УДК 004.822:514

INTEGRATED APPROUCH TO THE DEVELOPMENT OF TEXT SUMMARIZATION SYSTEM

Makarych M.V.
BNTU, Minsk, Belarus, 2348843@tut.by

In this information and communication technology era designing interactive computer systems that are effective, efficient, easy and enjoyable to use is becoming increasingly important. Of the numerous ways explored by researchers to enhance Human-Computer Interaction and interfaces. People have to get and operate huge volume of information for different purpose. Therefore the main task is to design text-processing systems that can help a man to collect, process and summarize the content of a great number of texts.

Automatic summarization has become an increasingly popular research topic in recent years. This area of research is a part of machine learning and data mining. The main idea of summarization is to find a subset of data that contains the "information" of the entire set. Such techniques are widely used today not only in scientific sphere but also in industrial sector. Search engines are an example of it. A lot of research include summarization of documents, image collections and videos. Document summarization tries to create a representative summary or abstract of the entire document by finding the most informative sentences while in image summarization the system finds the most representative and important images. For surveillance videos one might want to extract the important events from the uneventful context.

There are two general approaches to automatic summarization: *extraction* and *abstraction*. Extractive methods work by selecting a subset of existing words, phrases or sentences in the original text to form the summary. In contrast abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might express. Such a summary might include verbal innovations. Research to date has focused primarily on extractive methods, which are appropriate for image collection summarization and video summarization.

As regards the second principle of summarization, it is the view coming from the level of processing distinguishes. There are *surface-level* approach and *deeper-level* approach [1].

In surface-level approaches case information is represented in notions of shallow features and their combination. Shallow features include e.g. statistically salient terms, appositionally salient terms, terms from cue phrases, domain-specific or a user's query terms. Results have the form of extracts.

Deeper-level approaches may produce extracts or abstracts. The later uses synthesis involving natural language generation. They need some semantic analysis e.g. can use entity approaches and build a representation of text entities (text units) and their relationships to determine salient parts. Relationships of entities include thesaural relations, syntactic relations, meaning relations and others. They can as well use discourse approaches and model the text structure on the base of e.g. hypertext markup or rhetorical structure.

The first approaches of the automatic text summarization used only simple (surface level) indicators to decide what parts of a text include into the summary. The oldest sentence extraction algorithm was developed in 1958 by Luhn [2]. It used frequencies of terms as the sentence relevance criterion. The basic idea was that a writer would repeat certain words when writing about a given topic. The importance of terms is considered proportional to their frequency in summarized documents. The frequencies are used in the next step to score and select sentences for the extract. Other indicators of relevance used in [3] are the position of a sentence within the document and the presence of certain cue-words (i.e., words like "important" or "relevant") or words contained in the title. The combination of cue-words, title words and the position of a sentence was used in [4] to produce extracts and was demonstrated their similarity with human written abstracts.

Since that time all mentioned methods were developed into new scientific approaches to text summarization. For example, Luhn is considered as a pioneer of modern *statistical methods*. Main basic concept of these methods is that the relevance of document terms is inversely proportional to the number of documents in the corpus containing the term. Sentences can be subsequently scored for instance by summing relevance of terms in the sentence. An implementation of a more ingenious statistical method was described in [5]. It uses Bayesian classifier to compute the probability that a sentence in a source document should be included in a summary. To train the classifier the authors used a corpus of 188 pairs of full documents. The characteristic features used in Bayesian formula include except of word frequency also uppercase words, sentence length, phrase structure, in-paragraph position.

Another modern approach to text summarization is a *positional method*. It is based on text connectivity anaphoric expressions that refer to previously mentioned parts of the text need to know their antecedents in order to be understood. Extractive methods can fail to capture the relations between concepts in a text. If a sentence containing an anaphoric link is extracted without the previous context the summary can become difficult to understand. Cohesive properties comprise relations between expressions of the text. This approach is used in the works by Barzilay who has investigated lexical chains for text processing [6]. He uses the WordNet thesaurus for determining cohesive relations between terms (i.e., repetition, synonymy, antonymy, hypernymy, and holonymy) and composes the chains by related terms. Their scores are determined on the basis of the number and type of relations in the chain. Only those sentences where the strongest chains are highly concentrated are selected for the summary. A similar method where sentences are scored according to the objects they mention was presented in [7]. The objects are identified by a co-reference resolution system. Co-reference resolution is the process of determining whether two expressions in natural language refer to the same entity. The sentences where the occurrence of frequently mentioned objects overcomes the given limit are included into the summary.

