UDC 811.111’33
DOI https://doi.org/10.32782/folium/2025.6.6

OPTIMAL ALGORITHM OF LINGUISTIC INDEXATION

Liudmyla Vlasiuk
Postgraduate Student, Lecturer,
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»
ORCID ID 0000-0003-1020-0076
l.vlasiuk@kpi.ua

Olga Demydenko
Candidate of Pedagogical Sciences, Associate Professor,
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»
ORCID ID 0000-0002-0643-5510
olga.demydenko80@gmail.com

Key words: *linguistic indexation, semantics and syntax, comparative analysis, language system, applied linguistics and corpus, sentence structure, mediatext.*

One of the core features of the linguistics development under the period of XXI century is the emergence of large volumes of documents, publications and other information sources which need to be sorted and unified. It was during this period that the first information retrieval systems were developed. At the first stages, such search was carried out exclusively manually, however, the rapid development of the computer industry and, accordingly, the subsequent processes automation significantly contributed to digitizing the text information format and, consequently, developing the automatic information retrieval systems.

The article presents a comprehensive overview of the linguistic indexation phenomena, including the challenges posed by the unstructured textual data in the digital era. Highlighting the need for the improvement of information search process, the article deals with the low efficiency of existing information analysis systems, mainly caused by uncontrolled information overload. To pursue the key objective of this article, the authors provide a detailed outline of the existing systems for automatic language analysis to identify their main features and gaps to be further addressed.

While developing the methodology for optimal linguistic indexation algorithm, the authors analyze and integrate all levels of language analysis: morphological, syntactic, and semantic analysis.

The authors create a structured multi-step approach to enhance the quality of automated text analysis, encompassing grammatical parsing, morphological tagging, syntactic-semantic dependency analysis, and semantic modeling. The findings suggest that such approach enhances the accuracy of information analysis and contributes to structuring the information ecosystem more effectively and accurately.

ОПТИМАЛЬНИЙ АЛГОРИТМ ЛІНГВІСТИЧНОЇ ІНДЕКСАЦІЇ

Людмила Власюк
аспірантка, викладач,
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Ольга Демиденко
кандидат педагогічних наук, доцент,
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Ключові слова: *лінгвістична індексація, семантика та синтаксис, порівняльний аналіз, мовна система, прикладна лінгвістика та корпус, структура речень, медіатекст.*

Однією з ключових особливостей розвитку лінгвістики у XXI столітті є поява великих обсягів документів, публікацій та інших джерел інформації, які потребують сортування та подальшої уніфікації. Саме на цей період припадає поява перших інформаційно-пошукових систем. На перших етапах такий пошук здійснювався виключно вручну, проте стрімкий розвиток комп'ютерної індустрії та, відповідно, автоматизація всіх процесів значно сприяли оцифруванню текстового формату інформації і, як наслідок, розвитку автоматизованих інформаційно-пошукових систем.

У статті представлено всебічний огляд явища лінгвістичної індексації, включаючи поточні проблеми, які виникають в результаті неструктурованих текстових даних в цифрову епоху. Підкреслюючи необхідність вдосконалення процесу пошуку інформації, в статті особливу увагу приділено питанню низької ефективності існуючих систем аналізу інформації, головним чином через неконтрольоване інформаційне перевантаження. Для досягнення основної мети цієї статті автори також надають детальний огляд наявних систем автоматичного аналізу мови з метою виявлення їхніх основних особливостей, функцій та потенційних недоліків, які потребують подальшого вирішення.

Розробляючи методологію оптимального алгоритму лінгвістичної індексації, автори ретельно аналізують та інтегрують усі рівні мовного аналізу: морфологічний, синтаксичний та семантичний.

Автори створюють структурований, комплексний, багатоетапний підхід з метою підвищення якості автоматизованого аналізу тексту, який включає в себе граматичний розбір, морфологічне тегування, аналіз синтаксично-семантичних залежностей та семантичне моделювання. Результати дослідження свідчать, що такий підхід підвищує точність аналізу інформації та сприяє більш ефективному та точному структуруванню інформаційної екосистеми текстів.

Introduction. With the advent of the Internet, electronic information has taken a prominent place in every sphere of modern life because it provides access to any information. The world's information repositories contain terabytes of information, a significant percentage of which is textual information. However, on the other hand, the emergence of the Internet has also led to an uncontrolled information overload. Rough estimates suggest that the share of unstructured data on the Internet is at least 90%. In other words, the actual structured data indexed in the database management systems is only 10%. This figure is critically low and indicates the impossibility of conducting an adequate search for relevant information due to the exceedingly big amount of unstructured information.

The problem of finding information is now increasingly being replaced by the problem of selecting the right information. This is due to the fact that users spend a huge amount of time searching for relevant information within the information flow. In turn, there is a need to create the so-called intelligent search systems, in other words, technologies for in-depth text analysis.

Material and methods. The aim of the article is to develop a working algorithm of linguistic indexation to foster structuring of the information ecosystem and improve information search. The methodological base of this research is determined by its objectives and aim.

