UDC 811.111'33
DOI https://doi.org/10.24919/2308-4863/91-1-27

Liudmyla VLASIUK,
orcid.org/0000-0003-1020-0076
Senior Lecturer, PhD Student at the Department of Theory, Practice and Translation of English
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
(Kyiv, Ukraine) l.vlasiuk@kpi.ua

Olga DEMYDENKO,
orcid.org/0000-0002-0643-5510
PhD,
Associate Professor at the Department of Theory, Practice and Translation of English
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
(Kyiv, Ukraine) olga.demydenko80@gmail.com

AI-POWERED TEXT ANALYSIS SYSTEMS

The article explores the development, principles, and applications of AI-powered text analysis systems as a response to the exponential growth of unstructured information in the modern digital environment. It outlines the historical trajectory from manual information retrieval towards highly automated intelligent systems, emphasizing how artificial intelligence has transformed text processing into a structural, semantic, and functional discipline within linguistics and computational science. The study employs a comprehensive methodological framework, integrating analysis and synthesis, taxonomy, induction and deduction, comparative and contextual methods, corpus approaches, distributive and component analysis, as well as frame and compositional analysis. Drawing on English and Ukrainian media texts, the research highlights the acute challenges of information overload, where unstructured data accounts for nearly 90% of available online content, making effective retrieval and selection of relevant information increasingly difficult.

The discussion reviews key AI-driven tools such as OBSERVER, OntoSeek, and TextAnalyst, demonstrating their ability to build semantic networks, thematic hierarchies, conceptual graphs, and perform automated clustering, indexing, and semantic search. Particular attention is given to the integration of metadata-based technologies like the Open Archives Initiative (OAI-PMH), which enable unified access to distributed text repositories. The article systematically describes the multi-level process of AI text analysis, including graphematic segmentation, morphological parsing, syntactic structuring, and semantic modeling. While these technologies represent a significant advance in enabling intelligent information retrieval, they continue to face major limitations in accuracy, semantic understanding, and contextual relevance due to the complexity of natural language.

The findings indicate that current AI-powered systems remain constrained in their ability to achieve deep syntactic-semantic comprehension, often producing results with insufficient correlation to user queries. This underscores the necessity for developing new approaches that better capture structural and semantic features of natural language, as well as improved databases capable of handling vast unstructured text arrays. The article concludes that future progress in AI-powered text analysis will depend on advancing semantic algorithms, refining linguistic modeling, and addressing the critical problem of structuring the digital information space.

Key words: *language learning, research methodology, English language, compound word, compound noun, information reliability, software development field.*

Людмила ВЛАСЮК,
orcid.org/0000-0003-1020-0076
старший викладач, аспірантка кафедри теорії, практики та перекладу англійської мови
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»
(Київ, Україна) l.vlasiuk@kpi.ua

Ольга ДЕМИДЕНКО,
orcid.org/0000-0002-0643-5510
кандидат педагогічних наук,
доцент кафедри теорії, практики та перекладу англійської мови
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»
(Київ, Україна) olga.demydenko80@gmail.com

СИСТЕМИ АНАЛІЗУ ТЕКСТУ НА БАЗІ ШТУЧНОГО ІНТЕЛЕКТУ

У статті досліджується розвиток, принципи та застосування систем аналізу тексту на базі штучного інтелекту у відповідь на експоненціальне зростання неструктурованої інформації в сучасному цифровому середовищі. Вона окреслює історичну траєкторію від ручного пошуку інформації до високоавтоматизованих інтелектуальних систем, підкреслюючи, як штучний інтелект перетворив обробку тексту на структурну, *семантичну та функціональну дисципліну в рамках лінгвістики та обчислювальної науки. У дослідженні використовується комплексна методологічна база, що інтегрує аналіз і синтез, таксономію, індукцію та дедукцію, порівняльні та контекстуальні методи, корпусні підходи, дистрибутивний та компонентний аналіз, а також фреймовий та композиційний аналіз. Спираючись на англійські та українські медіатексти, дослідження підкреслює гострі проблеми інформаційного перевантаження, де неструктуровані дані становлять майже 90% доступного онлайн-контенту, що робить ефективний пошук і відбір релевантної інформації дедалі складнішим.*

