DOI: https://doi.org/10.31341/jios.44.1.8 UDC 004.912:81
Open Access Original Scientific Paper

Analysis of Text Collections for the Purposes of Keyword Extraction Task

Alexander Vanyushkin *alexmandr@mail.ru*
Pskov State University
Pskov, Russia

Leonid Graschenko *graschenko@mail.ru*
Institute of Mathematics named A. Juraev of
the Academy of Sciences of the Republic of Tajikistan
Dushanbe, Tajikistan

Abstract

The article discusses the evaluation of automatic keyword extraction algorithms (AKEA) and points out AKEA's dependence on the properties of the test collection for effectiveness. As a result, it is difficult to compare different algorithms who's tests were based on various test datasets. It is also difficult to predict the effectiveness of different systems for solving real-world problems of natural language processing (NLP). We take in to consideration a number of characteristics, such as the text length distribution in words and the method of keyword assignment. Our analysis of publicly available analytical exposition text which is typical for the keywords extraction domain revealed that their length distributions are very regular and described by the lognormal form. Moreover, most of the article lengths range between 400 and 2500 words. Additionally, the paper presents a brief review of eleven corpora that have been used to evaluate AKEA's.

Keywords: text corpus, corpus linguistics, keyword extraction, text length distribution, natural language processing, information retrieval

1. Introduction

The number of digital documents available is growing on a daily basis at an overwhelming rate. As a consequence, there is a need to increase the complexity of the structure and software solutions in the field of NLP which are based on a number of basic methods and algorithms. The algorithms of automatic keyword and key phrase (KW) extraction are among them. This task has been analyzed over the past sixty years from different perspectives. There has been a significant increase in the number of research that took place in the last twenty years, of which many have been publications of different AKEA's [1]. The reason for this is the increasing amount of computing research, data resources and especially the development of internet services. It also simplifies the development and evaluation of new algorithms. This trend is clearly illustrated in Figure 1, obtained using Google Books Ngram Viewer[1].

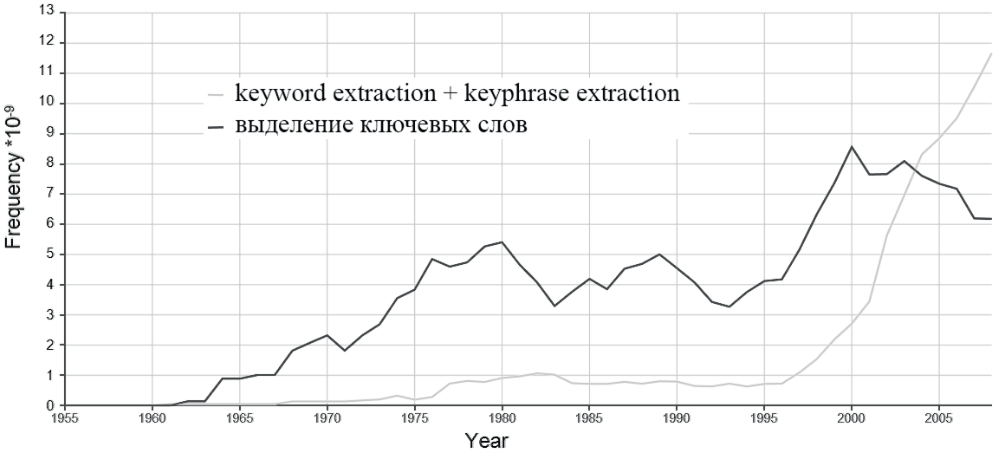


Figure 1. Usage of phrases 'keyword extraction', 'keyphrase extraction' (Russian: 'выделение ключевых слов') found in the Google Books Dataset.

The term "keyword" is interdisciplinary and above all, is used in works on psycholinguistics and Information Retrieval [2] that causes the existence of different approaches to its definition. Summarizing the numerous opinions, we can conclude that the keywords (phrases) are words (phrases) in the text that are especially important, commonly understood, capacious and representative of a particular culture. The set of which can give a high-level description of its content for the reader and providing a compact representation and storage of its meaning in mind [1]. In practice, the terms keyword and key phrase have the same meaning.

Despite the large amount of specialized and interdisciplinary work there has not been a consistent technique developed for detecting keywords yet. Experiments confirmed that this is done intuitively by people, and is personality, and even gender-based [3]. This implies the non-triviality of the development of formal methods and KW extraction algorithms for computing. Therefore, the current efforts of researchers are focused on the development and implementation of hybrid learning-based AKEA's which assumes the use a variety of linguistic resources. Thus, the accuracy of training and control datasets has great importance on the effectiveness of development.

