UDC 81'33(082)
DOI https://doi.org/10.24919/2663-6042.16.2021.2

APPROACHES TO AUTOMATIC SUMMARIZATION AND ANNOTATION

Golub T. P., Kovalenko O. O., Nazarenko O. I., Zhygzhytova L. M.
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

The article is devoted to the study of different approaches of digital texts automatic summarization and annotation. Due to a rapid increase in the volume of textual information on the Internet during the last years and the appearance of big databases of information, research in the field of computational linguistics devoted to the development of methods and approaches in automatic summarization and annotation have become and are still relevant and really vital. The development of algorithms, methods and approaches, and the creation of systems for automatic summarization and annotation of texts remain difficult and up-to-date tasks.

The paper considers such widespread approaches to automatic summarization of digital texts as single and multi-document summarization. Among the task of multi-document summarization are: creation of monolingual abstracts from sources in different languages; construction of abstracts from hybrid sources; creation of abstracts based on arrays of documents.

Also, knowledge-rich, knowledge-poor and hybrid approaches are studied. Thus, the knowledge-poor approach is based on the principle of creating an abstract based on the most frequently used words from the text, and combining sentences from the texts where these words are used, and not generating new sentences for the summary. So, the process of summary generation with the help of this approach consists of just two stages: analysis and synthesis. The knowledge-based approach is semantic summarization and it lets the system create a summary of texts by generalizing concepts, detecting main topics, and composing new sentences. In this case, the process of summarization goes through three stages: analysis, transformation and synthesis. And the hybrid approach is the combination of the best features of the knowledge-poor and knowledge-based approaches.

Besides the approaches, some problems in automatic summarization which are actual nowadays are named. They are: ensuring the completeness of the presentation of information, reduction of repetitions when presenting information, ensuring the coherence and clarity of the information provided.

Key words: *automatic summarization, automatic annotation, applied linguistics, knowledge-based approach, knowledge-poor approach.*

Голуб Т. П., Коваленко О. О., Назаренко О. І., Жигжитова Л. М. Підходи до автоматичного реферування та анотування. *Статтю присвячено дослідженню підходів до автоматичного реферування та анотування цифрових текстів. У зв'язку зі стрімким зростанням обсягів текстової інформації в Інтернеті протягом останніх років та появою великих за обсягом інформаційних баз даних, дослідження в галузі комп'ютерної лінгвістики, присвячені розробці методів і підходів до автоматичного реферування та анотації, стали досить актуальними. Складним залишається завдання розробки алгоритмів, методів і підходів, створення систем автоматичного реферування та анотування текстів.*

У роботі розглянуто поширені підходи до автоматичного реферування цифрових текстів, реферування на базі одного чи багатьох документів. З'ясовано, що серед завдань реферування на основі декількох документів є: створення одномовних анотацій з джерел різними мовами; створення анотацій з різнорідних джерел; створення анотацій на основі масивів документів.

Вивчено підходи, базовані на знаннях, підходи, не базовані на знаннях, а також гібридні підходи. Простежено, що основою підходу, який не базується на знаннях, є принцип створення анотації з використанням найбільш часто вживаних слів із тексту та об'єднання речень із текстів, де ці слова вживаються, а не створення нових речень для анотації. Отже, процес формування підсумків за допомогою цього підходу складається всього з двох етапів: аналізу та синтезу. Зі свого боку, підхід, заснований на знаннях, – це семантичне узагальнення, яке уможливлює системі створювати анотації до текстів за рахунок узагальнення понять, виявлення основних тем і створення нових речень. У цьому випадку процес узагальнення проходить три етапи: аналіз, перетворення та синтез. Гібридний підхід поєднує в собі найкращі аспекти обох із зазначених підходів.

Крім підходів, у статті також окреслено деякі проблеми автоматичного реферування, які є актуальними на сьогодні. Серед них – забезпечення повноти викладу інформації, зменшення повторів при поданні інформації, забезпечення зв'язності та чіткості поданої інформації.

Ключові слова: *автоматичне реферування, автоматичне анотування, прикладна лінгвістика, підхід на основі знань, підхід не на основі знань.*

Defining the problem and argumentation of the topicality of the consideration. Nowadays, the volumes of information are constantly increasing, which forces specialists in various fields of knowledge to use a variety of methods for its search, processing, presentation and transmission within their professional community. The social role of science grows, the volume of information constantly increases, a large number of publications on a variety of problems of science and technology appear in the world, and specialists do not have enough

time to follow the latest books and articles in their field of knowledge, therefore, the main content of the latest publications must be transmitted in a compressed form – in the form of annotations and abstracts, which include only the main essence, basic meaning-bearing words, phrases and sentences typical for a particular field of knowledge.