У обговоренні розглядаються ключові інструменти на базі штучного інтелекту, такі як OBSERVER, OntoSeek та TextAnalyst, демонструючи їхню здатність будувати семантичні мережі, тематичні ієрархії, концептуальні графи, а також виконувати автоматизовану кластеризацію, індексацію та семантичний пошук. Особлива увага приділяється інтеграції технологій на основі метаданих, таких як Ініціатива відкритих архівів (OAI-PMH), які забезпечують єдиний доступ до розподілених текстових сховищ. У статті систематично описується багаторівневий процес аналізу тексту за допомогою штучного інтелекту, включаючи графематичну сегментацію, морфологічний розбір, синтаксичне структурування та семантичне моделювання. Хоча ці технології є значним прогресом у забезпеченні інтелектуального пошуку інформації, вони продовжують стикатися з серйозними обмеженнями в точності, семантичному розумінні та контекстуальній релевантності через складність природної мови.

Результати дослідження показують, що сучасні системи на базі штучного інтелекту залишаються обмеженими у своїй здатності досягати глибокого синтаксико-семантичного розуміння, часто даючи результати з недостатньою кореляцією із запитами користувачів. Це підкреслює необхідність розробки нових підходів, які краще фіксують структурні та семантичні особливості природної мови, а також удосконалених баз даних, здатних обробляти величезні неструктуровані текстові масиви. У статті робиться висновок, що майбутній прогрес у аналізі тексту за допомогою штучного інтелекту залежатиме від розвитку семантичних алгоритмів, удосконалення лінгвістичного моделювання та вирішення критичної проблеми структурування цифрового інформаційного простору.

Ключові слова: *вивчення мови, методологія дослідження, англійська мова, складне слово, композит, надійність інформації, галузь розробки програмного забезпечення.*

Introduction. One of the key features of the development of linguistics in the 21st century is the emergence of large volumes of documents, publications and other information sources that require sorting and unification. It was during this period that the first information retrieval systems were developed. At the first stages, such a search was carried out only manually, but the rapid development of the computer industry and, accordingly, the automation of subsequent processes significantly contributed to the digitization of the format of text information and, as a result, the development of automatic information retrieval systems.

With the growth of the number of various information sources, the need to structure the information ecosystem has become more acute than ever and has quickly become one of the most pressing problems of modern linguistics. Therefore, automatic text analysis and synthesis, text clustering, linguistic databases and their automation, and the improvement of information retrieval systems are among the most important areas of linguistic research.

Furthermore, this rapid growth in the number of information sources has led to uncontrolled information overload: only according to rough estimates, the share of unstructured data on the Internet is at least 90%. That is, in fact, structured data that is indexed in database management systems is only 10%. This indicator is critically low and indicates the impossibility of adequately searching for relevant information through huge data arrays of unstructured information. The problem of searching for information is increasingly being replaced by the problem of selecting the necessary information. This is due to the fact that users spend a huge amount of time searching for relevant information among the information flow. In turn, there is a need to create so-called intelligent search systems, in other words, deep text analysis technologies.

The article **aims** to explore the principles and applications of AI-powered text analysis systems, with particular attention to their structural, semantic and functional capacities, highlighting their advantages over traditional methods, addressing key challenges and outlining future prospects for their use in research and practice.

Materials and methods. Given the deepening of the current processes of destructuring the information ecosystem, the need for a step-by-step methodology for conducting an analysis of AI-powered text analysis systems is growing. Given the multifaceted and multidisciplinary nature of the issue, our study uses a comprehensive approach that allows for analysis in different linguistic sections. The material of our study is a sample of English-language and Ukrainian-language media texts. Thus, the methodology of our study integrates the following methods: analysis and synthesis, taxonomy, induction and deduction, comparative method, contextual method, corpus method, method of direct components, distributive analysis, frame analysis, compositional.

Discussion. The need to optimize search results and structure the digital information ecosystem is growing every day, because quick search for the necessary information or data is needed in every field, starting from simple daily news in the media and ending with individual areas: science, technology, medicine, business, etc. Whatever field we consider, the trend of uncontrolled growth in information volumes is observed everywhere and every minute the information market generates more and more new information. For this reason, the ability to extract the most significant and relevant fragments from a huge array of information sources in the information space has become a necessity. Of course, modern science offers various tools for extracting keywords, but taking into account the constantly growing volume of information, obtaining specific results is becoming an increasingly difficult task. Among the wide variety of tools, we can highlight the Autosummarize function in Microsoft Office, automatic systems IBM Intelligent, Text Miner, Oracle Content and Inxight Summarizer. However, working with these tools demonstrates a high level of limitations in their functionality. Basically, the entire functionality of such tools comes down to highlighting and selecting original fragments of the source text/document and combining them into text that is smaller in size.