Our analysis reveals number problematic areas. The author's results in testing AKEA's are often different from those obtained by other researchers, since they use different control data in the evaluation of algorithms [1]. Independent testing of KW extraction algorithms is a difficult task because there is a lack of implemented system and source code of algorithms in open access. This problem is partially solved by carrying out workshops when the organizers propose test data for all

[1] https://books.google.com/ngrams

participants. At the same time the number of available and well-proven corpora for KW extraction evaluation is small (10-20) and the criteria for their formation are not methodologically well enough investigated. The possibility of transferring the results of the algorithms in other languages remains an open question. The remarkable thing is that most of the known results are obtained for the English language, and the rules for the interpretation of them to the Slavic languages, especially to Russian, have not been established.

Indeed, preliminary empirical data show that for the graph-based algorithms with increased text size the precision of AKEA's might reduce. Therefore, the effectiveness of the algorithms depends on the type and parameters of the text lengths distribution (in words) that constitute research data. Homogeneity of the data by genre and text difficulty probably has some influence on the effectiveness of AKEA's too, Figure 2.

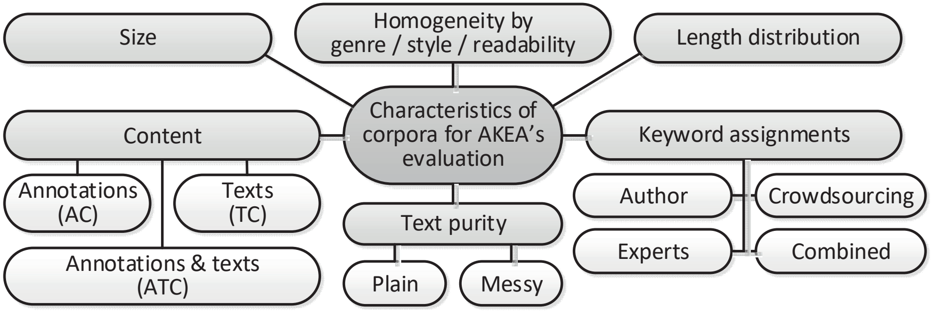


Figure 2. The specifications of research corpora for keyword extraction evaluation.

A separate discussion is necessary to explore the characteristics of experimental corpora such as size, existence and the methods of KW assignment (who and how many authors assigned them), the subject and the type of text (abstracts and full articles). KW assignment can be performed by authors, experts on the topic or by crowdsourcing. In this case, questions arise such as what kind of assignment is considered optimal, is it possible to rely on public opinion and what is a minimum number of participants that must specify the word as a keyword to assign it as such. It should be noted that the quality of KW assignment depends on the size of a corpus. As the size increases, the complexity of assignment rises.

But first of all it is necessary to investigate existing text collections (those used for KW extraction) for the length distribution parameters (in words).

2. Methodology and Research Tools

Articles from six web sites were selected as the statistical and research database subset that contains a voluminous collection on various English topics. This choice is due to the assumption that the main sphere of work for KW extraction is mostly with topical or subject-based text, especially those that contain elements of analytical themes. The eleven corpora (test and trial), that were used in some or other research or scholarly articles, were found using a search engine.

Many sites block automatic downloading for article collection or don't have freely available archives for use at all. So sites with freely available resources were used. After downloading the collection of articles, automatically parsing of the pages was made and the text was extracted. Then the tokenization and a count of the number of words in each article was made. Stanford Log-linear Part-Of-Speech Tagger[2] was used for tokenization of English texts, which is widely used in both research and commercial sectors [4].

The text lengths distributions in words were presented for every collection. We used Pearson's chi-squared test to evaluate the fitness of observed data to some theoretical distributions using advanced analytics software package Statistica[3] and EasyFit[4] software. It is worth pointing out that the form distribution depends on the mode of data grouping [5]. Calculating the number of bins *k* in different ways leads to a wide range of its possible values. For the expected Gaussian distribution, the Sturges formula is normally used, but if the data are not normal or there are more than 200 cases, it's poorly applied [6].