Analysis of recent research and publications. Researches in the field of creating automatic summarization and annotation systems have a big history. For more than half a century, there has been active search for effective methods of automatic summarization and annotation. Although the developments in this area started in the 1950s by the work of H. P. Luhn [7], the development of the Internet, of textual databases and the international evaluation efforts like the Document Understanding Conferences (DUC) [9], the Text Summarization Challenge by M. Okumura, T. Fukusima, H. Nanba and T. Hirao [8] have fuelled research in this field.

The analysis of recent research and publications showed that nowadays the researches devoted to automatic summarization and annotation are focused mainly on the following aspects:

– abstractive and extractive summarization (based on summary results) (N. Bhatia, A. Jaiswal) [2];

– supervised / unsupervised summarization (K. Lanyo, A. Wausi [6], M. Xiangke, Y. Hui, H. Shaobin, L. Ye, L. Rongsheng [16]);

– topic-based / query-based summarization (Y. Wei, Y. Zhizhuo [15]);

– single and multi-document summarization (M. S. Bewoor, S. H. Patil [1]);

– multilingual and cross-lingual summarization (N. Jhaveri, M. Gupta, V. Varma [5], E. Zosa, M. Granroth-Wilding, L. Pivovarova [18]);

– domain summarization (H. Hayashi, P. Budania, P. Wang, C. Ackerson, R. Neervannan, G. Neubig [4], A. M. Pujar [11], L. Scanlon, Sh. Zhang, X. Zhang, M. Sanderson [13]);

– real-time summarization (V. Paramanantham, S. S. Kumar [10]).

Setting the goals and tasks of the article. The **aim** of the article is to make an overview of the approaches in automatic summarization and annotation of text.

The outline of the main research material. Since a huge amount of the knowledge that has been accumulated by humanity nowadays is presented mostly in the form of digital texts, automatic summarization and annotation have gained significant relevance in connection with the development of the Internet and databases of information resources. To shorten search time and to improve search efficiency, users are offered catalogues of annotations and abstracts of sources on the topics they are interested in. Manual formation of them requires really enormous time and human resources, and therefore the task of creating methods for automatic summarization and annotation of texts arose.

First of all, let us define the difference between an abstract and an annotation. According to Cambridge dictionary [19], an abstract is a short form of a speech, article, book, etc., giving only the most important facts or ideas. So, we can conclude, that an abstract is a report on a specific topic, including an overview of relevant literature and other sources, or a presentation of the content of scientific work, books, etc. An annotation is a short explanation or note added to a text or image [20]. So, an annotation is a brief description of a print text or some text in a digital form. Usually, an annotation is given after a bibliographic description of a source. An annotation differs from an abstract by a significantly smaller volume and an obligatory statement of the purpose of an annotated work. As for the basic requirements for the abstract, they are compression and depiction of all the main ideas of the source text. Researchers usually name three types of abstracts: narrative, informational, critical (reviews).

The manual abstract formation (made by a human) includes the following stages: source analysis, highlighting the most important and informative fragments in the source, formation of conclusions (see Fig. 1).

Manual abstract formation is usually done for just one source while automatic abstract formation is based on different principles, approaches and methods, and gives much more possibilities. Besides, automatic summarization may be done not just for one, but for several or even for a huge amount of different sources. So, among the most frequent tasks of multisource automatic summarization are the following:

1. Creation of monolingual abstracts from sources in different languages.

2. Construction of abstracts from hybrid sources, including both text and numerical data in different forms (tables, diagrams, graphs, etc.).

3. Creation of abstracts based on arrays of documents. For example, the construction of a single abstract on the collection of abstracts of scientific conference reports. One of the areas of application of such tools is the formation of news reports from newspaper sources.