We proposed the extension of the method to process a cluster of documents written about the same topic. Multi-document summarization is one step more complex task than single-document summarization. It brings into new problems we have to deal with. Our text summarization system TRT can processes the desired amount of English journalistic texts and represent the summary in the

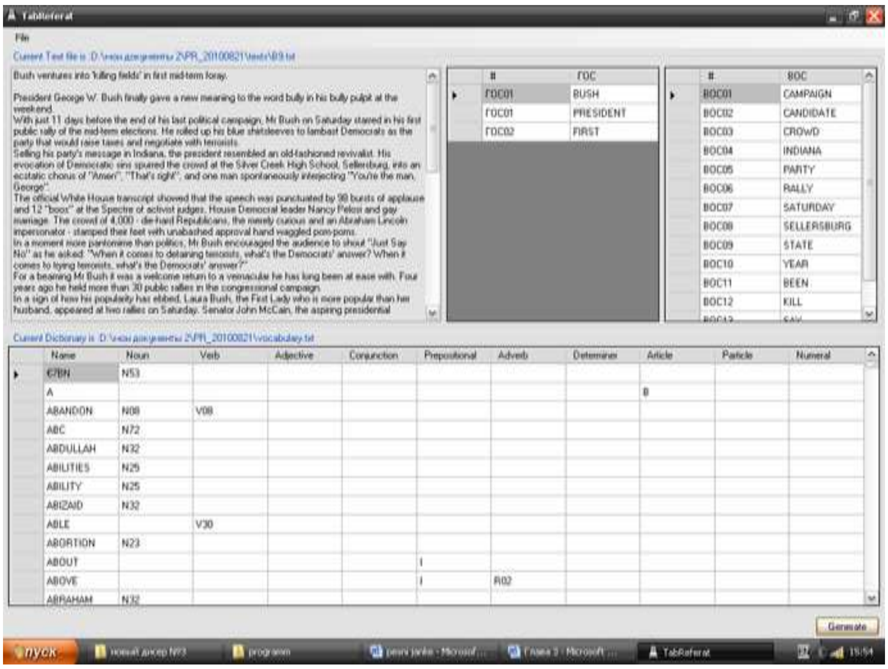

Figure 1 – View port of TRT program

form of a table. The table slots shows: Who acts? Whom does he / she act towards? In what way does he / she act? Where does he / she act? When does he / she act? The system TRT differs from the usual “total system” approach to such kind of systems. An integrated design of a linguistic database implements both principles for extracting: statistical and positional methods. Basic building blocks for a new linguistic processor are lexical-semantic, syntactical and semantic-syntactical blocks.

Lexical-semantic block is used for the semantic analysis of all the words from a processed text and contains an alphabetical dictionary with special semantic codes. These codes have been developed in accordance with a lexical-semantic classification [9].

Syntactical block performs parsing of sentences. It is based on bounda-ry signals lists for main configurations of syntax groups.

Semantic-syntactical block correlates identified syntax groups with 52 specified semantic functions in accordance with “case grammar” by C.Fillmore [8, p. 117]. This method helps to precisely define semantic functions of the keywords that are situated in parsing groups and allows the automatic system to avoid typical mistakes.

View port of TRT program is given in Fif.1. Main Menu allows to load a processed text that can be found on the left top of the TRT window. There is a score of selected terms for the extract on the right top. An alphabetical dictionary with special semantic codes is situated at the bottom of the page. Fig. 2 shows the table that summarizes given text contents.

