Multi-purpose search to determine the context of a text message based on the dictionary data structure

Andrii Yarovyi
Department for Computer Science
Vinnytsia national technical university
Vinnytsia, Ukraine
a.yarovyy@vntu.edu.ua

Dmytro Kudriavtsev
Department for Computer Science
Vinnytsia national technical university
Vinnytsia, Ukraine
2kn15b.kudryavtsev@gmail.com

***Abstract*— This article issues related to determining the context of textual information were analyzed. The issue of the effectiveness of the analysis of text messages is identified and resolved. Solutions to reduce the time spent on determining the context of the information were presented. Testing of the proposed solutions was performed and the results of the application of the proposed ideas were compared.**

Keywords—text-processing, intelligent information system, semantic text analysis, terminological knowledge base.

I. INTRODUCTION

This research aims to determine the maximum set of keywords that identify the text message in the adequate time and use the maximum amount of available information.

Previous research had identified the impact of the number of subject areas for identifying a text message by keywords and using semantic text analysis [1-3]. As the number of subject areas increases, the processing time of a text message increases linearly. In view of this, the total processing time of a text message is defined as the largest among all subject areas, namely their terminological knowledge bases. The structure of the terminological base for any subject area, presented in the previous research [2, 3].

In turn, the dependence on the number of terminological knowledge bases is linear and does not lead to complications in the interaction with each other. Each terminological database is isolated from the others, even in the case of duplication of data, but nevertheless, the threshold value of information correspondence must be defined as common to all terminological knowledge bases. Based on this and the results of previous studies, the optimal threshold value of information compliance should be within 15-20 percent [3, 4]. The analysis of the text messages also revealed cases when the input information has been distorted in the form of spelling errors and abbreviations, which are often used to reduce the time of writing the message by the user [5, 6]. But comparing the results of text analysis with errors or reductions and without them, the accuracy of recognition was reduced by no more than 3-5 percent, as noted in the previous research. [3]. When analysing the impact of the number of subject areas, as well as the forms of their presentation, it was found that the presence of dependencies between different terminological databases, negatively affects the analysis time of the message context and significantly slows down the overall text processing time [2, 4]. From which it was concluded that it is necessary to isolate all terminological databases. Only one repository should be used to access them and creating a dictionary with shared data from two or more subject areas will require synchronization to avoid duplication of dictionaries [3]. The choice of the dictionary as the basic unit of the terminological database was made because of the speed of information retrieval and the possibility of the automated grouping of dictionaries among themselves. In view of this, as the result of the research the text processing time was reduced up to 20 percent from the base. The datasets used in this research were taken from the Kaggle data storage resource [7-9]. Due to the decrease in recognition accuracy, there was a problem checking the quality of the terminological knowledge base and its similarity to others. To reduce the processing time of a text message, it is also advisable to use a data structure in the form of a dictionary, consisting of a key and a value that will be formed as a set of keywords from the selected subject area [10]. As you know, plain text consists of sentences that have a semantic connection and are interconnected in content [2, 11]. Accordingly, the definition of the context of information in a text message can be determined based on even a partial analysis, and then only detailing the result.

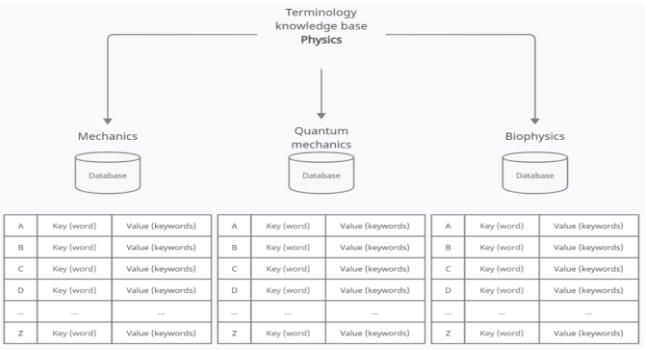


Fig. 1. Structure of terminology knowledge database.

TABLE I. TEXT ANALYSIS ACCURACY

Entry points for analyze	1	2	3
The accuracy of recognition the context of text (up to 100 words)	81.5%	89.5%	91.3%
The accuracy of recognition the context of text (100-250 words)	84.4%	84.9%	87.2%
The accuracy of recognition the context of text (250-400 words)	85.1%	88.3%	90.4%

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

The accuracy of recognition the context of text (400+ words)	87.8%	91.6%	92.1%
Tests	10000	10000	10000

II. MAIN RESEARCH

The process of analysing textual information requires semantic analysis to identify keywords and a semantic kernel, which presented as a key to get a value-answer [12]. The main disadvantage of this method of text analysis is a sharp change in the subject of the text or a weak semantic connection between lexical constructions [13]. To eliminate this shortcoming, you should use a search from different parts of the text. If several results coincide, namely the correspondence of their keys in the dictionary of the terminological database of the subject area, it is advisable to draw a conclusion on the definition of the subject of the text [14, 15]. If the key values do not match for each of the searches in each terminology knowledge base, the context should be defined as the sum of all values for the keys of the vocabulary of the subject area terminology databases.

