculating frequencies not just in one document, but in a collection of them (for example, TF-IDF). The same principles underlie KPMiner [7] that additionally filters n-grams. The RAKE algorithm (Rapid Automatic Keyword Extraction) was introduced in [8] and extracts multilingual keywords represented by n-grams. It is characterized by assigning more weight to longer sequences of words. The algorithm applies stop words and a delimiter list calculating statistics to search for multi-word terms [ibid]. Based on frequencies and count of relations between words and phrases, the method estimates the weight for each candidate and ranks them according to the values. YAKE is a corpus- and language-independent algorithm that employs a mixture of linguistic and statistical features such as casing, word position, relatedness to context, frequency, and dispersion of lexical items in different sentences [9,10]. Automatic extraction with YAKE is based upon the assumptions about the behavior of terms in documents. Relevant keywords are supposed to be concentrated more at the beginning of a document. Or a large number of different terms that co-occur with the candidate word can be crucial to indicate its meaningless character.

Graph-based ranking methods have been successfully used in a number of applications, keyword extraction being one of them. TextRank [11] was proposed for two language processing procedures, namely, unsupervised keyword and sentence extraction. It ranks keyword candidates according to their positions in graphs. One of the recent algorithms is RaKUn [12] that merges similar words into meta-vertices, reducing the number of vertices, as well as edges. It computes load centrality measure that is based on the number of shortest paths for a given vertex and thus estimates the importance of vertices in graphs (i.e.

keywords). Most recent and state-of-the-art approaches belong to machine learning. Methods based on transformers seem to be the most promising ones, BERT being one of them. Its modification KeyBERT belongs to embedding-based methods that use word distributions and sentence representations. It was proposed in [13] and is based on the bidirectional pretrained BERT model. Candidate keyphrases are ranked according to the cosine similarity.

Keywords "capture" the essence of texts and thus their extraction from academic papers is in a focus of attention in a number of works. In [14], the authors extract noun phrases in scientific abstracts in English (the Inspec dataset) based on pos-tags to use the results in academic search systems. The authors evaluate approaches that allow them to weight candidate phrases and then apply the metrics to rank them in terms of average geometric mean, pointwise mutual information, tf-idf, and entropy-based measures. Bruches et al. [15] study methods for entity recognition and relation extraction applied to Russian texts on information technologies. The authors collected a corpus of abstracts and annotated manually fragments with terms (about 2,000 items) represented by noun phrases and semantic relations (620 items involving "cause", "compare", "isa", "partof", "synonyms" and "usage"). Nguyen & Zaslavskiy [16] deal with keyphrase extraction in papers written in Russian and English using sentence embeddings. They propose a supervised learning model that calculates scores estimating the quality of every keyword. LanAKey_Ru was proposed in [17] for keyword extraction in Russian papers on mathematical modeling. Based on n-grams, the algorithm employs stop lists for their filtering and evaluates relevance of noun phrases using statistical and linguistic features.

3 Data and Methodology

3.1 Data Collection

We collected a corpus of abstracts in Russian for the papers submitted to "Dialogue" (from 2017 to 2022). It is the largest conference on computational linguistics and intellectual technologies that focuses on the Russian language and is held annually in Russia [18].

The corpus of abstracts comprises about 27,000 words and contains texts of different lengths (as authors do not always follow the template). We encountered, oddly enough, a certain challenge in collecting texts in Russian: many high-ranked professional conferences in Russia pursue a widest possible audience, as well as indexing in international databases, and hence the majority of talks is given in English. Therefore, most articles are submitted in English, as well as texts published in proceedings. However, papers may contain abstracts in the Russian language upon authors' consideration.

3.2 Methods

In our study, we deal with a number of unsupervised methods for keyword extraction as they require no labeled training data. These methods rely on

statistics, embeddings, and graphs representing different approaches. Statistical measures involved joint frequency (chosen as baseline) and log-likelihood. Among other approaches we used YAKE, RAKE, TextRank, and KeyBERT. Preprocessing included stop words removal, as well as lemmatization and morphological annotation that were performed with pymorphy2 [19].

The quality of the first 100 candidates was evaluated in two ways: by comparing with author keywords (a predefined set of terms assigned by the authors themselves) and by expert evaluation.