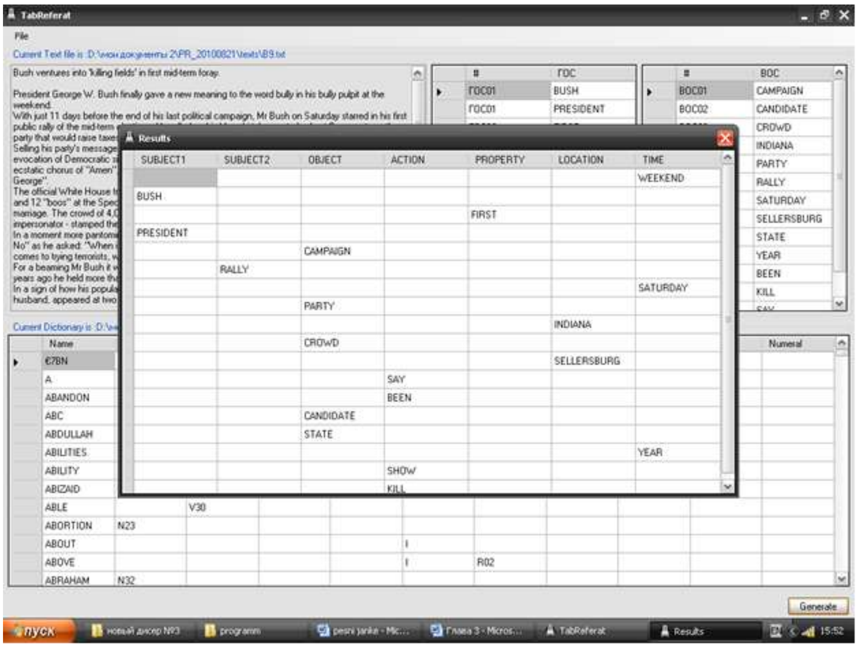

Figure 2 – View port of TRT program with table referat

This multi-document summarization system is an automatic procedure extraction of information from multiple texts written about the same topic. Resulting summary report in a form of a table allows individual users, such as professional information consumers, to familiarize themselves with information contained in a large cluster of documents. In such a way TRT system is complementing a news aggregator performing information tables that are both concise and comprehensive. Being put together and outlined every topic is described from multiple perspectives within a single document. While the goal of a brief summary is to simplify information search and cut the time by pointing to the most relevant source documents, our comprehensive multi-document table summary contains the required information, hence limiting the need for accessing original files to cases when refinement is required. Linguistic database and software of the system is opened for redesigning and can be applied in any sphere of activity.

References

1. Ježek, K. Automatic Text Summarization / K. Ježek, Steinberger, J. //Lecture Notes in Computer Science – [Electronic resource]. – Mode of access: http://textmining.zcu.cz/publications/Z08.pdf – Date of access: 5.11.2017.
2. Luhn, H. The Automatic Creation of Literature Abstracts / H. Luhn // IBM Journal of Research Development, 1958. – V 2(2). – P. 159–165.
3. Baxendale, P. Man-made Index for Technical Literature - an experiment / P.Baxendale // IBM Journal of Research Development, 1958. – V 2(4). – P. 354–361.
4. Edmundson, H. New Methods in Automatic Extracting / H. Edmundson // Journal of the Association for Computing Machinery, 1969. – V16(2). – P. 264–285.
5. Kupiec, J., A Trainable Document Summarizer / J.Kupiec, J.Pedersen, F. Chen // Research and Development in Information Retrieval, 1995. – P. 68 – 73.
6. Barzilay, R., Using Lexical Chains for Text Summarization / R. Barzilay, M. Elhadad // Proceedings of the ACL/EACL’97 Workshop on Intelligent Scalable Text Summarization, 1997. – Madrid, Spain. – P. 10 – 17.
7. Boguraev, B. Salience-based content characterization of text documents / B.Boguraev, C. Kennedy, I. Mani and M. Maybury // Advances in Automatic Text Summarization, 1999. – The MIT Press.
8. Fillmore, C. Frame semantics / C.Fillmore // Linguistics in the Morning Calm. Seoul: Hanshin Publishing Co, 1982. – P. 111–137.
9. Makarych, M. Automatic text summarization system / M.Makarych – Germany: LAP LAMBERT Academic Publishing, 2012 – 145p.