Now let us study the approaches of automatic summarization. So, some authors also divide automatic summarization and annotation methods into surface-level and deep learning methods [17]. Surface-level methods are based on text "extraction". The deep-learning methods are based on the use of thesauri and developed mechanisms for parsing text. Luis Gonçalves [3] divides the approaches to automatic summarization into extraction and abstraction ones. For extraction, summarization sentences are taken directly from the document by the principle of identifying important sections of the text and combining them into a short summary. In abstractive summarization, the summary is produced while interpreting the text. Therefore, parts of the summary

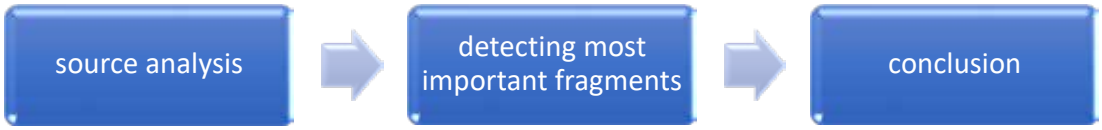

Fig. 1. A process of an abstract formation by a human

in this case are not sentences from the source text. So, abstractive methods are depth methods.

Udo Hahn and Inderjeet Mani [14] name another three categories of approaches to automatic summarization: knowledge-rich, knowledge-poor and hybrid ones. The idea of knowledge-rich or some authors call it knowledge-based approaches is that when you make out the meaning of the text, you can shorten it more efficiently and get a better summary. Knowledge-poor approaches are based on the principle of not adding new rules to each new application domain or language. Hybrid ones have the features of both of them.

Let us study these approaches to the theory of automatic summarization more precisely:

1. The knowledge-poor approach does not imply reliance on knowledge related to the text. Systems of this type use a universal rule base that does not depend on the software and the language of the text.

2. The knowledge-based approach provides for the allocation of different levels of understanding of the text, which requires along with the universal rules the rules of a knowledge base of the software and a base of linguistic rules that depend on the language.

3. The third approach is hybrid. It combines the best of the first two.

In the knowledge-poor approach (Fig. 2), the method of extracts constitution is used. It is implemented in two stages:

1. Analysis. The text and phrasal patterns are compared, and as result, the blocks of the highest lexical and statistical relevance are allocated.

2. Synthesis. The final document is formed by connecting the selected fragments.

To implement the analytical stage, a linear weighting coefficient model is used. In accordance with it, each block U of the original text is automatically assigned such weight coefficients:

• k_1, depending on the location of the U block in the original;

• k_2, depending on the frequency of the block appearance in the source text;

• k_3, depending on the frequency of using the block in key sentences;

• k_4, reflecting the indicators of the statistical significance of the block.

Then, the block importance coefficient B (U) = $\alpha_1 k_1 + \alpha_2 k_2 + \alpha_3 k_3 + \alpha_4 k_4$ is calculated from the values of k_1, k_2, k_3 and k_4 and the coefficients of setting the summarization program α_1, α_2, α_3 and α_4. According to the importance factors, the selection of blocks in the abstract is carried out.

A different rule group is used to calculate each weighting factor:

– For k_1 the location of the block is taken into account.

– For k_2, the rules take into account the results of automatic document indexing.

– For k_3, the presence in the block of such key phrases and expressions as “in conclusion …”, “according to the results of the analysis …”, “different from …”, “insignificant …” etc. is taken into account.

– For k_4, the rules take into account the occurrence of the term in headers, headers and footers, the first paragraph of the text, the user’s request profile, etc.

The main advantage of the described model of linear weights lies in the simplicity of its implementation, and the main disadvantage is associated with the possibility of forming incoherent abstracts that do not take into account the context. To eliminate it, the stage of manual editing of the results is introduced.

For a person who has grasped the general meaning of the information, it is easier to highlight the main ideas and summarize the content. This leads to the creation of referencing systems of the knowledge-based type (Fig. 3).

The knowledge-based systems require:

• powerful computing resources;

• developed grammars and dictionaries;

• advanced parsing tools;

• means of generating natural language constructions;

• ontological reference books.

These systems use three methods:

1) the traditional method of parsing;

2) a method based on the understanding of natural language;

3) combined method.

The stages of abstract synthesis in all these methods are almost the same – the usage of a text generator. The functioning of such systems requires:

