ANNOTATION

for dissertation of Akhmetov Iskander Rafailovich work of on the topic: "Development of a method for informative extractive abstracting of scientific texts in English", submitted for the degree of Doctor of Philosophy (PhD) in the specialty "8D06101 - Informatics, Computer Engineering and Management"

Introduction

Rapid processing of information is a vital function that is currently necessary for every modern person. The Automatic Text Abstracting (ART) process faces many challenges despite the fact that the technology in this area is constantly evolving, and this problem has been studied since 1958. For example, how to assess the quality of the abstracts received, and what should serve as a benchmark for comparison? There are two main tasks that are solved in the ART process:

1. Select critical information from a given text.
2. Presenting this information in a condensed form.

ART is a complex task in the field of natural language processing because it involves a thorough semantic and lexical analysis of the text to create an informed concise representation of the original text data. A high-quality abstract should contain basic information, be accurate in relation to the facts, relevant, readable and not redundant. Research in this area began in 1958 and numerous new works and methods have appeared every year since 2003, when large amounts of data for this purpose and the necessary computing equipment became available, reviving interest in this research topic.

From the very beginning of the study of the problems of abstracting the text, many different methods were developed; For details, see Chapter 3.2. The methods vary in the number of documents to which they apply; thus, there is single-document and multi-document autoreferencing defined two classes of methods for generalizing text:

1. **Extractive abstract**, which includes steps to extract specific sentences from the source text, without any changes.
2. **Abstract abstract** - involves a coherent and concise presentation of the source text in free form.

If we compare these two methods, the second type is more similar to human thinking, since there is a need to replace words with synonyms and rearrange them in places. In contrast, the extractive method is to compose a resume from the source text, finding the most important sentences. Thus, extractive resumes are easier to obtain and are expected to produce better results than abstract summaries.

From this it follows that the second class is more complicated, since it involves such complex techniques as natural language generation.

Currently, research around the world has refocused on abstract auto-abstraction. Nevertheless, extractive automatic abstracting is still in trend, as can be seen from the scientific works of the last two years. In addition to the complexity of forming a resume, an open question in the scientific community is its assessment. The quality metric of texts should take into account the ambiguity of natural language.

Topicality

Work towards the development of methods for automatic informative abstracting of scientific texts is now more relevant than ever as a result of the exponential growth of information in general and scientific information in particular in our world today.

The most modern models of automatic abstracting of texts at the moment are built on the basis of complex architectures of neural networks with billions of parameters, and trained on huge amounts of data and use embaddings from pre-trained language models such as BERT, GPT-3 and others. This raises questions about the possible excessive complexity of such models, their ability to generalize, because billions of neural network parameters allow models to simply "memorize" the correct answers, and ultimately questions about the economic efficiency and environmental safety of these models.

Currently, every person, and a researcher in the first place, urgently needs tools for effective work with information, one of which can be a system for automatic abstracting of texts. This topic is studied in detail in classical works.

Object of research

Automatic, informative and extractive abstracting of scientific texts in English.

Subject of research

Methode of automatic abstracting of texts.

Purpose of the study

The purpose of the dissertation work is to develop a method for informative extractive abstracting of scientific texts in English, to save time and reduce the amount of information for processing.

Objectives of the study

1. Develop an automatic referencing method based on the greedy algorithm approach.
2. Select the control parameters of the developed method.
3. To experimentally estimate the highest value of the ROUGE metric using extractive automatic text abstracting techniques.
4. Conduct a comparative analysis of the results of the developed method with existing methods.

Materials of the study

The solution of the tasks set in the work was carried out on the basis of the use of general scientific research methods in the framework of experiments with textual data and a quantitative assessment of the results obtained. The programming and source codes were written in Python 3.6 (Pandas, Numpy).

Scientific novelty

The novelty of the proposed model lies in the unique application of a greedy algorithm in the method of extractive, informative abstracting of texts.

Also, the model demonstrates performance at the level of modern abstract models, in the development of which neural networks and a huge amount of data for training were used. At the same time, the proposed model is relatively simple, and requires much less time and data for training.

The contribution of our study to scientific knowledge is as follows: 1) the identification of the upper limit for extractive summation methods (VNS, greedy algorithm, genetic algorithm) and the discovery that the VNS initialized by the greedy algorithm works even better than any of the algorithms on its own for a given task, 2) the proposal of an extractive summation method based on a greedy algorithm that works at a high level, despite its relative simplicity, 3) A cleaned dataset with different types of resumes with high ROUGE and useful text statistics.

Practical significance

We present an extractive summation approach that uses simple and old methods, but at the same time works at the level of modern models that use complex neural network architectures and huge amounts of data for training; see Fig. 1.1 for a brief description of our approach. The arXiv extract dataset of the 17,000 articles we used in our experiments is available at https://data.mendeley.com/datasets/nvsxfcbzdk/1. Some of the other advantages of the proposed approach include:[1]

- Computational simplicity.
- No training of the Machine Learning model is required, but statistical output is used.
- The summaries generated by the algorithm are rich in useful information from the text.

The developed model of automatic extractive abstracting of the text has the widest range of practical applications in science, education and business.

Science:

- Automation of literature review.
- Generation of the abstract of the article.

[1] The source code is available at GitHub at https://github.com/iskander-akhmetov/Greedy-Summarization

- Multimodal autoreferencing.
- Popularization of science.
- Updating scientific information.
- Use autrafering in other NLP tasks.
- Thematic modeling.
- Sentiment analysis.

Education:

- Automatically create notes.
- Memo.
- Mind maps.
- Generate presentation slides.
- Generation of test questions.
- Essay writing.

Business:

- Summary of large volumes of text (reports, studies, business plans).
- Formation of the minutes of the meeting.
- Query-oriented abstracting.
- Optimization of contextual advertising.

Guidelines for protection

The following are put forward for protection:

1. The method of heuristic assessment of the level of quality of the abstract by the ROUGE-1 metric, achievable with the help of extractive methods of automatic abstracting of texts, gives a result of 0.59, which is significantly higher than the current level of 0.46 for the most modern methods using neural networks.
2. The developed Method of Automatic Extractive Abstracting (AER) GreedSum, which shows the result of 0.42 on the ROUGE-1 metric on the arXiv dataset.
3. The technique of fine-tuning the control parameter of the minimum document frequency (min_df) of the GreedSum work, or the minimum frequency of occurrence of words in the sentences of the refereed text to take them into account in the creation of a dictionary for constructing a TFIDF matrix. In a sample of 376 texts from the arXiv dataset, by simple search, the optimal value of the min_df was determined as 0.042 (i.e. the word must appear in at least 4.2% of the sentences).

Approbation of the obtained results

As a result of the dissertation research, 13 scientific articles were published, including 10 articles published in foreign publications (4 journal and 6 conferences)

included in the international citation database Scopus, with percentiles from 27 to 80, one author's certificate was obtained.

Structure

The structure of the dissertation research is determined by its purpose and objectives: the dissertation consists of an introduction, a review of the literature, the main part, the conclusion and a list of references used.

In the "Main Part", in addition to data and methods, experiments are written to assess the upper limit of the quality of abstracts achievable using extractive methods of automatic abstracting of texts, as well as an autoreferencing model based on a greedy algorithm.

The total volume of the study is 147 pages, the list of references includes 144 titles.