For demonstrate the ability of using several entry points to search the context in the text, four test scenarios were created with up to four entry points. The positions for each point are selected as beginning, middle and the end of the text message on the same distance between each other which provide the best effectiveness of the search.

To confirm this, a series of experiments were performed to determine performance improvements over standard word processing methods and to determine the context of textual information.

TABLE II. TEXT ANALYSIS WITH DIFFERENT ENTRY POINTS OF SEARCH

Entry points for analyze	**1**	**2**	**3**	**4**
Average time of context recognition (up to 100 words), sec	0.842	0.672	0.593	0.574
Average time of context recognition (100-250 words), sec	3.026	2.303	**1.227**	0.983
Average time of context recognition (250-400 words), sec	6.329	3.510	**2.682**	2.221
Average time of context recognition (400+ words), sec	7.328	4.205	**3.574**	3.297
Tests, thousand	100	100	100	100

From the table, the best ratio is getting when using three entry points and the average length of a text message from 100 words and more.

The selection of intellectually valuable elements from the information stream is formed by searching among the information of structures already identified by the repository as constructs that belonging to the terminology database of the chosen subject area or regions [16-18]. The described search is carried out by applying the intellectual means of information analysis.

Technic tools and instruments which were used during the research are –

Software:

- .NET Framework 5.0.
- Microsoft Visual Studio 2019.

Hardware:

- GPU NVIDIA 2070 RTX
- CPU Intel Core i7-8700
- RAM 16 Gb 3000 MHz

Also, the structure of stored data for each subject area must be the same for fast search with a minimum of the getting data logic changes. Formatting data needs to be simple and clearly understandable. The most preferable structures are the hash table and key-value dictionary. Implementing part is related to end-user development [19]. Pay attention to the size of each table in the terminology base. In the experiment section will be presented several examples of terminology knowledge bases for different subject areas.

After all, when using multiple data stores at the same time to select the best result, each repository will provide its own best answer with the corresponding recognition factor [16, 20]. In this case, the data storage response that is not related to the subject of the session should be rejected as inappropriate. It is for this purpose that this threshold is introduced, which aims to reject such answers and to avoid such situations in general. Also using several subject areas with their terminology knowledge bases can have unions of keywords which will create an issue of understanding the context of information. In this case, will be better to create a combination of dictionaries which will store the parts of terminology knowledge databases and their unions with each other to resolve this issue. Using this data structure prevents performance issues because the time spent on getting the item from the dictionary data structure is equal to O (1) [21]. The result of adding the dictionaries is presented on the Table 3.

TABLE III. COMPARISON TABLE OF USING DICTIONARIES OF TERMINOLOGICAL KNOWLEDGE BASES

Entry points for analyze	**1**	**2**	**3**	**4**
Average time of context recognition (up to 100 words) without dictionaries, sec	0.836	0.628	0.593	0.526
Average time of context recognition (up to 100 words) with dictionaries, sec	0.734	0.569	0.427	0.434
Average time of context recognition (100-250 words) without dictionaries, sec	3.148	2.321	1.583	0.837
Average time of context recognition (100-250 words) with dictionaries, sec	2.415	1.793	1.204	1.186

The accuracy of recognition the context of text (up to 100 words) with dictionaries	81.2%	82.4%	82.5%	82.5%
The accuracy of recognition the context of text (up to 100 words) without dictionaries	79.7%	80.8%	81.1%	81.9%
Average count of keywords set for text (up to 100 words)	35	33	32	30
Average count of keywords set for text (100-250 words)	104	87	85	85
The accuracy of recognition the context of text (100-250 words) with dictionaries	92.7%	94.1%	94.7%	95.2%
The accuracy of recognition the context of text (100-250 words) without dictionaries	87.3%	91.4%	92.1%	93.7%
Tests, thousand	10	10	10	10

The above comparisons testify to the effectiveness of the use of combined data sets transformed from terminological knowledge bases into sets of dictionaries. Moreover, the effect is magnified using prior distribution on dictionaries of terminological knowledge bases as the number of terminological knowledge bases increases. The results of using up to four terminology databases are presented on the Figures 2-4.