4 Results

4.1 Author Keywords

Specialized dictionaries can be used for evaluation as a benchmark. For example, dictionaries of linguistics terminology. Dictionary by Akhmanova [20] is a recognized source for Russian, though unsuitable for such a rapidly developing field as computational linguistics due to a broad linguistic scope, on the one hand, and outdated material, on the other.

In our evaluation we consider the keywords that are given in the papers and were attributed by the authors themselves. This predefined set of terms is compared to extracted candidate keywords. In total, we collected 822 author terms. The most frequent ones are: *BERT* (14), *klassifikacija tekstov* 'text classification' (6), *korpus* 'corpus' (6), *korpusnaja lingvistika* 'corpus linguistics' (11), *lemmatizacija* 'lemmatization' (7), *morfologicheskij analiz* 'morphological analysis' (7), *nejronnye seti* 'neural networks' (6), *rechevoj korpus* 'spoken corpus' (5), *russkij jazyk* 'Russian language' (42), *semantika* 'semantics' (7). Among the examples we find both general linguistic terms and highly specialized ones that are typical for computational linguistics.

The authors assign keywords inconsistently and in their own way. The analysis revealed synonyms in the lists of keywords. For example, *vybor zagolovkov* 'choice of titles' vs *generacija zagolovkov* 'title generation', *generacija zagolovkov* 'title generation' vs *generacija novostnyh zagolovkov* 'news title generation', *summarizacija* 'summarization' vs *summarizacija tekstov* 'text summarization', *predobuchennye modeli* 'pretrained models' vs *predobuchennye jazykovye modeli* 'pretrained language models', *diskursivnye markery* 'discourse markers' vs *diskursivnye slova* 'discourse words', *semanticheskaja blizost'* 'semantic similarity' vs *semanticheskaja blizost' tekstov* 'semantic similarity of texts'. Different word forms within the same node term are identified, e.g. singular vs plural (*generacija teksta* 'generation of text' vs *generacija tekstov* 'generation of texts'). Shortenings and standard forms represent another example of using the same terms, e.g. *avtomaticheskaja morfologicheskaja razmetka* 'automatic morphological analysis' vs *avtomaticheskaja morforazmetka* 'automatic morphoanalysis', *avtomaticheskoe referirovanie tekstov* 'automatic summarization of texts' vs *avtoreferirovanie tekstov* 'autosummarization of texts'.

In computational linguistics, a large number of terms come from the English language, so in some cases we can find a transliteration of terms (for example,

gepping 'gapping', *embeddingi* 'embeddings'), and in some cases, duplication of existing ones (for example, *evaljuacija* 'evaluation' instead of *ocenka* 'evaluation', *simplifikacija* 'simplification' instead of *uproshhenie* 'simplification'). On the one hand, this may indicate that there is no established term, or the authors are influenced by their English-based scientific background and want to clarify what material the study is being conducted on, as well as indicate certain methods and separate their studies from previous ones. On the other hand, this inconsistency could be eliminated if an automatic system were used that would allow the selection of suitable words or phrases from a precompiled list.

4.2 N-grams and joint frequency

The most frequent bigram candidate terms in the "Dialogue" corpus of abstracts include (frequencies are given in parentheses): *russkij jazyk* 'Russian language' (134), *nabor dannyh* 'data set' (34), *jazykovaja model'* 'language model' (31), *imenovannaja sushhnost'* 'named entity' (26), *estestvennyj jazyk* 'natural language' (18), *nacional'nyj korpus* 'national corpus' (16), *semanticheskij sdvig* 'semantic shift' (15), *rechevoj akt* 'speech act' (15), *vektornoe predstavlenie* 'word embedding' (13), *mashinnoe obuchenie* 'machine learning' (12), *znachenie slova* 'word meaning' (12), *nejronnaja set'* 'neural network' (11), *baza dannyh* 'database' (9), *semanticheskij sketch* 'semantic sketch' (8), *morfologicheskij analiz* 'morphological analysis' (7), *diskursivnoje slovo* 'discourse marker' (7), *mehanizm vnimanija* 'attention mechanism' (7), *rechevoj sboj* 'speech failure' (7), *individual'noje razlichije* 'individual difference' (7), *kljuchevoje slovo* 'keyword' (6).