To achieve the objectives set, we use a comprehensive approach to the study of the linguistic indexation phenomenon on the material of English and Ukrainian mediatexts. Thus, we integrate the following methods: analysis and synthesis, taxonomy, induction and deduction, comparative method, contextual method, corpus method, frame analysis, distributional analysis, compositional analysis, immediate constituent analysis. This approach allows for the analysis and description of key features and functions of linguistic indexation, and, consequently, development of the appropriate algorithm.

Guided by the specifics of the research, we've created a corpus of English and Ukrainian mediatexts of

Table 1

Programs for text analysis and linguistic processing

№	Name of the system	Description
1	Cognitive Dwarf	A program for text analysis and linguistic text processing. The software package includes a parser for English and an automatic translation system.
2	Core Language Engine	It uses quasi-logical forms to represent knowledge, knowledge here is represented as a set of forms that are weakly dependent on the context. The system is used to translate texts, database management, and interpretation of natural language search queries. The system allows you to limit the range of possible interpretations of the syntactic structure of a sentence by cutting off options that do not have a quasi-logical form for transformation.
3	Link Grammar Parser	Language syntactic parser. As a result of parsing a sentence, the system determines its syntactic structure, which consists of a set of marked relations connecting pairs of words.
4	Cíbola/Oleada	Cíbola/Oleada projects implement systems for linguistic analysis of texts, including tools for working with multilingual texts, performing statistical analysis and automatic translation.
5	Text Analyst	The program allows you to build a semantic network of concepts with references to the context, and carries out a meaningful search for fragments of the text, taking into account the hidden connections in the query words. The text is analyzed by building a hierarchical topic tree.
6	Langsoft	Software for processing natural language, performs grammatical sentence parsing, spelling and grammar, logical inference, audio and video translation of sentences.
7	*Quintura Searchcrystal*	Metasearch engines that present search results in graphical form. The results are clustered by static criteria. Morphological analysis is used to build a visual cluster; syntactic and semantic analysis are not implemented.
8	*Vivisimo Nigma*	Metasearch engines with clustering of search results provide the ability to highlight words that often occur with the words of the search query. The systems perform grammatical and morphological analysis.
9	*Oracle Text*	The software package allows you to work efficiently with queries related to unstructured texts, search, classify and cluster documents, extract key concepts, perform automatic annotation, and search for associative links in documents.
10	*ADVEGO*	Carries out semantic analysis of the text by calculating the ratio of unimportant words in the document to the total number of words, comparative analysis of texts using the "shingle" method.
11	*IBM Watson*	A supercomputer whose main task is to understand questions formulated in a natural language and find answers to them in a database.
12	*IBM Intelligent Miner for Tex*	It is a set of separate utilities that run independently of each other. For example, the Language Identification Tool automatically detects the language in which a document is written; the Categorization Tool automatically assigns text to a specific category; the Clusterization Tool divides a large number of documents into groups depending on the proximity of style, form or frequency characteristics; the Feature Extraction Tool detects keywords in documents based on the analysis of a predefined dictionary.

publicistic style – English and Ukrainian Mediatexts Corpus (EUMC) serving as a material basis for the research.

Results and discussion. Taking into account an ever-increasing number of various information sources, the issue of structuring the information ecosystem has substantially gained in its importance. Given the fact that this rapid development is happening in the digital era, automatic text analysis and synthesis, text clustering, linguistic databases and their automation, improvement of information retrieval systems are among the most important areas of linguistic research.

The first step in the process of structuring the information ecosystem is the analysis of the existing automatic systems for language analysis. This enables us to understand the current gaps in their functioning. Currently there is a whole range of different systems, however, none of those systems can ensure a detailed and accurate language analysis. In table 1 we provide a comparative analysis of the most widely used automatic systems for language analysis.

As it is clear from the comparative analysis above, none of the existing systems can provide an extensive and, more importantly, accurate language analysis. However, the level of analysis accuracy not only depends on the functions of a particular system, but also on the type of the language. In this respect, we take into account whether the language is high-resource or low-resource. For instance, English is a widely used high-resource language and, thus, automatic systems for language analysis have sufficient material to analyze and identify basic patterns, which is not the case with Ukrainian. Therefore, linguistic analysis of the text, in particular its grammatical and semantic components, plays an extremely important role and helps to replenish the systems with the necessary data. Automatic extraction of information from arrays of text documents is, of course, associated with artificial intelligence systems and adequate understanding of natural language text by an automated system (Steinbach, 2011: 34–37).

Automated natural language analysis goes through a number of stages, each of which we will consider in more detail. We start with a grammatical analysis, which divides the source text into separate words/sentences. At this stage, we create a sample of words from the text in the form of a table, assigning each word the sequence number of the sentence from which it was extracted. Since this is the initial stage of automated text analysis, we identify not only words, but also punctuation marks, abbreviations, conventions, etc., as all of these are part of the grammatical structure (Corazza, 2004: 21–32).

