DOI: https://doi.org/10.30837/csitic52021232889

Method of Annotating a Collection of Text Documents

Olesia Barkovska[1]

Vitalii Vodolazkyi[2]

[1,2]Kharkiv National University of Radio Electronics, 14 Nauky Ave, Kharkiv UA-61166, Ukraine, d_ec@nure.ua

Abstract. *The work is devoted to the analysis of methods for annotating text documents, the relevance of which is due to the fact that when familiarizing with the information object presented in text form, reading the annotation by the reader is very much in demand, since it can reduce the time for selecting the necessary sources by several times. The paper considers the method of automatic annotation SumBasic, based on the probabilistic approach. An approach to data decomposition in each of the separate modules that ensure the operation of the method is proposed..*

Keywords: *methods, summarization, text documents, parallelization, speedup, partition, efficiency.*

I. Introduction

The development of information technology and the emergence of the Internet have led to the exponential growth in the volume of electronic information, which began about two decades ago and is rapidly continuing today. Among the tasks related to working with textual information, which provide ease of selection of the necessary sources for the reader, one can single out text annotation, translation, search for a word-image in the text, categorization, etc [1,2]. Most of the early work on automatic annotation was about annotating a single document, that is, a single document acts as input. Later, with the development of research in the field of automatic annotation, as well as the emergence of a large number of new sources of information and an increase in information flows in general, a new type of automatic annotation problem arose: preparing a review abstract for a collection of documents). This type of annotation is most in demand when processing a large number of text documents related to some storyline, theme, or some other parameter. The document name consists of the following.

II. Research Task Rationale

The task of automatic annotation - the creation of a short version of a text document or a collection of documents, which represents the most relevant and most significant information that the user needs, in a concise, concise form is an urgent task, since the amount of information has already reached such dimensions that a person is not able to independently familiarize yourself with materials from all information sources, often even in the context of specialized information needs. A brief summary of a text document - an annotation - comes to the rescue. The potential range of applications for automatic annotation systems is already extremely wide and continues to grow, along with the development of artificial intelligence systems, computational linguistics and automatic information processing systems in general.

III. Aims and Tasks for the Work

The purpose of this work is to study and accelerate the methods of automatic annotation of a collection of texts using the available functions of the programming language, without deteriorating the quality of the annotation received after processing.

IV. Task Fulfillment

Word likelihood is the simplest use of frequency to determine the significance of a word [3]. It is calculated as the ratio of the number of occurrences of a word to the total number of words in a document or collection of documents. This weighting system is the basis of the automatic annotation method SumBasic [3-5], which selects sentences for annotation based on the average probability of the words that are included in it. Considering a method based on the use of frequency characteristics of words, the execution speed can be increased by breaking up parts of the program into separate parallelized blocks. Figure 1 shows a simplified model of the interaction of the main stages of the SumBasic method, taking into account data decomposition based on data parallelism. The modules work asynchronously, which allows you to get results immediately after the completion of the modules, and eliminates unnecessary delays and expectations. The model has four modules that can be implemented in any available programming language: – module for the accumulation of initial data. The module is responsible for receiving a raw collection of documents and placing them in a queue for execution, taking into account the availability of threads, for the number of threads available for simultaneous operation we will count the number of central processor cores; – module for calculating the frequency of occurrence of a word in a document; – module for constructing annotations based on the frequency of occurrence of a word in a document; – module for matching annotations with source text documents. An additional task implemented in the module is writing the generated annotation to the resulting text files.

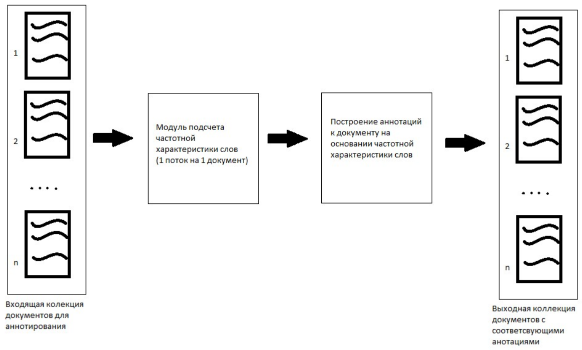
1
2
....
n
Модуль подсчета частотной характеристики слов (1 поток на 1 документ)
Построение аннотаций к документу на основании частотной характеристики слов
1
2
....
n
Входящая колекция документов для аннотирования
Выходная коллекция документов с соответсвующими анотациями

Figure 1. Simplified model of interaction of modules taking into account data decomposition for the SumBasic method.