The most typical frequency lexemes in abstracts are: *jazyk* 'language' (266), *model'* 'model' (214), *russkij* 'Russian' (212), *tekst* 'text' (211), *korpus* 'corpus' (192), *zadacha* 'task' (177), *stat'ja* (163) 'paper', *rezul'tat* 'result' (149), *slovo* 'word' (146), *metod* 'method' (128), *rabota* 'work' (116), *dannye* 'data' (115), *issledovanie* 'syudy' (111), *znachenie* 'meaning' (103), *sorevnovanie* 'competition' (94), *kachestvo* 'quality' (94), *semanticheskij* 'semantic' (93), *osnova* 'base' (80), *podhod* (77) 'approach', *tip* 'type' (75).

We also outlined typical trigrams: *korpus russkogo jazyka* 'corpus of Russian' (17), *obrabotka estestvennogo jazyka* 'natural language processing' (9), *raspoznavanije imenovannyh sushhnostej* 'named entities recognition' (8), *nositel' russkogo jazyka* 'speaker of Russian' (6), *predobychennyje jazykovyje modeli* 'pre-trained language models' (5), *metody mashinnogo obuchenija* 'machine learning methods' (4), *vektornye predstavlenija slov* 'word embeddings' (4), *ponimanie estestvennogo jazyka* (4) 'natural language understanding', *upotreblenie roditel'nogo partitivnogo* (4) 'usage of partitive genitive', *rekurrentnye nejronnye seti* 'recurrent neural networks' (3), *izmenenie znachenija slova* 'change in word meaning' (3), *verbal'naja reakcija slushajushhego* 'hearer's verbal response' (3), *Odin rechevoj den'* 'One speaker's day' (a title of the project) (3), *semanticheskaja slozhnost' slova* 'semantic complexity of words' (3), *obnaruzhenie semanticheskih sdvigov* 'semantic shift detection' (3).

Most of the top-list for bigrams and trigrams do reflect the terminology of computational linguistics, while the list for unigrams reveal broader linguistic and science terms.

4.3 Log-likelihood

For log-likelihood measure, the output almost completely coincides with n-grams mentioned above, with a difference in ranking only: *russkij jazyk* 'Russian language', *nabor dannyh* 'data set', *jazykovaja model'* 'language model', *rechevoj akt* 'speech act', *semanticheskij sdvig* 'semantic shift', *nacional'nyj korpus* 'national corpus', *vektornoe predstavlenie* 'word embedding', *nejronnaja set'* 'neural network', *estestvennyj jazyk* 'natural language', *mashinnoe obuchenie* 'machine learning', *individual'noje razlichije* 'individual difference', *fonovoe znanie* 'background knowledge', *mehanizm vnimanija* 'attention mechanism', *kommunikativnaja neudacha* 'communication failure', *imenovannaja sushhnost'* 'named entity', *roditel'nyj partitivnyj* 'partitive genitive', *rechevoj sboj* 'speech failure', *izvlechenije otnoshenoj* 'relation extraction', *predmetnaja oblast'* 'subject area', *semanticheskij sketch* 'semantic sketch'. The measure achieved the best score for the top list of candidates (however, with bigrams only), when evaluated against author keywords, outperforming other methods.

4.4 YAKE

Opposed to other algorithms, YAKE ranks candidate terms in ascending order, i.e. the lower the score, the more relevant the keyword is. The algorithm outperformed the above-mentioned two methods by suggesting unigrams, bigrams and trigrams as candidates. Among the first 100 candidates, we found more than 50 percent represented by verb phrases and simple clauses (e.g. *stat'ja predstavljaet rezul'taty* 'the paper presents the results', *ispol'zovat' korpus tekstov* 'to use a text corpus', *predstavljat' rezul'taty sorevnovanija* 'to present competition results', *dannaja rabota posvjashhena* 'the work deals with', *rezul'taty pokazala model'* 'the model showed results'). This may indicate that YAKE can be used, for example, for summarization and similar tasks, since it extracts prefabricated and frequent chunks.