For the unification of the calculation the bin sizes in the histograms we used the Freedman and Diaconis rule, which gives the value agreed with the recommendations on standardization[5] and then convert it into the number of bins:

$$h = 2(IQ)n^{-\frac{1}{3}}, \tag{1}$$

where h is the bin size, IQ is the interquartile range of the data and n is the number of observations. At the same time according to the Pearson's chi-squared test (p-value = 0.05) we did not obtain a satisfactory fit of the results in all cases. Our hypothesis was confirmed by varying k in a small range with respect to the calculated value. To improve the accuracy of estimates of the form and parameters of the probability density function further research is needed. For example, the Levenberg-Marquardt algorithm was used by other researches to solve similar problems [7].

3. A Review of Existing Information Resources

3.1. Text Length Distributions in Analytical Articles Collections

The issue of natural length distribution and optimal lengths are taken into consideration by many researches. Most studies have been devoted to investigate blog post sizes [8], [9], [10], which describes the text length distribution with fat tails. This is true for the user comments [7], e-mail messages [11] and for the length

[2] http://nlp.stanford.edu/

[3] https://www.quest.com/

[4] http://www.mathwave.com/

[5] R 50.1.033-2001. Applied statistics. Rules of check of experimental and theoretical distribution of the consent. Part I. Goodness-of-fit tests of a type chi-square

of the texts that are stored on users' computers [12]. It is proposed [13] to consider the length of the articles from Wikipedia encyclopedia as an indicator of their quality, and the overall length of the English papers described by the lognormal form [14]. Figure 3 presents the probability density function distributions for the six data-sets.

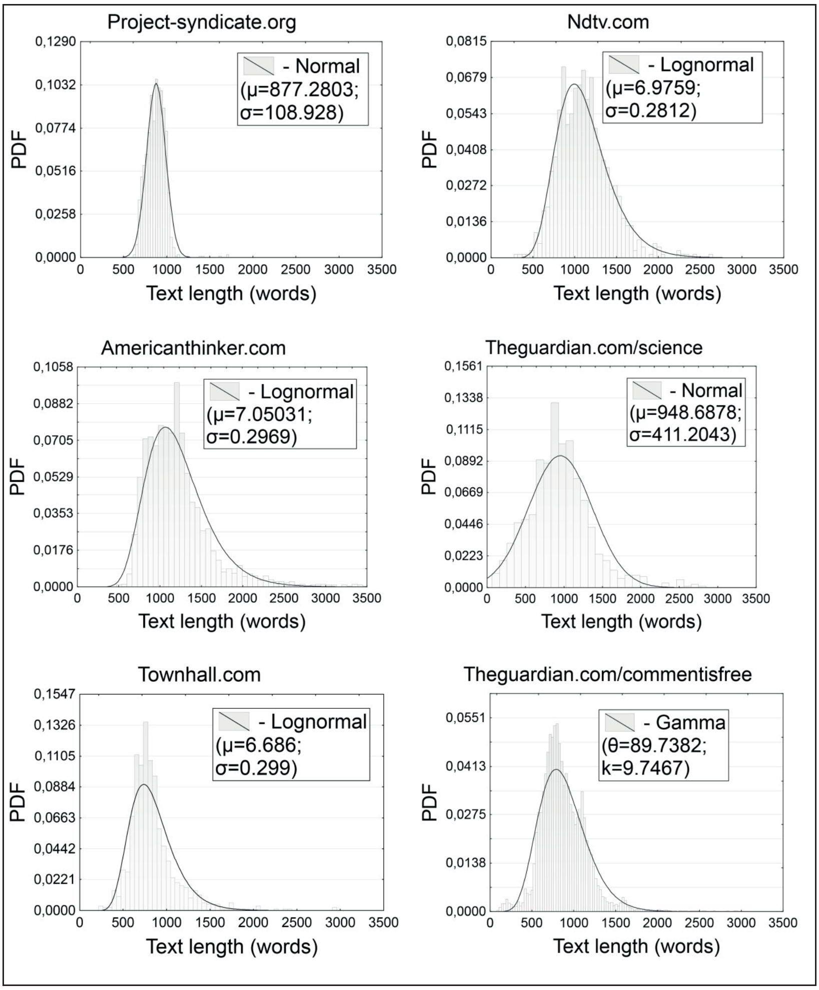


Figure 3. Distribution of analytical articles lengths in words.

As can be seen from the graphs, the majority of the length distribution of analytical articles can be comparative to the normal or lognormal form. The majority of texts are in the range of 400 to 2500 words. Figure 4 summarize probability density function distributions for the considered collections.