For automated methods for assessing the quality of automatic annotation, correlation with expert estimates is important. The paper [6] provides an assessment of the cross-correlation of various ROUGE quality measures, a pyramid assessment, as well as a manual assessment of automatic annotations (responsiveness). The following ROUGE quality measures were compared: ROUGE-n (n = 1, 2, 3, 4), ROUGE-L, ROUGE-W-1.2, ROUGE-SU4, presented in table 1. The Pyramid Scoring Method for Automatic Annotations was developed by Columbia University in 2005. This method is based on manual selection by experts of "information units" from the reference annotations - Summary Content Units (SCUs). Each SCU represents a quantum of information that, in the opinion of the expert, should also be reflected in the automatic annotation. The ROUGE quality measure comparison procedure is based on comparison with manual scores of automatic annotations: for each pair of competition systems, results obtained from manual scores, pyramid scores and scores on various ROUGE quality measures are cross-checked. The main characteristic when assessing test pairs was prediction accuracy, which characterizes the percentage of agreement between these test documents. In addition to the accuracy characteristic, similarity measures such as precision, recall and balanced accuracy were calculated [6].

Table 1. Results of assessing the quality of automatic annotation based on ROUGE quality measures according to the characteristics of Accuracy (A), Precision (P), Recall (R) and Balanced Accuracy (BA)

	Responsiveness				Pyramid			
Metric	Acc	P	R	BA	Acc	P	R	BA
R1	0.58	0.24	0.64	0.57	0.62	0.37	0.67	0.61
R2	0.64	0.28	0.60	0.59	0.68	0.43	0.63	0.64
R3	0.70	0.31	0.48	0.60	0.73	0.49	0.53	0.66
R4	0.73	0.33	0.40	0.60	0.74	0.50	0.45	0.65
RL	0.50	0.20	0.56	0.54	0.54	0.29	0.60	0.55
R-SU4	0.61	0.26	0.61	0.58	0.65	0.40	0.65	0.63
R-W-1.2	0.52	0.21	0.54	0.55	0.57	0.32	0.62	0.57

The work shows that all of the available ROUGE quality measures can give qualitative results of modeling manual estimates (depending on the specifics of the input text collections, different quality measures can show different results).

V. Conclusion

As a result of the work, methods of automatic annotation of text documents were investigated, a simplified model of interaction of the main stages of the SumBasic method was proposed, taking into account data decomposition based on data parallelism, which accelerates the work of methods for automatic annotation of a collection of texts using the available functions of the programming language, without deteriorating the quality of the annotation received after processing. The work also analyzed and evaluated the quality of automatic annotation based on ROUGE quality measures according to the characteristics of Accuracy (A), Precision (P), Recall (R) and Balanced Accuracy (BA), which showed that all ROUGE quality measures are necessary. for computation when conducting a comprehensive assessment of automatic annotation.

REFERENCES

[1] Serdechnyi, V., Barkovska, O., Rosinskiy, D., Axak, N., Korablyov, M., Model of the Internet Traffic Filtering System to Ensure Safe Web Surfing, Advances in Intelligent Systems and Computing, 2020, 1020, стр. 133–147.

[2] Olesia, B., Oleg, M., Daria, P., ...Vladyslav, D., Maxim, V. Local concurrency in text block search tasks, International Journal of Emerging Trends in Engineering Research, 2020, 8(3), стр. 690–694.

[3] Vanderwende L., Suzuki H., Brockett C., Nenkova A. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion // Information Processing and Management Journal. - 2007. - Vol. 43, № 6. – P. 1606-1618. URL: http://citeseerx.ist.psu.edu/viewdoc/down load?doi= 10.1.1.105.9491 &rep=rep 1 &ty pe=pdf

[4] Vanderwende L., Suzuki H., Brockett C. Microsoft Research at DUC2006: Task-Focused Summarization with Sentence Simplification and Lexical Expansion // Proceedings of the Document Understanding Conference. - 2007. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.114.2486&rep:=repl&ty pe=pdf

[5] Nenkova, A. and L. Vanderwende. The impact of frequency on summarization // Microsoft Research Technical Report, MSR-TR-2005-101. - 2005. URL: http://www.cs.bgu.ac.il/~elhadad/nlp09/sumbasic.pdf

[6] Rankel P., Conroy J., Dang H., Nenkova A. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art // Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics. - 2013.-P. 131-136. URL: http://aclweb.Org/anthology/P/P13/P13-2024.pdf