The next stage is morphological analysis, where we first identify the bases (i.e., the parts that do not change), then compare grammatical characteristics (parts of speech, gender, number, case, etc.) with individual words. Through morphological analysis, we determine the individual characteristics of a word as a part of speech, taking into account its context. In fact, firstly we determine the initial form and categorical meaning, and only then we determine the morphological characteristics of the word, which include various morphological categories, semantic-functional groupings, and lexical-grammatical categories. After completing the first two steps, it is possible to move on to the final one and determine the morphological characteristics of the word forms (Giorgi, 2010: 105). To increase the efficiency of indexation, it is advisable to perform post-morphological analysis after morphological analysis and, thus, partially eliminate grammatical homonymy, which complicates the process of automated linguistic indexation.

The next step is to perform syntactic analysis that is parsing, which requires searching for grammatical idioms; analyzing the sentence in terms of both grammar and lexis; identifying noun and verb groups; and separating the core and dependent elements. Automated parsing of natural language text in a grammatical context requires a parser whose main task is primarily to search for information in a structured way (Лобановська, 2011: 24). The main difficulty in this process is the interdependence of syntax and semantics, which is difficult for a parser to track, especially in case of syntactic homonymy. This problem can be solved by creating an explanatory combinatorial dictionary that contains information about the consistency of words in the context of syntax and semantics.

Parsing (syntactic analysis) and lexical analysis are an integral part of linguistic indexation. The key problem with linguistic indexation is information retrieval. To make such search effective, it is necessary to develop a parser that can track the interaction between the semantic and syntactic components and, accordingly, eliminate syntactic homonymy in whole or in part. In fact, when performing parsing, we compare a linear sequence of language tokens and formal grammatical forms, resulting in a parse tree (Сухий, Міленін, Тарадайнік, 2005: 60). We perform parsing simultaneously with lexical analysis. During parsing, the source text is converted into a data structure that fully reproduces the syntactic structure of the source text, which allows for further processing.

Understanding both the meaning of the sentence itself and its semantic relations with other elements requires an awareness of lexical elements and their correlation, which primarily correlates with the syntactic structure of the sentence. Hence, the main task of semantic analysis is to explain the correlation between natural language sentences and objects of the external world. Accordingly, the purpose of such analysis is to identify the semantic characteristics inherent in each word or phrase.

Semantic analysis is always based on isolating the semantic core of a sentence, which allows us to focus on the objective components of the content. Accordingly, semantic analysis makes it possible to improve information retrieval systems, as it allows us to identify keywords, organize them according to their weight in the document, and thus create a meaningful portrait of the document.

The next step is to automatically identify syntactic (semantic-syntactic) relations within the identified components. These relations can be displayed within the constructed dependency trees (Steinbach, 2011: 74). The corresponding automated process will be based on the construction of a logical and linguistic model of a natural language sentence that reproduces the syntactic structure of the sentence, taking into account also the semantic relationship that makes the meaning of the text clear.

The next important step is to build a semantic model of the text. Any text can be interpreted as a system of elements that can be formalized in terms of the properties of the text, which is always holistic, coherent, modal, dialogic, can be divided into smaller parts, and is characterized by the autosemantics of text segments. We define the semantic model of a text as an abstract model that combines the key textual properties and the interconnectedness of the structural elements of the text.

Currently, the indexing process is based on relatively traditional means of analysis by key parameters, but such template models are unable to perform a complete and detailed analysis of natural language texts, and therefore do not analyze the text in a meaningful way, taking into account the context. The need to create clear descriptions of natural language remains relevant. This would lay an important theoretical and practical foundation for further automated text analysis and synthesis using computer technology. Therefore, the aforementioned steps ensure an effective algorithm of linguistic indexation, which can create a basis for more accurate analysis results and, thus, contributes to structuring the text information ecosystem.

Conclusions. When working with textual information from a variety of different information resources, it is necessary to define a number of tasks, including identifying keywords and creating a conceptual textual model, further integrating this model into a full-text database, searching full-text databases, guaranteeing relevant search results, and summarizing information from multiple sources.

Despite the large number of programs for automatic text analysis, their functionality remains insufficient to provide highly relevant and accurate search results that would have a qualitative level of correlation with the user's search query. This process is complicated by the ever-growing amount of information, which requires further improvement of search engines. The algorithm suggested can contribute towards the structuring of information ecosystem, however, this area leaves space for further research. This includes further analysis of existing gaps that disable an accurate information search and, consequently, enrichment of the automation language analysis systems with the relevant language patterns to ensure an effective information retrieval and search via automatic means.

BIBLIOGRAPHY

1. Corazza, E. (2004). Reflecting the Mind: Indexicality and Quasi-Indexicality. Oxford : Oxford University Press.
2. Giorgi Alessandra. (2010). About the Speaker: Towards a Syntax of Indexicality. New York: Oxford University Press.
3. Steinbach, M. A. (2011). Comparison of Document Clustering Techniques. Minnesota : Minnesota Publishing.
4. Ticher, S., & Mejer, M. (2009). Methods for analyzing text and discourse. Oxford : Oxford University Press.
5. Лобановська, І.Г. (2011). Індексування документів ключовими словами. Київ: Нілан-ЛТД.
6. Сухий, О.Л., Міленін, В.М., & Тарадайнік, В.М. (2005). Алгоритми пошуку в інформаційних системах. Київ.

Отримано: 27.03.2025
Прийнято: 16.04.2025