Analytical processing of text information through automated systems also includes means of implementing electronic libraries. The development of a single unified interface to provide users with access to a set of autonomous sources is data integration in electronic systems (Sukhyi, 2005). The basis of integration systems is the Open Archives Initiative technology. Information resources of such systems are mainly collections of text documents that are independently formed in the nodes of the global network, and their owners support and administer them.

The Open Archives Initiative technology integrates in a unified repository not the information resources themselves, but the metadata of these resources, because through metadata it is possible to describe collections of information resources of archive sources. Metadata is collected on the basis of a specially developed protocol – Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH), which enables global search.

Effective processing of text information requires careful analysis of software tools for working with arrays of text documents. Today, there are a number of programs that enable automated text analysis, in our work we consider those that are most popular and offer the widest arsenal of functions.

The OBSERVER program allows you to use existing ontologies to gain access to heterogeneous, distributed and independent data repositories. The program generates a set of ontologies of subject areas. Accordingly, when the user creates a search query using the terms of one or more ontologies, the program searches for the corresponding texts by converting the query into ontologies. It is also possible to combine several ontologies in order to obtain more accurate search results.

The OntoSeek program helps to extract information contained on online pages and product catalogs taking into account the context. The high efficiency of the program is explained by its interaction with the linguistic database WordNet, which allows the user to specify the meaning of keywords. The WordNet database consists of synsets, that is, words that are equivalent in meaning. The uniqueness of the WordNet database lies in the combination of a lexical dictionary and an ontology that helps to trace the relationship between words in the dictionary. The description of a resource in a linguistic database can be schematically represented in the form of a lexical conceptual graph, in which words are placed at the vertex, and the arcs reflect the semantic relations between words, for example, this is how subclass relations can be displayed (Lobanovska, 2011). Thus, the search for information relevant to the user's search query occurs by matching ontologies. When selecting relevant resources, the OntoSeek program matches the conceptual graph of the query with existing resource graphs. The placement of databases of such graphs occurs on a centralized program server.

Another program for automated text content analysis, semantic information search, and electronic archive development is TextAnalyst, which allows you to (Giorgi, 2010):

– automatically form semantic networks with hyperlinks, which creates a semantic portrait of the text and allows you to analyze its content, taking into account semantic connections;

– automatically form a thematic tree with hyperlinks, i.e., present the semantic structure of the text in a hierarchical form with division into topics and subtopics, and accordingly analyze the content of the text;

– perform semantic search, taking into account semantic connections between the keywords of the query and the text itself;

– form a semantic portrait of the text, i.e., summarize it automatically;

– divide texts into thematic classes, in other words, perform automatic clustering of information;

– automatically index the text and convert it into hypertext;

– rank the text taking into account the semantic meaning of the keywords and their corresponding degree of importance in the text.

The key purpose of all the programs discussed above is to perform information search in huge text arrays. In their databases, such systems contain any sources of text information, starting from small media articles and ending with entire encyclopedias, directories, archives of periodicals, specialized archives, libraries of specialized literature, etc (Kushnarenko, 2006). That is why in response to his request the user always receives a lot of links, because the system processes each link and based on the data received outputs all the relevant texts. That is, the system must find not just documents, but information in them.

Modern automated electronic text processing is able to set a number of restrictions on search combinations of words in order to increase search efficiency. In this case, a full-fledged linguistic analysis of the text, in particular its grammatical and semantic components, plays an extremely important role. Automatic extraction of information from arrays of text documents is certainly associated with artificial intelligence systems and adequate understanding of natural language text by an automated system.

Any AI-powered system performs a text analysis during three phases. The analysis starts with graphematic analysis, which divides the source text into individual words or sentences. At this stage, the system creates a sample of words from the text in the form of a table, assigning each word the ordinal number of the sentence from which it was extracted (Selivanova, 2008).

The next stage involves performing morphological analysis, during which the system first isolates the bases (i.e., those parts that do not change), then compares grammatical characteristics (parts of speech, gender, number, case, etc.) with individual words. Thanks to morphological analysis, we can determine individual characteristics of a word as a part of speech, taking into account its context (Giorgi, 2010). In fact, first we determine the initial form and categorical meaning, and only then – the morphological characteristic of the word, which includes various morphological categories, semantic-functional groupings and lexical-grammatical categories.