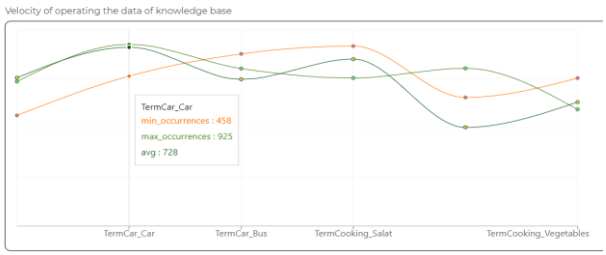


Fig. 2. Using two terminology knowledge databases.



Fig. 3. Using three terminology knowledge databases.

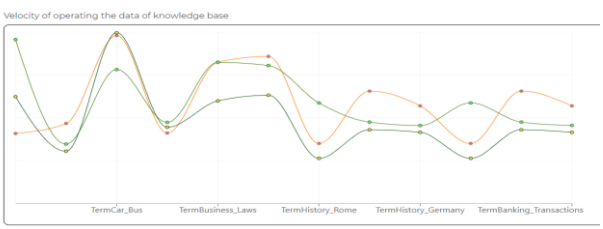


Fig. 4. Using four terminology knowledge databases.

According to the results of analysis, it is worth noting that most of the time is spent on retrieving data from the database and on forming a value-response to the semantic kernel of a text message, accounting for a total of 60% to 90% of the total processing time. In this case, reducing the time to retrieve information from a database is not the task of this research and requires further analysis of the use of the database. But the formation of the value-response must be analysed for the detection of processing defects. To do this, we need to determine all the dependencies between the input data and module of the intelligent text analysis.

As a result, the best combination for text analysis is using three entry points for text and up to four terminology knowledge bases which can contain up to 48 dictionaries with all keywords for adequate level of text recognition.

In our case, percent on Fig.5-7 are using for showing data capacity relation which taken from data sources to find the context of the dialog in parallel sessions between terminology databases.

By using several subject areas with their own terminology knowledge databases, the data distribution in the dictionaries must be structured by size of different parts of subject areas. The results of data distribution between dictionaries are presented in Figures 5-7:

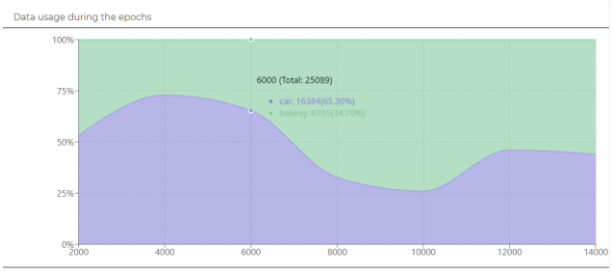


Fig. 5. Distribution between two terminology knowledge databases (up to 24 dictionaries).

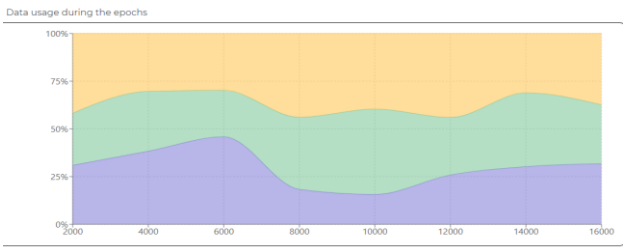


Fig. 6. Distribution between three terminology knowledge databases (up to 36 dictionaries).

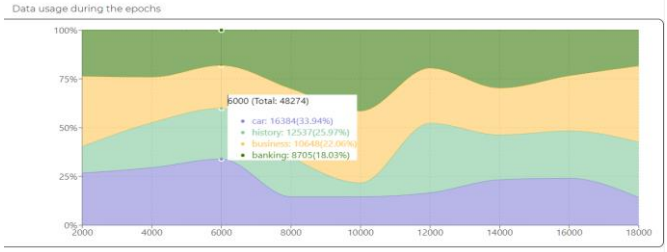


Fig. 7. Distribution between four terminology knowledge databases (up to 48 dictionaries).

III. Conclusion

According to these results, the advantages of using several entry points are proved. Performance issue about the parallel search the context for text parts is resolved by using modern parallel CPU architecture which can operate up to 16 threads simultaneously.

TABLE IV. Comparison of using different methods of text analysis

Title	Average processing time, s	Found keywords, N	Correctness, %	Tests, N
Semantic text analysis	0,602	381064	74,22	150000
Word2vec [10]	0,449	389082	83,43	150000
Multi-purpose text analysis (two subject areas)	0,376	393176	**85,37**	150000
Multi-purpose text analysis (three subject areas)	0,481	395362	**86,52**	150000
Multi-purpose text analysis (four subject areas)	0,619	395163	**89,03**	150000

Text analysis for multiple subject areas can be used for checking the topic of texts, find keywords for marketing purposes, for checking the emails or websites content to find necessary information. The main advantages of using the multi-purposed text analysis are finding the keywords which can be different by subject areas, the ability to scan user preference in the marketing sphere, making an analysis of the sense of the entered text, and provide a better semantic text analysis especially in the big social networks like Facebook, Twitter, etc. In experiments section was shown that increasing the subject areas count creates an issue relating to the performance of the search, and the best solution is to use up to three entry points for search the context. Using the fourth entry point is not decreased the time in the same proportion as for performance using of CPU.

