INTERACTIVE SUMMARIZATION WITH THE HELP OF MACHINE TRANSLATION

Kholmatova Feruza Dilshodovna
Master student of Samarkand State Institute of Foreign Languages

---

ABSTRACT

The article describes the methodology for developing an automated system for creating abstracts[1] in a form that meets the requirements and facilitates its subsequent translation into a foreign language. The technique combines empirical and rational methods of natural language processing and is illustrated by the example of an interactive system for generating abstracts of scientific articles on mathematical modeling. The abstract/abstract is generated in Uzbek and is accompanied by the issuance of the English equivalents of the used vocabulary.

Key words: automated abstracting, multilingual lexicon, automatic translation.

INTRODUCTION

Abstracting scientific and technical literature is an important type of professional communication, the purpose of which is the rapid exchange of information between specialists both within the same country and internationally [1]. The ever-increasing need for the publication of scientific and technical literature to be abstracted in various languages, and the increasing cost of manual processing of documents and their translation, turn the automation of abstracting into a social and economic necessity. Research on the automation of abstracting by both domestic and foreign linguists is being carried out in various areas and, despite a half-century history [6], is far from complete. In this paper, we propose a new method for developing an automated abstracting system, as well as its specific implementation on the example of a system for generating abstracts of scientific articles on mathematical modeling in Uzbek with the issuance of English equivalents of the used vocabulary.

Justification of the methodology The natural language is so vast and ambiguous that the creation of high-quality computer systems for its processing requires an infinite amount of time and effort from developers. At present, the achievement of the most correct results in automatic text processing is possible only within the rigid framework of the sublanguage due to the limitations of its vocabulary and grammar. The results of the analysis of the characteristics of the sublanguage make it possible to develop requirements for the final product of the system, help to identify the optimal ways of representing knowledge for each sublanguage, and also allow us to simplify or bypass many problems of automatic text processing that are not solvable for the entire language as a whole. Our methodology is based on sublanguage orientation. In order to save efforts and time of developers, the proposed method provides for the reuse of individual program components previously developed for other languages and applications, with their subsequent adaptation to new tasks and inclusion in the system being developed at certain stages of processing language material.

Much of the research on computer summarization concentrates on the development of fully automatic systems. However, in this process it is quite often impossible to do without the participation of a specialist, for at least two reasons. Firstly, it is necessary to introduce semantic knowledge (the content of the abstract) into the system, on the basis of which the text of the abstract should be synthesized, and secondly, it is the specialists who have the knowledge of what content should be reflected in the abstract. Introducing this knowledge into the system is not a trivial task. Therefore, one of the main characteristics of our methodology is the interactive interaction of the user with the computer. When developing linguistic support, which includes lexicographic and algorithmic components, we followed the lexicalist approach, in which the main part of linguistic knowledge is included in the lexicon, which increases the reliability of the system. The described approach was successfully tested in the

development of an interactive system for generating abstracts and annotations of scientific articles on mathematical modeling, which is described in the next section.

Description and implementation of the methodology on the example of the abstract system In this section, using the abstract system as an example, a computer implementation of the above approach to summarizing scientific and technical information is given. The abstract system is designed for the subject area of mathematical modeling. The sublanguage of abstracts on mathematical modeling reflects the requirements of State Standart [7] to the structure of the abstract as such and the specifics of the subject area ofmathematical modeling. According to the requirements of the State Standard, the text of the abstract should clearly state the main provisions of the article, avoiding complex language structures and observing the unity of terminology. This requirement is explained by the fact that incorrect or complex language design of the abstract, even with correctly selected content, can lead to misunderstanding of the abstract and errors in its translation into a foreign language. Long subordinate clauses inserted into the main clause, participial and adverbial constructions, etc. reinforce the ambiguity inherent in natural language, adding syntactic homonymy to lexical homonymy. Therefore, the system draws up the abstract of the article in the form of sentences with a simple syntactic structure and terminology used in the original article. The specificity of the subject area of mathematical modeling is reflected in the content of the linguistic knowledge base of the system, which is built on the basis of the analysis of the Uzbek corpus of articles on mathematical modeling published in the SUSU Bulletin in 2008–2012. and English-language articles on similar topics found on the Internet. The main part of linguistic knowledge is represented in the lexicon of the system. The lexicographic component of the abstract system contains a Uzbek-English lexicon with information necessary for a) formal knowledge fixation, b) algorithms for analyzing and synthesizing texts of abstracts in Uzbek, c) algorithms for translating single- and multi-component vocabulary. The algorithmic component contains: a) algorithms for accessing the lexicon, b) algorithms for analyzing scientific and technical documentation, providing for the translation of textual information into a formal language of meanings, c) algorithms for synthesizing texts of abstracts in Uzbek and d) algorithms for translating single and multicomponent vocabulary into English language. For interaction with the user, an interactive knowledge extraction module has been developed.