Table 1 presents general information and statistical characteristics of the reviewed text collections. Collection size ranges from 736 to 14529 articles and their publication dates cover the period from 2015 to 2016. Mean lengths of articles vary between 839-1212 words.

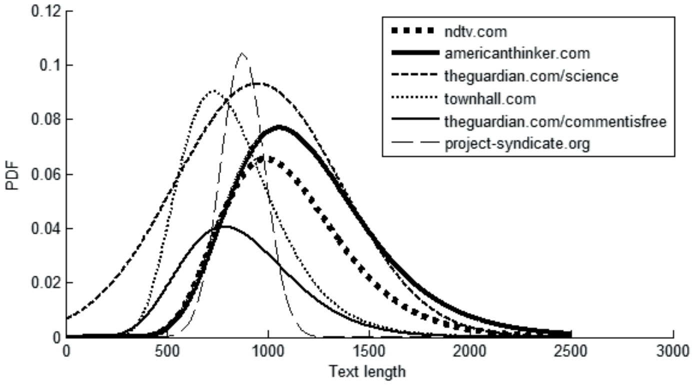


Figure 4. Distribution curves for analytical article collections lengths in words.

№	Source	Count	Text length				Publishing period
			Mean	Min.	Max.	Std. Dev.	
1	project-syndicate.org	1163	873,3	612	1721	108,9	01.15-12.15
2	ndtv.com	736	1112,5	274	2650	309,9	01.15-12.15
3	americanthinker.com	2268	1212,2	473	3703	410,4	01.15-02.15
4	townhall.com	905	839,5	217	2960	283,9	07.15-12.16
5	theguardian.com/ science	897	948,7	66	2848	411,2	01.15-12.16
6	theguardian.com/ commentisfree	14529	874,6	79	3045	278,8	01.15-12.16

Table 1. Characteristics of the analytical articles collections.

It is worth pointing out that there are possible restrictions authors can have on the length of published articles. For example, on project-syndicate.org a recommended article length by their editorial team is 1000 words.

3.2. Existing Corpora for Keyword Extraction Evaluation

Despite the large number of works devoted to keyword extraction evaluation the number of specially trained and public corpora are much less so. Some of them are used multiple times in different studies. Hulth-2003 [15] for example, consisting of abstracts of scientific articles, is one of the most popular and was used in the many academic papers [16], [17], [18], [19], [20], [21], [22]. Other datasets are used much less frequently, often only by their authors. One of the main drawbacks of such corpora is the "messy" texts, as many of them contain a bibliography, tables, captions and pictures in text files.

We surveyed eleven public corpora, which are significantly different from each other such as the text length distribution as well as other characteristics such as the size, themes and authorship of the keyword assignment. Table 2 summarizes the characteristics of reviewed corpora. The following are some explanations.

№	Corpus	Year	Contents	KW assign	Type	Resource
1	DUC-2001 [23]	2001	News articles	E-2	AT	github.com
2	Hulth-2003 [15]	2003	Paper abstracts from Inspec 1998-2002	E-?	A	researchgate.ne t
3	NLM-500 [24], [25]	2005	Full papers of PubMed documents	E-?	AT	github.com
4	NUS [26]	2007	Scientific conference papers	A+E-?	AT	github.com
5	WIKI-20 [27], [28]	2008	Technical research reports of computer science	E-15	AT	github.com
6	FAO-30 [28], [29]	2008	Documents from UN FAO[6]	E-6	T	github.com
7	FAO-780 [28], [29]	2008	Documents from UN FAO	E-?	T	github.com
8	KRAPIVIN [30]	2009	ACM[7] full papers 2003-2005	A	AT	disi.unitn.it
9	CiteULike [28], [31]	2009	Bioinformatics papers	O-3	T	github.com
10	SemEval-2010 [32]	2010	ACM full papers	A+E-0,2	AT	github.com
11	500N-KPCrowd-v1.1 [33]	2012	News articles	O-20	T	github.com

Note: notation of KW assignment: A-text authors, O-N – Crowdsourcing (N – number of people per one text, ? - n/a), E-experts.
Corpus type: A – annotation, AT – annotation + text, T – the main body of the text.

Table 2. Characteristics of the available corpora for KW extraction evaluation.