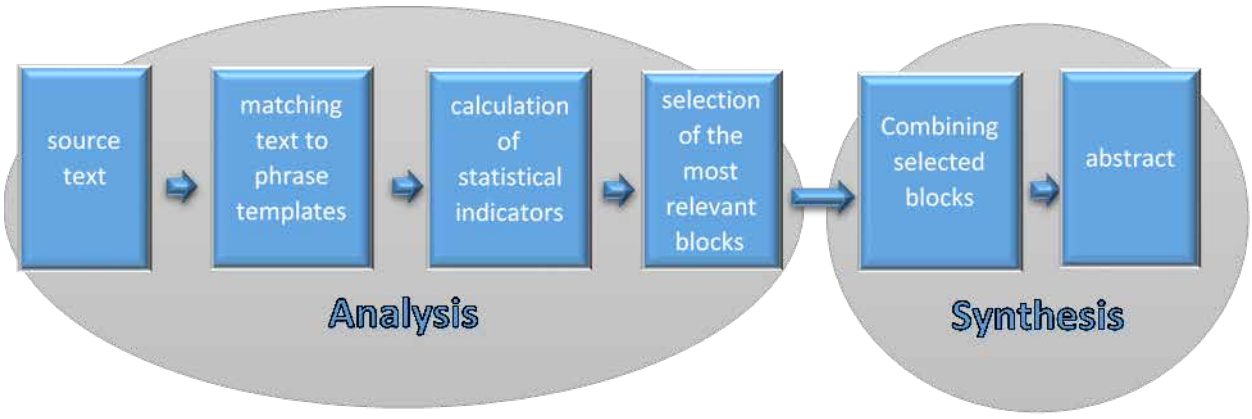

Fig. 2. The generalized architecture of automatic summarization on the knowledge-poor approach

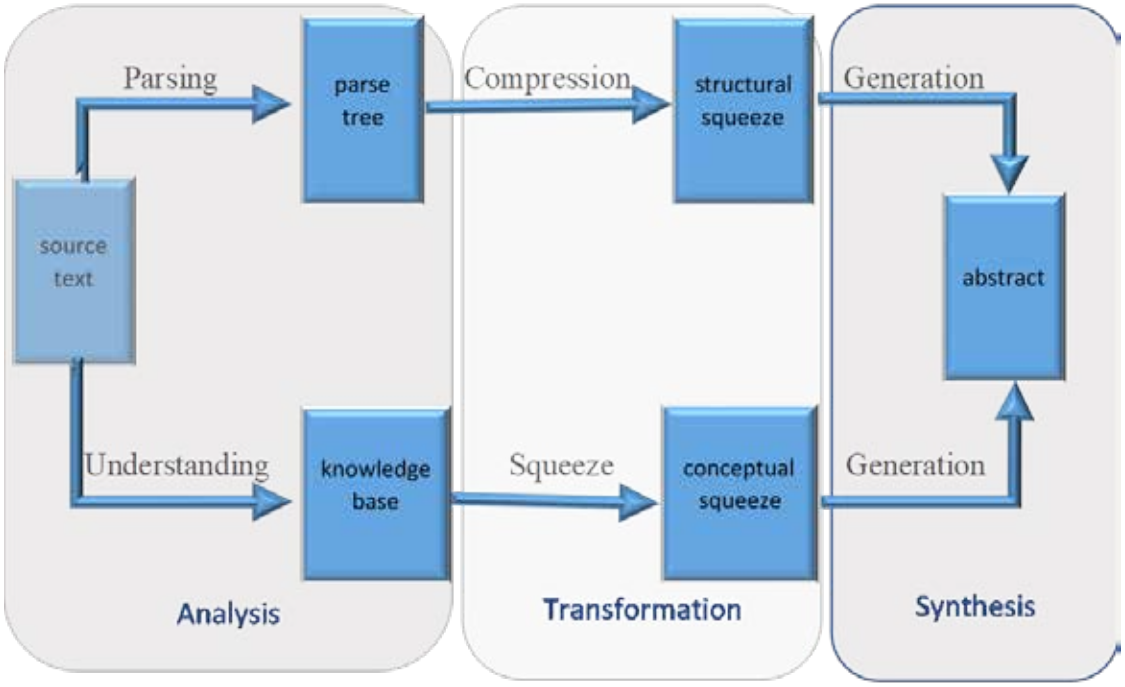

Fig. 3. The approaches to the formation of an abstract in knowledge-based systems

• comprehensive dictionaries (thesauri) of the WordNet type;

• ontological reference books such as Sousse and Penman Upper Model;

• large volumes of test files with texts (for example, The Wall Street Journal or Perm Treebank from the Linguistic Data Consortium).

As for the most significant problems automatic summarization and annotation, some researchers [12] name:

• ensuring the completeness of the presentation of information,

• reduction of repetitions when presenting information,

• ensuring the coherence and clarity of the information provided.