At the next stage, the system performs syntactic parsing, which involves searching for grammatical idioms; analyzing the sentence from the point of view of both grammar and vocabulary; determining noun and verb groups; isolating the core and elements dependent on it (Selivanova, 2008). For automated parsing of natural language text in a grammatical context, a syntactic parser is required, the main task of which is, first of all, a structured search for information. The main difficulty in this process is the interdependence of syntax and semantics, which is difficult for a syntactic parser to track, especially in the case of syntactic homonymy.

No less important stage is the semantic analysis of the text, the main result of which is the final formation of a formalized representation of the text, comparing during this process the knowledge from the simulation model and the acts that are present in the text.

The basis for semantic analysis is the results obtained during syntactic analysis, since at the output we get an exact reproduction of the syntactic structure of each sentence, represented by a set of dependency trees. However, given the insufficient effectiveness of syntactic analysis, semantic analysis solves a significant set of problems using the results of the analysis of individual words (Kushnarenko, 2006). Semantic analysis is based on a thesaurus of the language, with the help of which it is possible to identify semantic relations between words. The identification of these relations is possible due to the fact that the thesaurus builds binary relations using a set of words of natural language.

Since there is no single principle for implementing semantic structuring of digital information space, and each new program improves the previous ones, using new methods and working mainly with a specific subject area, semantic structuring algorithms are considered in the context of the systems in which they are used. Individual sentences and situations from which it is necessary to extract knowledge for further processing by the system must be submitted to the input in a certain ordered form.

Modern technologies of pre-processing of texts, which make up the digital information space, provide for multi-level representation of natural language. However, the functionality of these technologies remains insufficient to provide highly relevant and clear search results, which would have a qualitative level of correlation with the user's search query. This process is complicated by the constantly growing amount of information, which requires further improvement of search engines.

Conclusions. When working with text information from a variety of different information resources, it is necessary to define a number of tasks, which include identifying keywords and creating a conceptual text model, further integrating this model into a full-text database, performing searches in full-text databases, ensuring relevant search results, and summarizing information from multiple sources. An important step in this process is to take into account the structural and semantic features of the natural language text. A low level of preliminary structural and semantic analysis can lead to the system misinterpreting intratextual relationships.

The modern digital information space is becoming increasingly unstructured. This is primarily a consequence of the rapid growth of information data. Despite significant progress in the field of automated text analysis and information retrieval, modern technologies remain insufficient for a deep semantic and syntactic understanding of text data. This, in turn, requires the development of new approaches that will take into account the peculiarities of working with unstructured text arrays, the creation of appropriate databases, as well as the structural and semantic features of texts.

BIBLIOGRAPHY

1. Giorgi Alessandra. About the Speaker: Towards a Syntax of Indexicality. New York: Oxford University Press, 2010. 229 p.
2. Кушнаренко Г.М. Наукова обробка документів. Київ: Знання, 2006. 336 с.
3. Лобановська І.Г. Індексація документів ключовими словами. Київ: Нілан-ЛТД, 2011. 32 с.
4. Селіванова О.О. Сучасна лінгвістика: напрями та проблеми. Полтава: Довкілля-К, 2008. 711 с.
5. Сухий О.Л., Міленін В.М., Тарадайник В.М. Алгоритми пошуку в інформаційних системах. Київ, 2005. 70 с.

REFERENCES

1. Giorgi Alessandra. (2010) About the Speaker: Towards a Syntax of Indexicality. New York: Oxford University Press, 229 p.
2. Kushnarenko N.M. (2006) Naukova obrobka dokumentiv [Scientific processing of documents]. Kyiv: Znannia. 334 s. [in Ukrainian]
3. Lobanovska I.H. (2011) Indeksuvannia dokumentiv kliuchovymy slovamy [Indexing of documents by keywords]. Kyiv: Nilan-LTD. 32 s. [in Ukrainian]
4. Selivanova O.O. (2008) Suchasna linhvistyka: napriamy ta problemy [Modern linguistics: directions and problems]. Poltava: Dovkillia-K. 711 s. [in Ukrainian]
5. Sukhyi O.L., Milenin V.M., Taradainik V.M. (2005) Alhorytmy poshuku v informatsiinykh systemakh [Search algorithms in information systems]. Kyiv. 70 s. [in Ukrainian].

Дата першого надходження рукопису до видання: 23.09.2025
Дата прийнятого до друку рукопису після рецензування: 29.10.2025
Дата публікації: 24.11.2025