2. Соколова Т.А. Методика формирования словообразовательных стратегий при обучении чтению на французском языке студентов неязыкового вуза: дис. … канд. пед. наук : 13.00.02. – Иркутск, 2009. – 175 с.
3. Языкова Н.В. Методика формирования компенсаторных умений говорения в общеобразовательной школе / Н.В. Языкова, М.Р. Коренева // Иностранные языки в школе. – 2013. – № 9. – С.26-32.
4. Cook V. Second language learning and language teaching. 2nd Edition. – Bristol, 1996. – 262 p.
5. System and hierarchy in L2 compensatory strategies / E. Kellerman, A. Amerlaan, T. Bongaerts, N. Poulisse // Developing Communicative Competence in a Second Language. – New-York: Newbury, 1990. – Pp. xi–xii; 163-178.
6. Tarone E. Communication strategies, foreigner talk and repair in interlanguage, Language Learning. New-York: Newbury House, 1980. – Pp. 30–31.
7. Yule G. Eliciting the performance of strategic competence / G. Yule, E. Tarone // Developing Communicative Competence in a Second Language. – New York: Newbury House, 1990. – Pp. 179–193.

M. Makarych
Belarussian National Technical University

MODERN APPROACH IN NATURAL LANGUAGE PROCESSING SYSTEMS FOR SUMMARIZATION

The increasing availability of online information has necessitated intensive research in the area of automatic text summarization within the Natural Language Processing (NLP) community. Over the past half a century the problem has been investigated by applied linguistics and addressed from many different perspectives in varying domains and using various paradigms.

The subfield of summarization has been investigated by the NLP community for nearly the last half century. Dragomir R. Radev defines a summary as “a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that” [5, p.921]. This simple definition captures three important aspects that characterize research on automatic summarization:

– summaries may be produced from a single document or multiple documents;

– summaries should preserve important information;

– summaries should be short.

In fact many approaches differ on the manner of their problem formulations. But there are four main steps in summarization process [2, p. 30].

– *Extraction* is the procedure of identifying important sections of the text and producing them verbatim.

– *Abstraction* aims to produce important material in a new way.

– *Fusion* combines extracted parts coherently.

– *Compression* aims to throw out unimportant sections of the text.

Earliest instances of research on summarizing scientific documents proposed paradigms for extracting salient sentences from text using features like word and phrase frequency – *statistical method* (Luhn, 1958); position in the text – *positional method* (Baxendale, 1958) and key phrases – *lingvo-semantic method* (Edmundson, 1969). Various works published since then has concentrated on other domains, mostly on newswire data. Many approaches addressed the problem by building systems depending of the type of the required summary. While extractive summarization is mainly concerned with what the summary content should be, usually relying solely on extraction of sentences, abstractive summarization puts strong emphasis on the form, aiming to produce a grammatical summary which usually requires advanced language generation techniques. In a paradigm more tuned to information retrieval (IR), one can also consider topic-driven summarization which assumes that the summary content depends on the preference of the user and a final summary is focused on a particular topic.

There are broadly two types of extractive summarization tasks depending on what the summarization program focuses on. The first is *generic summarization*, which focuses on obtaining a generic summary or abstract of the collection (whether documents, or sets of images, or videos, news stories etc.). The second is *query-based summarization* which summarizes objects specific to a query. Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs.

An example of a summarization problem is document summarization, which attempts to automatically produce an abstract from a given document. Sometimes one might be interested in generating a summary from a single source document, while others can use multiple source documents (for example, a group of articles on the same topic). This problem is called multi-document summarization.

At a very high level summarization algorithms try to find subsets of objects which cover information of the entire set. This is also called the *core-set*. These algorithms model notions like diversity, coverage, information and representativeness of the summary. Query based summarization techniques, additionally model for relevance of the summary with the query.

Any NLP system should be implemented at a "functional building block" level. It requires a system with the following features [1, p.73]:

1. The system should have a well-defined set of primitive operators which can be combined as needed to perform the processes. The data model recognized by the operators should encompass operators themselves as a data type. This permits the system to be used to create new operators by defining them in terms of already existing ones.

2. The building blocks of a system should be made available in a processing environment which shields the application developer and user from the real computer system. It must support private data storage while providing access to a public data base and other external processes and devices. The command language must permit conversational invocation of operators in any desired sequence using names designated by the user.

3. Documents must be represented in the system so that their logical attributes are fully described.

The integrated text processing methodology for summarization was developed in Minsk state linguistic university in accordance with all based summarization techniques. A related application is summarizing automatically news articles on a given topic. The system pulls together a set of English publicistic articles and concisely represents them as a summary in a form of a table. The system called *Table R* differs from the usual "total system" approach to such kind of systems. An integrated design of a linguistic database implements three existing principles for extracting: *statistical, positional* and *lingvo-semantic methods.* Basic building blocks for a new linguistic processor are *lexical-semantic*, *syntactical* and *semantic-syntactical blocks* [4, p. 68].

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.

Syntactical block performs parsing of sentences. It is based on boundary 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 [3, 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.

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 *Table R* 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. Aone C., Okurowski, M., Gorlinsky, J., and Larsen, B. A trainable summarizer with knowledge acquired from robust NLP techniques // Advances in Automatic Text Summarization. MIT Press pages, 1999. – P. 71–80.
2. Goldfarb C., Mosher E., Peterson T. Integration of the Text Processing Functions in an Interactive Environment // Proc. Fourth Hawaii
3. International Conference on System Sciences. Honolulu: University of Hawaii, 1971. – P. 29–33.
4. Fillmore C. Frame semantics // Linguistics in the Morning Calm. Seoul: Hanshin Publishing Co, 1982. – P. 111–137.
5. Makarych M. Automatic text summarization system. – Germany: LAP LAMBERT Academic Publishing, 2012 – 145c.
6. Radev, D. R., Jing, H., Stys, M., and Tam, D. Centroid-based summarization of multiple documents // Information Processing and Management 40, 2004. – P. 919–938.