4.5 RAKE

We used *multi-rake* implementation [21] that supports Russian texts with minimum frequency for keywords equal to 2. The following candidates were extracted by implementing the algorithm: *obnaruzhenie novostnyh sobytij* 'event detection from news', *morfologicheski bogatyj jazyk* 'morphologically rich language', *verhnij sloj set'* 'top layer of a neural network', *predobuchennaja jazykovaja model'* 'pre-trained language model', *izvlechenie imenovannyh sushhnostej* 'named entity extraction', *sovremennyj russkij jazyk* 'modern Russian language', *znachenie obshhej neopredelennosti* 'value of the total uncertainty', *baza znanij wikidata* 'wikidata

database', *semanticheskij sdvig* 'semantic shift', *nabor dannyh* 'data set', *nejronnaja set'* 'neural network', *rechevoj akt* 'speech act', *analiz tonal'nosti* 'sentiment analysis', *semanticheskij klass* 'semantic class', *jazykovaja model'* 'language model', *imenovannaja sushhnost* 'named entity', *baza dannyh* ''database', *komp'juternaja lingvistika* 'computational linguistics', *trenirovochnye dannye* 'training data', *grammaticheskij priznak* 'grammatical feature'. As one can see, the algorithm extracts not only terms and keywords, but also free phrases. Nevertheless it produced one of the best results.

4.6 TextRank

TextRank was implemented with *summa* package [22]. This algorithm revealed a large number of unigrams (about 80 percent of the total candidate list) that represent such science terms as *model'* 'model', *zadacha* 'task', *rezul'tat* 'result', *metod* 'method', *issledovanie* 'study', etc. Despite preprocessing and lemmatization, TextRank revealed examples with typos and errors, thus showing the poorest results for both types of evaluation.

4.7 KeyBERT

KeyBERT algorithm was launched into two configurations – the default sentence transformers model (paraphrase-multilingual-MiniLM-L12-v2) and the distilled model (distiluse-base-multilingual-cased-v2). The models were already fine-tuned and could be applied to many languages, including Russian. The third scenario involved CountVectorizer with lists of stop words and pos-patterns.

KeyBERT with vectorizer extracted 4- and 5-grams (for example, *predobuchennaja transformennaja jazykovaja model'* 'pre-trained transformer language model' or *jazykovaja model' tipa transformer* 'transformer-based language model') and outperformed other algorithms. Better results compared to other KeyBERT implementations can be explained by more elaborate tuning.

4.8 Author keywords vs Expert evaluation

Expert evaluation shows higher precision for all algorithms compared to author keywords (Table 1 presents the results). This can be explained by the fact that instead of selecting the keywords to be assigned from a pre-set list, authors rely on their own consideration and occasionally may misindicate terminology units. In several cases, the candidate phrase was not labeled as a term by the expert, although it was marked among author terms (for example, the key phrase *dvizhenija golovy* 'head movements' that describes a paper focusing on records for a multimodal corpus).

Low results for comparison with author keywords in a number of cases deal with different lengths of the extracted candidates and author terms (we considered only a complete match). The latter group was mostly represented

Table 1: Precision for comparison with author keywords and expert evaluation.

Algorithm	Precision (author keywords)	Precision (expert)
Joint freq	0.11	0.28
Log-likelihood	**0.24**	0.35
YAKE	0.06	0.39
RAKE	0.32	0.54
TextRank	0.01	0.17
KeyBERT_paraphrase	0.01	0.12
KeyBERT_distiluse	0.00	0.13
KeyBERT_vectorizer	0.07	**0.58**

by bigrams or longer word combinations specified by authors themselves and, hence, comparison with this list of terms failed to show high scores. Moreover, author terms may not be found in abstracts, but only in papers themselves, and hence in future we need to evaluate the results across full texts.

5 Conclusion

Automatic keyword extraction cannot replace profound expert evaluation, but it can serve as an initial stage for analysis. Keywords extracted by way of automatic methods can be used to compile thesauri, as well as modern dictionaries, as numerous foreign vocabulary units and borrowings appear in scientific discourse. Lack of labeled data still makes it challenging to perform experiments using supervised machine learning methods. Thus, the collected data can be used for data annotation. At the same time, some models are often pre-trained on more general data (for example, news collections, Wikipedia, web texts), which may impair the quality of the results, making them different from what is desired. For this reason, the next step may suggest training the models on more relevant texts, including compiling a collection of academic papers on linguistics. To this end, the corpus requires significantly more data, considering how demanding are the accuracy requirements.

