AUTOMATED SYSTEM FOR EVALUATION OF TEXTS NATURALNESS

A.V. Yurasov, O.A. Degtiareva

Samara National Research University, Samara, Russia

Abstract. This paper describes research results on naturalness of texts as well as laws and algorithms the developed system is based upon. This paper should provide understanding of change of quantitative characteristics of texts after application of Zipf's first law.

Keywords: text naturalness, Zipf's laws, abstract, sentence, word, stemming, statistics.

Citation: Yurasov AV, Degtiareva OA. Automated system for evaluation of texts naturalness. CEUR Workshop Proceedings, 2016; 1638: 857-863. DOI: 10.18287/1613-0073-2016-1638-857-863

Introduction

Most texts we encounter in real life are natural. An example of an artificial text is a text created for web crawlers to improve site's rank on the search engine results page. Search engines fight this 'black' optimization actively, to the point of excluding such sites from the search index.

Zipf's laws describe regularities of frequency distribution of words in a text written in any natural language. These laws are empirical: they have no strict mathematical proof and are based on statistical distribution of words in large corpora of texts in various languages. Nevertheless, their correctness is proven statistically.

1 Task formulation

The object of the research is a text with a large percentage of repeating words and abstracts of it. The purpose of the research is to study changes of percentage of key words in the texts of an original abstract and an abstract based on statistics calculated using Zipf's first law.

It is the authors' opinion that the percentage of key words defined by the user separately for each given text should increase in the target abstract as compared with the original abstract.

2 Zipf's first law

The Zipf's first law [1, 2] correlates notions of word rank and frequency, where "word frequency" is the number of appearances of a word in a text and "rank" is the position of a word in the total list of words ranked by frequency. For any text written by a human, this law is true from the statistical, not mathematical point of view. [2] This means that deviations are possible for small texts, but the more words a text contains, the smaller such deviations are.
Zipf's first law (1) states that the probability of discovery of any word multiplied by its rank is a constant (C).

$$C = P * r, \tag{1}$$

where C is a constant, r is word rank, P is probability of discovery of a given word in a text. And C coefficient is different for different languages [2]. In this paper coefficient for Russian language was used.
Probability P is defined by (2).

$$P = \frac{f}{N}, \tag{2}$$

where P is probability, f is frequency, and N is the total number of words.

3 Stemming algorithm

The first published stemmer was written by Julie Beth Lovins in 1968. A later stemmer was written by Martin Porter and was published in 1980. This algorithm is implemented in stemmer [3]. Some stemmers can be automatically generated with the specified algorithms [4, 5].
An approximate heuristic process of removing suffices and inflections from a word is usually called stemmatization [6]. Stemming often involves removal of derivate affixes. An affix is a morpheme that is attached to a word root to create new words.
Words in Russian may be very short [3]. There are also many particles, conjunctions, etc. These words fall under the natural 'stop words' category. This gives a reason to create a filter to exclude these words from analysis. In this research, words were filtered by length. No ideal and recommended numbers of their appearances in texts were calculated for words that were filtered out. These words did not influence the process of generating abstracts.

4 System operation order

At the first stage, the system uses the stemming algorithm to flag words and calculates numbers of their appearances.
At the second stage, the system uses the generated statistics of numbers of words in the text to calculate ideal and recommended number of appearances of every word in

the text according to (3). According to Zipf's first law, the chart of words distribution in a natural text must be approximate to the graph of negative correlation.

$$y = \frac{a}{x} + b, \tag{3}$$

where x is word rank, y is word frequency.

The least square method was used in the research to find coefficients a and b.

If the ideal value is higher than the current number of appearances of a word in a text, then the recommended value will be the ideal one rounded down to the nearest whole number. If the ideal value is less, then the recommended one will be rounded up to the nearest whole number.

Fig. 1 shows a chart of word distribution in a text. The red line shows initial rank-frequency distribution of the analyzed document. The green line shows ideal values of word appearances in the text. According to Zipf's first law, those are non-integral values. The blue line shows integral values used for generation of the second abstract.