Let us explain the features of the KW assignment of the given corpora. DUC-2001 was prepared for text summarization evaluation within the Document Understanding Conferences, but KW assignment was made by two only graduate students in 2008 for the study of AKEA's [23]. A feature of the Hulth-2003 assignment is the presence of two sets of KW – a set of controlled, i.e. terms restricted to the Inspec thesaurus, and a set of uncontrolled terms that can be any terms. NLM-500 sets of keywords restricted to the thesaurus of Medical Subject Headings. WIKI-20 assigned by 15 teams consisting of two senior computer science undergraduates each. These KW sets were restricted to the names of Wikipedia

[6] Food and Agriculture Organization

[7] Association for Computing Machinery

articles. NUS has the author's assigned KW lists as well as KW lists assigned by student volunteers.

FAO-30 and FAO-780 differ in size and composition of the experts, but both KW sets were restricted to the Agrovoc[8] thesaurus. In KRAPIVIN parts of the articles are separated by special characters, which makes it convenient to their separate processing. CiteULike KW's were assigned by 322 volunteers but the authors noted that for this reason the high quality of the KW assignment is not guaranteed. For assignment of 500N-KeyPhrasesCrowdAnnotated-Corpus (500N-KPCrowd-v1.1) the researchers used the crowdsourcing platform Amazon's Mechanical Turk[9].

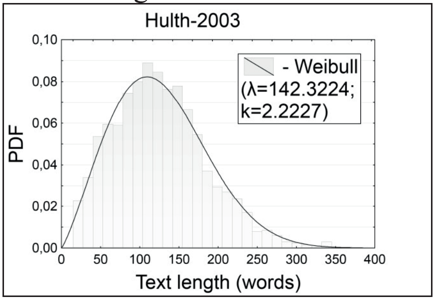
SemEval-2010 has been specially prepared for the Workshop on Semantic Evaluation 2010, where 19 systems were evaluated by matching their KW's against manually assigned ones. It consists of three parts: trial, training and test data. The authors note that on average 15% of the reader-assigned KW and 19% of the author-assigned KW's did not appear in the papers.

Table 3 shows the statistical characteristics of text length distributions in the reviewed corpora.

№	Name	Count	Mean	Min.	Max.	Std. Dev.
1	DUC-2001	307	769,1	141	2505	435,1
2	Hulth-2003	2000	125,9	15	510	59,9
3	NUS	211	6731,7	1379	13145	2370,6
4	NLM-500	500	4805	436	24316	2943,3
5	WIKI-20	20	5487,8	2768	15127	2773,4
6	FAO-30	30	19714,3	3326	70982	16101,6
7	FAO-780	779	30106,5	1224	255966	31076,5
8	KRAPIVIN	2304	7572,8	144	15197	2092,3
9	CiteULike	180	6454,1	878	23516	3408,9
10	SemEval-2010	244	7669,1	988	13573	2061,9
11	500N-KPCrowd-v1.1	447	425,9	38	1478	311,7

Table 3. Statistical characteristics for the datasets used in this paper.

Figures 5 - 9 shows the text length distribution of the reviewed corpora.



[8] http://www.fao.org/agrovoc

[9] https://www.mturk.com/

Figure 5. Distribution of annotation lengths in words.

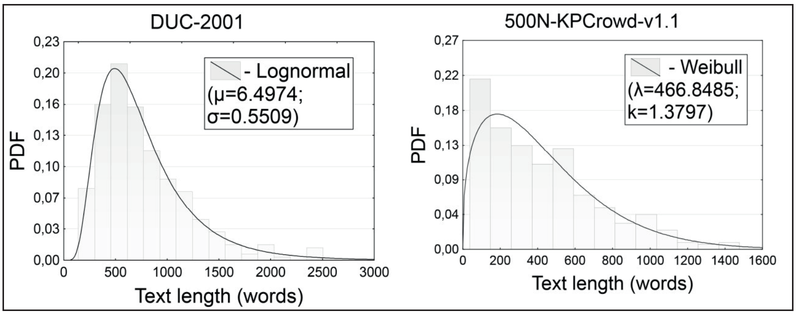


Figure 6. Distribution of news article lengths in words.