During the testing of dictionary using, the impact of sublevels of terminology knowledge databases were noted and after comparison between non-using dictionaries like a monolith structure and distributed data structure was found that time spent for analyse the text is better with distributed data structure on 15%-18%.

References

[1] Andrii Yarovyi, Dmytro Kudriavtsev, Serhii Baraban, Volodymyr Ozeranskyi, Liudmyla Krylyk, Andrzej Smolarz, and Gayni Karnakova "Information technology in creating intelligent chatbots", Proc. SPIE 11176, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019, 1117627 (6 November 2019); https://doi.org/10.1117/12.2537415.

[2] A. Yarovii, D. Kudriavtsev and O. Prozor, "Improving the Accuracy of Text Message Recognition with an Intelligent Chatbot Information System," 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT), Zbarazh, Ukraine, 2020, pp. 76-79, doi: 10.1109/CSIT49958.2020.9322036.

[3] Yarovyi, A., Kudriavtsev, D. Method of multi-purpose text analysis based on a combination of knowledge bases for intelligent chatbot (2021) CEUR Workshop Proceedings, 2870, pp. 1238-1248.

[4] Werder K., Heckmann C. S. Ambidexterity in Information Systems Research: Overview of Conceptualizations, Antecedents, and Outcomes, Journal of Information Technology Theory and Application. – 2019. – Т. 20. – №. 1.

[5] Alarcon, Rodrigo & Moreno, Lourdes & Martinez, Paloma. (2020). Word-Sense disambiguation system for text readability. 10.1145/3439231.3439257.

[6] Kanagarathinam, Karthick & Ravindrakumar, K & Francis, R & Ilankannan, S. (2019). Steps Involved in Text Recognition and Recent Research in OCR; A Study. 8. 3095-3100.

[7] Bhat G., "Chatbot Data" dataset, https://www.kaggle.com/fungusamongus/chatbot-data, 2018.

[8] Liling Tan, "Old Newspapers" dataset, https://www.kaggle.com/alvations/old-newspapers, 2018.

[9] Jeet J., "US Financial News Articles" dataset, https://www.kaggle.com/jeet2016/us-financial-news-articles, 2018.

[10] Word2vec algorithm, @vishwasbhanawat, medium.com, [Web source]: https://medium.com/@vishwasbhanawat/the-architecture-of-word2vec-78659ceb6638

[11] Research Gate article, "The multiple dimensions of information quality", [Web source]: https://www. researchgate. net/publication/ 242929284 The Multiple Dimensions of Information Quality

[12] M. Tamer Ozsu and Patrick Valduriez, "Distributed and Parallel Database Systems", [Web source]: https://www.researchgate.net/publication/220566423_Distributed_and_Parallel_Database_Systems

[13] BeyondTheBlocks, "Data storage complexity", [Web source]: https://beyondtheblocks.reduxio.com/enterprise-data-storage-needs-complexity-2017

[14] Chatbots magazine, "Types of chatbots how to determine the right one for your needs", [Web source]: https://chatbotsmagazine.com/the-3-types-of-chatbots-how-to-determine-the-right-one-for-your-needs-a4df8c69ec4c

[15] Yu D. et al. Towards Accurate Scene Text Recognition with Semantic Reasoning Networks //arXiv preprint arXiv:2003.12294. – 2020.

[16] Comparebusinessproducts, "The biggest data storages", [Web source]: https://www.comparebusinessproducts.com/fyi/10-largest-databases-in-the-world

[17] Guthrie D. Unsupervised detection of anomalous text – University of Sheffield, – 2008.

[18] "Intelligent data analysis and data mining", Universitat Politècnica de Catalunya [Web source]: http://www.lsi.upc.edu/~avellido/teaching/1314/Intro4_IDADM_DAKD_09-10-13.pdf

[19] Tsmots, I.,Skorokhoda, O., Tesliuk, T., Rabyk, V. Designing Features of Hardware and Software Tools for Intelligent Processing of Intensive Data Streams. Processing of the 2016 IEEE First International Conference on Data Streams and Processing, DSMP 2016. Lviv. Pp. 332-335.

[20] Pearlson K. E., Saunders C. S., Galletta D. F. Managing and using information systems: A strategic approach. – John Wiley & Sons, 2019.

[21] Serban I. V. et al. A deep reinforcement learning algorithm //arXiv preprint arXiv:1709.02349. – 2017.