The abstract system reuses, as separate blocks, some software modules previously developed for the English language. These modules have been adapted for processing the Uzbek language in accordance with the objectives of the described application. In general, the abstract system includes the following components: domain-specific knowledge base that includes lexicographic and algorithmic components domain-specific analyzer Uzbek texts, consisting automatic modules highlighting nominal (NP) and verbal [10] terminology in the text of the article. The output of this module is the text of the article in an interactive format, marked up for nominal terminology and predicates; o an interactive module for syntactic analysis of the generated abstract, which represents the content of the abstract selected by the author in the form of formal knowledge representation structures; o automatic module of morphological analysis; automatic generator of essay sentences in Uzbek.

The abstract text is generated based on the conceptual scheme of the abstract sentences from the system knowledge base and on the basis of information interactively retrieved from the user. Working with the software tool is as follows. The input of the system is the text of the article. the abstract automatically analyzes the received text and presents it in a marked-up form, focusing the author's attention on the terms used (noun phrases and verbs). At the same time, the system automatically generates the first sentence of the abstract and issues an interactive list of verbs used in the article, reduced to a form that allows using these verbs as predicates of the

remaining sentences of the abstract. After the author selects a specific verb, the appropriate template for the future abstract sentence is issued. The relevant template slots are filled in by the author by automatically transferring phrases from the text parsed by the system into the template slots. Based on the completed template, a grammatically correct sentence and a list of Uzbek-English equivalents of the single and multi-component vocabulary used are generated. At the same time, the English equivalents of Uzbek predicate sentences are given in the form corresponding to the text form (time, number and gender) of Uzbek predicates, and the rest of the words and phrases (up to six words long) are given in the main singular form of the nominative case.

CONCLUSION

The article describes a methodology for developing computer systems for summarizing and annotating, based on their orientation to specific areas of science and technology, as well as to a pre-standardized text. The possibility of reusing software resources for new languages and applications and, thus, reducing all types of costs when extrapolating the system to new scientific and technical areas is shown.

REFERENCES

[1] Kapterev, A.I. Informatization of socio-cultural space / A.I. Kapterev. - M.: FAIR-PRESS, 2004. - 512

[2] Trevgoda, S.A. Methods and algorithms for automatic text summarization based on the analysis of functional relations: Ph.D. dis. ... cand. tech. Sciences / S.A. Anxiety. - St. Petersburg, 2009. - 18 p.

[3] Yatsko, V.A. Symmetrical referencing: theoretical foundations and methodology / V.A. Yatsko // NTI. Ser. 2. - 2002. - No. 5. - P. 18–28.

[4] Loret, E. A Gradual Combination of Features for Building Automatic Summarization Systems Text / E. Lloret, M. Palomar // Speech and Dialogue. – Heidelberg, 2009. – P. 16–23.

[5] Luhn, H.P. The Automatic Creation of Literature Abstracts / H.P. Luhn // IBM Journal of Research and Development. - 1958. - V. 2, No. 2. - P. 159–165.

[6] Saggion, H. A classification algorithm for predicting the structure of summaries / H. Saggion // Proceedings of the 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLPЬ 2009. – Suntec, 2009. – P. 31–38.

[7] State Standart 79–95. System of standards on information, librarianship and publishing. Abstract and abstract. General requirements. - Input. 1997–07–01 - M.: Publishing House of Standards, 1995. - 8 p.

[8] Sheremetyeva, S. On Extracting Multiword NP Terminology for MT / S. Sheremetyeva // Proceedings of the Thirteen Conference of the European Associa- tion of Machine Translation (EAMT-2009). – Barcelona, Spain. May 14–15, 2009

[9] SHERZODOVICH, A. S., & KIZI, R. Z. D. (2020). Interpretation and Written Translation: Related Learning. *INTERPRETATION*, *6*(6).

[10] Аслонова, Ш. И. (2020). ПРОБЛЕМЫ ПЕДАГОГИЧЕСКИХ ТЕХНОЛОГИЙ В ОБУЧЕНИИ МОЛОДЁЖИ В ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЯХ. *Интернаука*, (21-1), 59-60.

[11] Sherzodovich, A. S. (2020). The role of online teaching and innovative methods. *Science and education*, *1*(3), 524-528.

[12] Aslonov, S., & Ruzimurodova, Z. (2020). INGLIZ TILINI O ‘QITISHNING INNOVATSION USULLARI. *Студенческий вестник*, (12-5), 72-74.

[13] Shahram, A., Umida, K., & Zarina, R. (2020). Information technology's role in the study of foreign languages. *Asian Journal of Multidimensional Research (AJMR)*, *9*(3), 96-98.’

[14] Аслонов, Ш. Ш. (2020). КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ФИЛОЛОГИЯ: ПРОБЛЕМЫ И РЕШЕНИЯ. *Гуманитарный трактат*, (84), 17-19.