Conclusions and directions for further research in the area. Currently, there is a problem of information overload. Abstracts and annotations make it possible to establish the main content of the document and determine the need to refer to the original source. Automatic summarization and annotation help a person to efficiently process large amounts of information. Modern approaches to automatic summarization and annotation differ in the variety of methods and approaches used. These technologies are developing very rapidly nowadays, so further research in the area may be devoted to the new methods or technologies of automatic summarization and annotation.

BIBLIOGRAPHY

1. Bewoor M. S., Patil S. H. Empirical Analysis of Single and Multi Document Summarization using Clustering Algorithms. *Engineering, Technology & Applied Science Research*. 2018. Vol. 8. Issue 1. P. 2562–2567.

2. Bhatia N., Jaiswal A. Trends in Extractive and Abstractive Techniques in Text Summarization. *International Journal of Computer Applications*. 2015. Vol. 117. P. 21–24.

3. Gonçalves L. Automatic Text Summarization with Machine Learning – An overview. 2020. URL: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25.

4. Hayashi H., Budania P., Wang P., Ackerson C., Neervannan R., Neubig G. WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. *Transactions of the Association for Computational Linguistics*. 2021. Vol. 9. P. 211–225.

5. Jhaveri N., Gupta M., Varma V. Clstk: The Cross-Lingual Summarization Tool-Kit. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. 2019. P. 766–769. DOI: https://doi.org/10.1145/3289600.3290614.

6. Lanyo K., Wausi A. A Comparative Study of Supervised and Unsupervised Classifiers Utilizing Extractive Text Summarization Techniques to Support Automated Customer Query Question-Answering. *5th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. 2018. P. 88–92. DOI: 10.1109/ISCMI.2018.8703237.

7. Luhn H. P. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*. 1958. Vol. 2. Issue 2. P. 159–165.

8. Okumura M., Fukusima T., Nanba H., Hirao T. Text Summarization Challenge 2 Text summarization evaluation at NTCIR workshop 3. *SIGIR Forum*. 2004. Vol. 38. Issue 1. P. 29–38.

9. Over P., Dang H., Harman D. DUC in context. *Information Processing & Management*. 2007. Vol. 43. Isuue 6. P. 1506–1520.

10. Paramanantham V., Kumar S. S. A Real Time Video Summarization for YouTube Videos and Evaluation of Computational Algorithms for their Time and Storage Reduction. *International Journal on Recent and Innovation Trends in Computing and Communication*. 2018. Vol. 6. Issue 4. P. 176–186. DOI: ttps://doi.org/10.17762/ijritcc.v6i4.1540.

11. Pujar A. M. An Efficient Domain-Specific Text Summarization Using Combined Statistical & Linguistic Methods. *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*. 2021. DOI: http://dx.doi.org/10.2139/ssrn.3852820.

12. Radev D., McKeown K., Hovy E. Introduction to the Special Issue on Summarization. *Computational linguistics*. 2002. Vol. 28. Issue 4. P. 399–408.

13. Scanlon L., Zhang Sh., Zhang X., Sanderson M. Evaluation of Cross Domain Text Summarization. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 2020. P. 1853–1856. DOI: https://doi.org/10.1145/3397271.3401285.

14. Udo H., Inderjeet M. The Challenges of Automatic Summarization. *Computer*. 2000. Vol. 33. P. 29–36. URL: https://www.researchgate.net/publication/2955348_The_Challenges_of_Automatic_Summarization.

15. Wei Y., Zhizhuo Y. Query based summarization using topic background knowledge. *13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. 2017. P. 2569–2572. DOI: 10.1109/FSKD.2017.8393180.

16. Xiangke M., Hui Y., Shaobin H., Ye L., Rongsheng L. Extractive Summarization Using Supervised and Unsupervised Learning. *Expert Systems with Applications*. 2019. Vol. 133. P. 173–181. DOI: https://doi.org/10.1016/j.eswa.2019.05.011.

17. Ye A. A Guide to Text Annotation – the Key to Understanding Language Named Entity Recognition, Sentiment Analysis, and More. 2020. URL: https://towardsdatascience.com/a-guide-to-text-annotation-the-key-to-understanding-language-e221a69ee90e.

18. Zosa E., Granroth-Wilding M., Pivovarova L. A Comparison of Unsupervised Methods for Ad hoc Cross-Lingual Document Retrieval. *Proceedings of the LREC 2020 Workshop on Cross-Language Search and Summarization of Text and Speech*. 2020. P. 32–37.