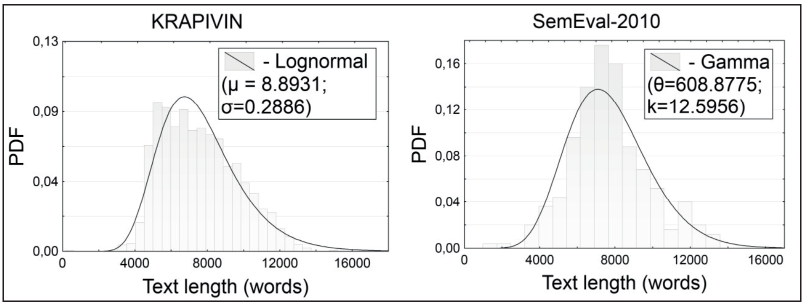


Figure 7. Distribution of ACM article lengths in words.

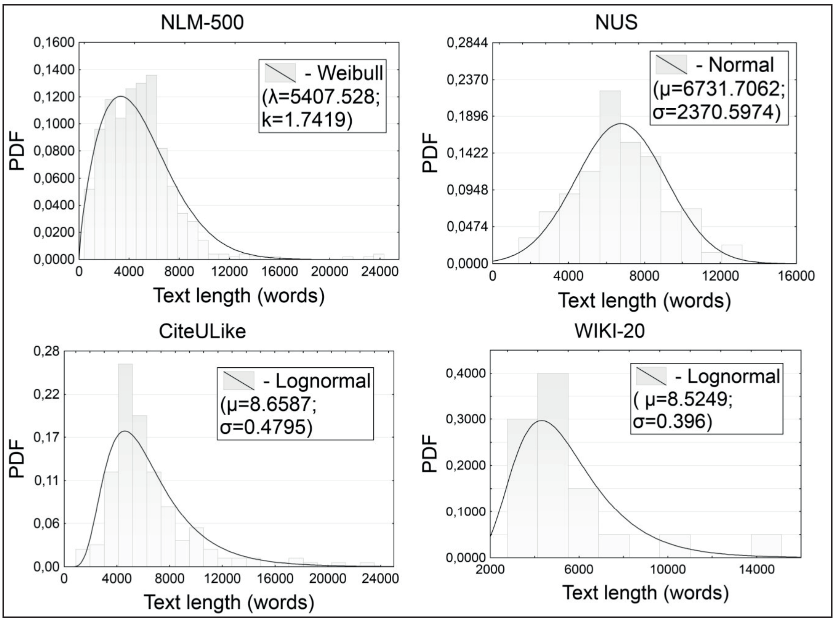


Figure 8. Distribution of Scientific paper lengths in words.

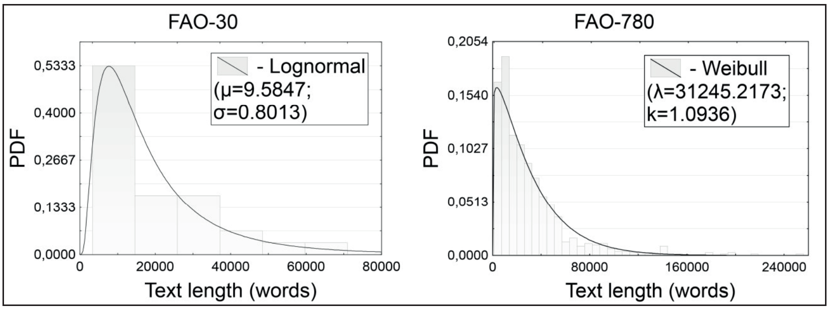


Figure 9. Distribution of FAO document lengths in words.

A review of test corpora revealed that they differ significantly on the sizes, the themes, and the method of keyword assignment. The difference of text lengths for some couples is three orders of magnitude. The text length in the tens of thousands of words questioned the possibility and the meaning of the use of AKEA's at its entire length, without division into semantic parts. In contrast, annotation in definition contain a higher percentage of KW's than text containing a few thousand words. Figure 10 summarize probability density function distributions for the considered datasets.