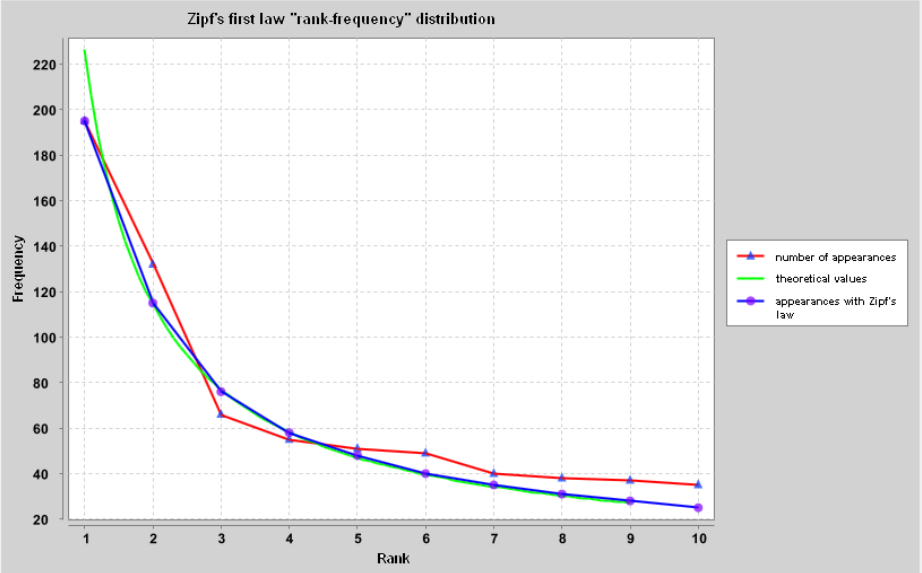

Fig. 1. Rank-frequency distribution

Next, two abstracts are generated: the first is based on initial numbers of appearances of words in the text, and the second is based on recommended values of appearances of words. The abstract generation algorithm is described in detail in [6, 7]. In both cases, the abstracts are generated by selection of a specified number of sentences with the highest weight in order of appearance in the text. Weight of a sentence is the sum of the number of appearances of its words.

Next, key words characteristic of the analyzed text are chosen. The number of appearances of these words in generated abstracts is calculated. Next, the statistics is displayed. An example of the statistics is shown in Fig. 2.

The red column shows the number of appearances of a specific word in the original abstract. The yellow column shows the number of appearances of the word in the target abstract generated using new values of appearances of the word in the text.

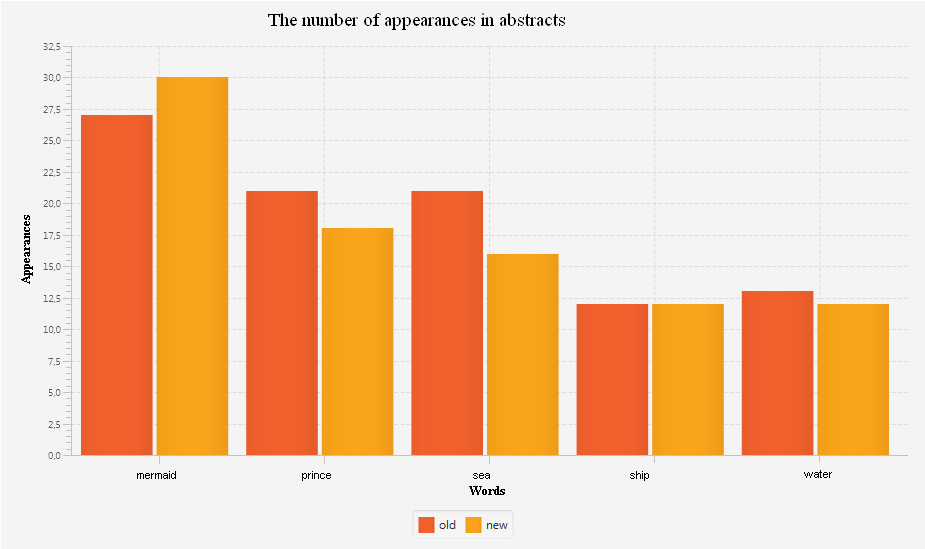

Fig. 2. Statistics of appearances of key words in abstracts

5 Results

Fairytales were chosen as objects of research of naturalness of texts because, despite a relatively small volume, key words appear frequently in them and are different for every separate fairytale. The key words were: Ivan, Tsar, Koschei, Vasilisa, princess, etc. These key words were supposed to vividly demonstrate changes in the content of an abstract after words in the texts were approximated to natural distribution.

5.1 Dependence of text naturalness on length of ignored words

Table 1 shows research results on dependence of text naturalness on length of ignored words during system operation. It must be remembered that short words appear more frequently in a text. The table shows the impact of frequent short words on the naturalness of text.

If the length of ignored words is increased by one character, text naturalness increases correspondingly by 1.8% on average. The text naturalness was calculated as percentage of words to all words of the text for which the recommended number of appearances corresponded with number of appearances in the original text.

Table 1. Dependence of text naturalness on length of ignored words

Length of ignored words, characters	Naturalness, %
2	73.1
3	74.4
4	76.8

5.2 Dependence of the number of key words in abstracts on the size of abstracts for small texts

Fig. 3 shows dependence of the number of key words in original (old) and new abstracts on the number of sentences in abstracts. It is evident from the figure that the assumption of increase of key words in the second abstract is not true for short (up to 20,000 words) texts. Changes are rather random. This mostly depends on the way of generation of an abstract. Frequent words, such as to be, to go, to become, etc., and their forms continue to influence the sentence weight.