DICTIONARIES

19. Cambridge dictionary. URL: https://dictionary.cambridge.org/dictionary/english/abstract. Accessed 19 November 2021.

20. Cambridge dictionary. URL: https://dictionary.cambridge.org/dictionary/english/annotation. Accessed 19 November 2021.

REFERENCES

1. Bewoor, M. S., Patil, S. H. (2018). Empirical Analysis of Single and Multi Document Summarization using Clustering Algorithms. *Engineering, Technology & Applied Science Research*, 8 (1), 2562–2567.

2. Bhatia, N., & Jaiswal, A. (2015). Trends in Extractive and Abstractive Techniques in Text Summarization. *International Journal of Computer Applications*, 117, 21–24.

3. Gonçalves, L. (2020). Automatic Text Summarization with Machine Learning – An overview. Retrieved from: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25.

4. Hayashi, H., Budania, P., Wang, P., Ackerson, C., Neervannan, R., Neubig, G. (2021). WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. *Transactions of the Association for Computational Linguistics*, 9, 211–225.
DOI: https://doi.org/10.1162/tacl_a_00362.

5. Jhaveri, N., Gupta, M., Varma, V. (2019). Clstk: The Cross-Lingual Summarization Tool-Kit. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*, 766–769. DOI: https://doi.org/10.1145/3289600.3290614.

6. Lanyo, K., Wausi, A. (2018). A Comparative Study of Supervised and Unsupervised Classifiers Utilizing Extractive Text Summarization Techniques to Support Automated Customer Query Question-Answering. *5th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 88–92. DOI: 10.1109/ISCMI.2018.8703237.

7. Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2 (2), 159–165.

8. Okumura, M., Fukusima, T., Nanba, H., Hirao, T. (2004). Text Summarization Challenge 2 Text summarization evaluation at NTCIR workshop 3. *SIGIR Forum*, 38 (1), 29–38.

9. Over, P., Dang, H., Harman, D. (2007). DUC in context. *Information Processing & Management*, 43(6), 1506–1520.

10. Paramanantham, V., Kumar, S. S. (2018). A Real Time Video Summarization for YouTube Videos and Evaluation of Computational Algorithms for their Time and Storage Reduction. *International Journal on Recent and Innovation Trends in Computing and Communication*, 6 (4), 176–186. Retrieved from: https://doi.org/10.17762/ijritcc.v6i4.1540.

11. Pujar, A. M. (2021). An Efficient Domain-Specific Text Summarization Using Combined Statistical & Linguistic Methods. Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021). Retrieved from: http://dx.doi.org/10.2139/ssrn.3852820.

12. Radev, D., McKeown, K., Hovy, E. (2002). Introduction to the Special Issue on Summarization. *Computational linguistics*, 28 (4), 399–408.

13. Scanlon, L., Zhang, Sh., Zhang, X., Sanderson, M. (2020). Evaluation of Cross Domain Text Summarization. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, 1853–1856. DOI: https://doi.org/10.1145/3397271.3401285.

14. Udo, H., & Inderjeet, M. (2000). The Challenges of Automatic Summarization. *Computer*, 33, 29–36. Retrieved from: https://www.researchgate.net/publication/2955348_The_Challenges_of_Automatic_Summarization.

15. Wei, Y., Zhizhuo, Y. (2017). Query based summarization using topic background knowledge. *13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2569–2572. DOI: 10.1109/FSKD.2017.8393180.

16. Xiangke, M., Hui, Y., Shaobin, H., Ye, L., Rongsheng, L. (2019). Extractive Summarization Using Supervised and Unsupervised Learning. *Expert Systems with Applications*, 133, 173–181. Retrieved from: https://doi.org/10.1016/j.eswa.2019.05.011.

17. Ye, A. (2020). A Guide to Text Annotation – the Key to Understanding Language Named Entity Recognition, Sentiment Analysis, and More. Retrieved from: https://towardsdatascience.com/a-guide-to-text-annotation-the-key-to-understanding-language-e221a69ee90e.

18. Zosa, E., Granroth-Wilding, M. & Pivovarova, L. (2020). A Comparison of Unsupervised Methods for Ad hoc Cross-Lingual Document Retrieval. *Proceedings of the LREC 2020 Workshop on Cross-Language Search and Summarization of Text and Speech*, 32–37.

DICTIONARIES

19. Cambridge dictionary. URL: https://dictionary.cambridge.org/dictionary/english/abstract. Accessed 19 November 2021.

20. Cambridge dictionary. URL: https://dictionary.cambridge.org/dictionary/english/annotation. Accessed 19 November 2021.