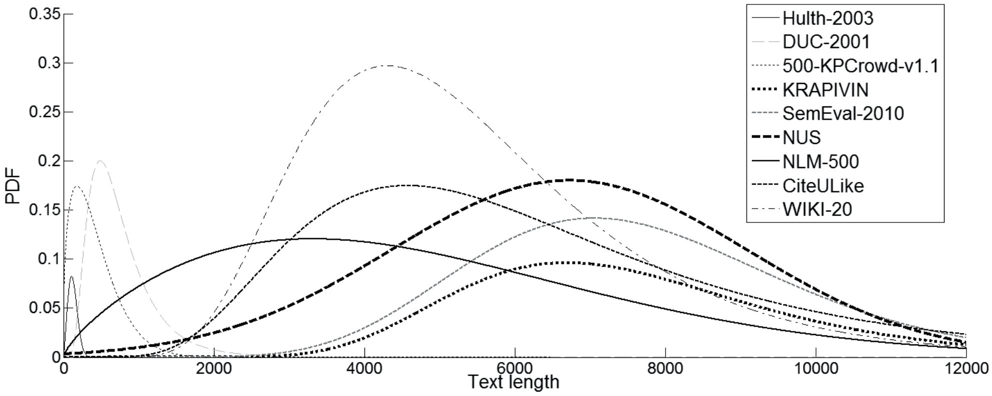


Figure 10. Distribution curves for datasets text lengths in words.

The text length distribution histograms of the most reviewed corpora have outliers, and does not correspond to the established in Section 3.1 principles, that is their apparent drawback. DUC-2001 has the most relevant form and distribution parameters (LN (6.49, 0.55)) but its disadvantage is the small number of experts participating in the KW assignment (only two). Moreover, all the above corpora are monolingual and do not allow carry cross-language study of KW extraction.

4. Conclusions

As can be seen from the above, the majority of the texts for which KW extraction is relevant are in the range of 400 to 2500 words and their text length distribution is quite well described by the lognormal form. Thus in practice it is advisable to use AKEA's that show a good performance in certain text length ranges. However, in general a comparison of existing AKEA's was performed on corpora with different characteristics. Moreover, the length of the manually assigned KW lists in them varies widely, and KW assignment was made by different categories of people such as students, volunteers and experts for example. Thus, for an objective comparison of existing AKEA, it is necessary to use corpora, whose characteristics are close to those of natural collections.

References

[1] A.S. Vanyushkin and L.A. Graschenko, “Methods and algorithms of keyword extraction [Metody i algoritmy izvlecheniya klyushevyh slov]” New information technology in the automated systems [Novye informacionnye tekhnologii v avtomatizirovannyh sistemah], pp. 85–93, 2016.

[2] E.V Yagunova, “Experiment and computation in the analysis of literary text’s keywords [Eksperiment i vychisleniya v analize klyuchevyh slov hudozhestvennogo teksta]” Collection of scientific works of the department of foreign languages and philosophy of PSC of UB RAS. Philosophy of Language. Linguistics. Linguodidactics [Sbornik nauchnyh trudov kafedry inostrannyh yazykov i filosofii PNC UrO RAN. Filosofiya yazyka. Lingvistika. Lingvodidaktika], pp. 85-91, 2010.

[3] T.G. Nozdrina, “Reconstructing original texts by key words [Osobennosti vosstanovlenija tekstov – originalov na osnove kljuchevyh slov]” Modern problems of science and education [Sovremennye problemy nauki i obrazovanija], Vol. 1-2, pp. 167-174, 2015.

[4] C.D. Manning, J.Bauer, J.Finkel, and S.J. Bethard, “The Stanford CoreNLP Natural Language Processing Toolkit” Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, pp. 55–60. 2014.

[5] B.Y. Lemeshko and S.N. Postovalov, “Limit distributions of the Pearson $\chi 2$ and likelihood ratio statistics and their dependence on the mode of data grouping” Industrial laboratory, Vol. 64, Iss. 5, pp. 344–351, 1998.

[6] R.J. Hyndman, “The problem with Sturges’ rule for constructing histograms” 1995. [Online]. Available at: http://robjhyndman.com/papers/sturges.pdf.

[7] P. Sobkowicz, M. Thelwall, K. Buckley, G. Paltoglou, and A. Sobkowicz, “Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law?” EPJ Data Science, Vol. 2, pp. 1–20, 2013.

[8] N. Kagan, “Why Content Goes Viral: What Analyzing 100 Million Articles Taught Us” 2013. [Online]. Available at: http://okdork.com/why-content-goes-viral-what-analyzing-100-millions-articles-taught-us.

[9] N. Patel, “Why 3000+ Word Blog Posts Get More Traffic (A Data Driven Answer).” [Online]. Available at: http://neilpatel.com/blog/why-you-need-to-create-evergreen-long-form-content-and-how-to-produce-it.

[10] T. Tunguz, “The Optimal Blog Post Length to Maximize Viewership” 2013. [Online]. Available: http://tomtunguz.com/content-marketing-optimizatio.