References

1. Scott, M. PC analysis of key words - and key key words. System, 25(2), 233–45 (1997)
2. Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics – Special Issue on Using Large Corpora, vol. 19, no. 1, 61–74 (1993)
3. Scott, M.: WordSmith Tools version 8 (64 bit version) Stroud: Lexical Analysis Software (2022)
4. Kilgarriff A. Simple maths for keywords, In: Proceedings of Corpus Linguistics Conference CL2009, University of Liverpool, UK (2009)
5. Sketch Engine, `http://sketchengine.eu`. Last accessed 6 Nov 2022
6. AntConc, `https://www.laurenceanthony.net`. Last accessed 6 Nov 2022

7. El-Beltagy, S.R., Rafea, A. KP-Miner: A Keyphrase Extraction System for English and Arabic Documents. Information Systems, 34(1), 132–144 (2009)
8. Rose, S., Engel, D., Cramer, N., Cowley, W. Automatic Keyword Extraction from Individual Documents. In M. W. Berry & J. Kogan (Eds.), Text Mining: Theory and Applications: John Wiley & Sons, 1–20 (2010)
9. Campos R., Mangaravite V., Pasquali A., Jorge A.M., Nunes C., Jatowt A. YAKE! Collection-independent Automatic Keyword Extractor. In: Pasi G., Piwowarski B., Azzopardi L., Hanbury A. (eds). Advances in Information Retrieval. ECIR 2018 (Grenoble, France. March 26 – 29). Lecture Notes in Computer Science, vol 10772, pp. 806–810 (2018)
10. Campos, R., Mangaravite, V., Pasquali, A., Jatowt, A., Jorge, A., Nunes, C. and Jatowt, A. YAKE! Keyword Extraction from Single Documents using Multiple Local Features. In: Information Sciences Journal. Elsevier, Vol 509, pp. 257–289 (2020)
11. Mihalcea, R., Tarau,P. TextRank: Bringing Order into Texts. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, vol. 4. pp. 404-–411 (2004)
12. Skrlj, B., Repar, A., Pollak, S. Rakun: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation. In International Conference on Statistical Language and Speech Processing, pp. 311–323 (2019)
13. Grootendorst M. KeyBERT: Minimal keyword extraction with BERT, `https://doi.org/10.5281/zenodo.4461265`. Last accessed 6 Nov 2022
14. Popova S., Khodyrev I. Ranking in keyphrase extraction problem: is it suitable to use statistics of words occurrences? In: Proceedings of the Institute for System Programming of RAS 26(4), 123–136 (2014)
15. Bruches E., Pauls A., Batura T., Isachenko V. Entity Recognition and Relation Extraction from Scientific and Technical Texts in Russian. In: Proceedings of the Science and Artificial Intelligence Conference, p. 41-–45 (2020)
16. Nguyen, Q. H., Zaslavskiy, M. Keyphrase Extraction in Russian and English Scientific Articles Using Sentence Embeddings. In Proceeding of the 28th Conference of Fruct Association, pp. 334–340 (2021)
17. Sheremet'eva, S., Osminin P. Metody i modeli avtomaticheskogo izvlechenija kljuchevyh slov [Methods and models for automatic keyword extraction] . In Vestnik of Sounth Ural State University. Series "Linguistics", vol. 12, no 1, pp. 76-–81 (2015)
18. International Conference on Computational Linguistics and Intellectual Technologies "Dialogue", `https://www.dialog-21.ru/en/`. Last accessed 6 Nov 2022
19. Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: Analysis of Images, Social Networks and Texts, pp. 320–332 (2015).
20. Akhmanova, O.: Slovar' lingvisticheskih terminov [Dictionary of linguistic terminology] . Izdatel'stvo "Sovetskaja enciklopedija", Moscow (1969)
21. Multi_rake package, `https://pypi.org/project/multi\protect\discretionary{\char\hyphenchar\font}{}{}rake/`. Last accessed 6 Nov 2022
22. Summa package, `https://pypi.org/project/summa/`. Last accessed 6 Nov 2022