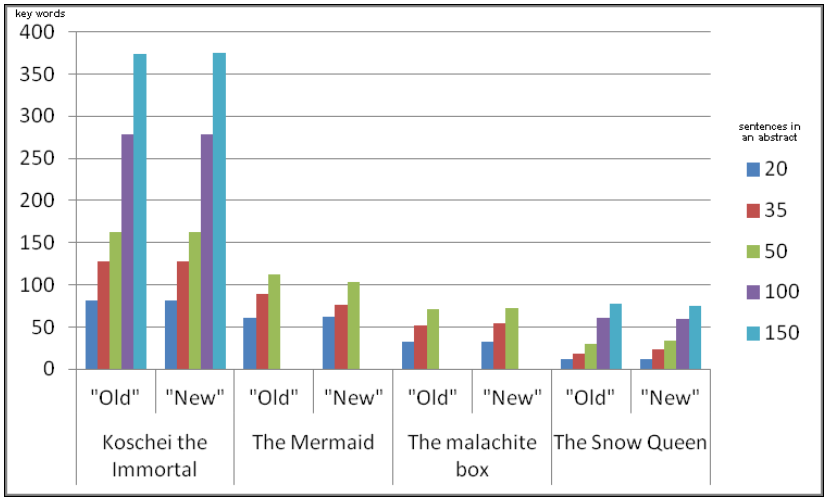

Fig. 3. Dependence of the number of key words in abstracts on the size of the abstracts for short texts

However, in long texts, such as “Tale of Cipollino” (Fig. 4), the content of key words increases correspondingly with the increase of the abstract length. This is mainly due to increase in appearances of the first three most frequent key words.

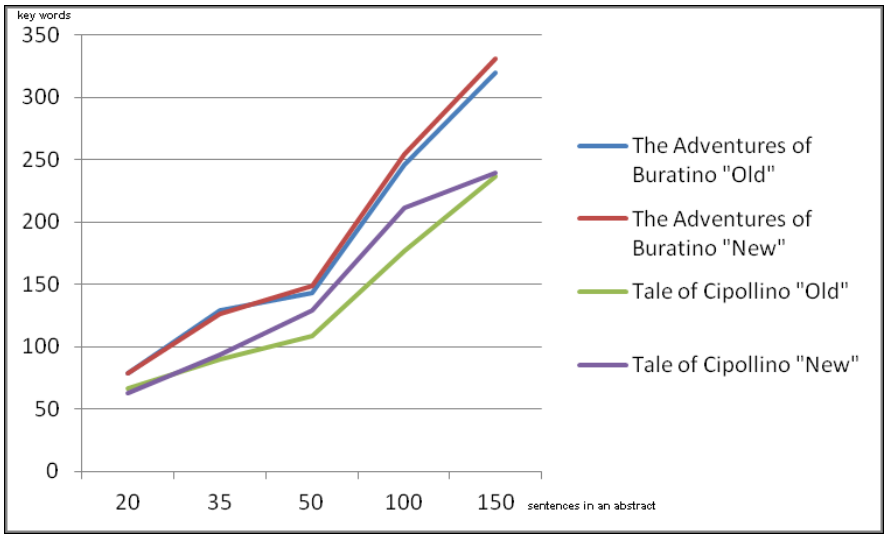

Fig. 4. Dependence of the number of key words in abstracts on the size of the abstracts for long texts

Conclusion

The paper has reviewed Zipf's first law on distribution of words in natural texts as well as an algorithm of stemming used to flag words. The research results on dependence of key words percentage in abstracts generated on the basis of recommended number of word appearances in the text were presented. This research allows to conclude that the application of Zipf's first law for short texts does not lead to increase of key words percentage in new abstracts. It is reasonable to apply this law to longer texts. Also, the work on mitigating the influence of non-informative words on content of the abstract, such as exclusion of frequent verbs and other parts of speech, should be continued. The stemming algorithm provides an opportunity to find parts of speech by specific inflections.

References

1. Ferrer I, Cancho R, Sole RV. Zipf's Law and Random Texts.Advances in Complex Systems, World Scientific Publishing Co , 2002; 5(1); 1-6.
2. Gelbukh A, Sidorov G. Zipf and Heaps Laws' Coefficients Depend on Language. Lecture Notes in Computer Science, Springer-Verlag GmbH, 2001; 2004: 332-335.
3. Russian stemming algorithm. URL: http://snowball.tartarus.org/algorithms/russian/ stemmer.html (access date 18.02.2016). [in Russian]
4. Fox B, Fox CJ. Efficient Stemmer Generation. Information Processing & Management, Elsevier Science Publishing Company, Inc., 2002; 38(4): 547.

5. Silva G, Oliveira C. A Lexicon-Based Stemming Procedure. Lecture Notes in Computer Science, Springer-Verlag GmbH, 2003; 2721:159-166.
6. Trusov V. Construction of thesauruses, classification and thematic categories for finding information in distributed information systems. URL: http://www.aselibrary.ru /digital_resources/journal/irr/irr2725/irr27253027/irr272530273030/irr272530273030303 (access date 18.02.2016). [in Russian]
7. Yatsko V, Vishniakov T. Some problems of development of modern automatic text abstracting systems. Scientific and Technical Information. Series 2: Information Processes and Systems, Moscow: VINITI RAS, 2007; 9: 7-13. [in Russian]