[11] V. Paxson, “Empirically-Derived Analytic Models of Wide-Area TCP Connections” IEEE/ACM Transactions on Networking, Vol. 2, Iss. 4, pp. 316–336, 1994.

[12] J.R. Douceur and W. J. Bolosky, “A Large-Scale Study of File-System Contents” Proceedings of the 1999 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, Atlanta, pp. 59–70, 1999.

[13] J.E. Blumenstock, “Size matters: word count as a measure of quality on wikipedia” Proceedings of the 17th International Conference on World Wide Web, Beijing, pp.1095–1096, 2008.

[14] M.A. Serrano, A. Flammini, and F. Menczer, “Modeling statistical properties of written text” PLoS One, Vol. 4, Iss. 4, pp. 1–8, 2009.

[15] A. Hulth, “Improved Automatic Keyword Extraction Given More Linguistic Knowledge” Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing, Sapporo, pp. 216–223, 2003.

[16] K.S. Hasan and V. Ng, “Conundrums in unsupervised keyphrase extraction: making sense of the State-of-the-Art” Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference, Beijing, pp. 365–373, 2010.

[17] R. Mihalcea and P. Tarau, “TextRank: Bringing order into texts” Proceedings of EMNLP 2004,. Barcelona, pp. 404–411, 2004.

[18] S.V Popova and I.A. Khodyrev, “Tag lines extraction and ranking in text annotation [Izvlechenie i ranzhirovanie kljuchevyh fraz v zadache annotirovanija]” Scientific and technical journal of information technologies, mechanics and optics [Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki], Vol. 1, pp. 81-85, 2013.

[19] F. Rousseau and M. Vazirgiannis, “Main core retention on graph-of-words for single-document keyword extraction” Advances in Information Retrieval, Vienna, pp. 382–393, 2015.

[20] N. Schluter, “Centrality Measures for Non-Contextual Graph-Based Unsupervised Single Document Keyword Extraction,” In Proceedings of TALN 2014, Marseilles, pp. 455–460, 2014.

[21] G. Tsatsaronis, I. Varlamis, and K. Norvag, “SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs” Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, pp. 1074–1082, 2010.

[22] T. Zesch and I. Gurevych, “Approximate Matching for Evaluating Keyphrase Extraction” Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets, pp. 484–489, 2009.

[23] X. Wan and J. Xiao, “Single Document Keyphrase Extraction Using Neighborhood Knowledge” Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, Chicago, pp. 855–860, 2008.

[24] A.R. Aronson, J.G. Mork, W.G. Clifford, S.M. Humphrey, and W.J. Rogers, “The NLM Indexing Initiative’s Medical Text Indexer” Studies in Health Technology and Informatics, Vol. 107, pp. 268–272, 2004.

[25] C.W. Gay, M. Kayaalp, and A.R. Aronson, “Semi-automatic indexing of full text biomedical articles” AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, pp. 271–275, 2005.

[26] T. Nguyen and M. Kan, “Keyphrase extraction in scientific publications” Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers, Hanoi, pp. 317–326, 2007.

[27] O. Medelyan, I.H. Witten, and D. Milne, “Topic Indexing with Wikipedia” Proceedings of the Wikipedia and AI workshop at AAAI-08, Chicago, pp. 19–24, 2008.

[28] O. Medelyan, I.H. Witten, and D. Milne, “Topic Indexing with Wikipedia (Thesis)” The University of Waikato, Hamilton, 2009.

[29] O. Medelyan and I.H. Witten, “Domain Independent Automatic Keyphrase Indexing with Small Training Sets” Journal of the American Society for Information Science and Technology, Vol. 59, Iss 7, pp. 1026–1040, 2008.

[30] M. Krapivin, A. Autayeu, and M. Marchese, “Large Dataset for Keyphrases Extraction” 2009. [Online]. Available at: http://eprints.biblio.unitn.it/archive/00001671/01/disi09055-krapivin-autayeu-marchese.pdf.

[31] O. Medelyan, E. Frank, and I. H. Witten, "Human-competitive tagging using automatic keyphrase extraction" Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 1318–1327,2009.

[32] S. Kim, O. Medelyan, M. Kan, and T. Baldwin, "Semeval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles" Proceedings of the 5th International Workshop on Semantic Evaluation, Los Angeles, pp. 21–26, 2010.

[33] L. Marujo, A. Gershman, J. Carbonell, and R. Frederking, "Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization" In 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, pp. 399–403, 2012.