

DEVELOPMENT TRENDS IN THE ANALYSIS OF SENTENCES IN NATURAL LANGUAGES AND THE PROBLEM OF PARAPHRASING IN THE UZBEK LANGUAGE

Khusniya Khusanovna Akhmedova

Abstract: This article analyzes the development trends of the automatic, computer-assisted analysis of sentences in natural languages, as well as the problem of paraphrasing within the framework of Natural Language Processing (NLP). It examines the morphological, syntactic, and semantic stages of linguistic analyzers and their role in the automatic processing of text. In addition, it reviews the paraphrasing systems and approaches that have emerged in international practice, together with the tasks of paraphrase extraction, identification, and generation. The agglutinative nature of the Uzbek language, its lexical synonymy, and the richness of its phraseological units are described as the principal linguistic factors that influence the paraphrasing process.

Keywords: natural language processing, NLP, computational linguistics, linguistic analyzer, morphological analyzer, syntactic analyzer, semantic analyzer, paraphrase, paraphrase identification, paraphrase generation, synonymy, semantic similarity, Uzbek language

Introduction

In recent years, the rapid development of information technology, artificial intelligence, and natural language processing has significantly increased the need for the automatic, computer-assisted analysis and processing of human language. Natural language processing systems are regarded as a set of technologies that make it possible to receive human speech in textual or audio form and to understand, interpret, and process it. These technologies are applied in numerous practical areas, such as voice assistants, machine translation, intelligent search engines, chatbots, sentiment detection, and automated information analysis.

The complexity of automatic natural language processing is connected with the multifaceted nature of language and the absence of a single universal model that fully describes it. For this reason, the use of linguistic analyzers is of great importance in text processing. Linguistic analyzers serve to formalize natural language, extract knowledge, form the semantic model of a text, generate vector representations, and support the subsequent stages of intelligent processing.

Among these tasks, paraphrasing occupies a special place. In paraphrasing, the core meaning of a sentence is preserved, while its lexical composition, syntactic structure, or certain means of semantic expression are changed. The growth in the volume of digital information and the need to detect duplicate content, check for plagiarism, improve search engines, and process texts automatically increase the relevance of the tasks of paraphrase identification and generation.

1. Systems for the Analysis of Sentences in Natural Languages

Linguistic analyzers are among the fundamental components of the architecture of intelligent text systems. According to their functional tasks, they are divided into morphological, syntactic, and semantic analyzers. These stages operate in an interconnected manner and constitute the sequential process of complex text analysis.

The morphological analyzer serves to determine the grammatical features of words. In this process, the word form is compared with dictionary units, the stem and affixes are identified, and the morphological features of the word are formed. The problem of creating morphological analyzers for the Turkic languages has been the subject of research for a long time. In Uzbek, too, there are studies devoted to the morphological analyzer and its linguistic support. In particular, the works of Sh. Hamroyeva focus on the formation of the linguistic support of the Uzbek morphological analyzer.

The syntactic analyzer determines the structure of an incoming sentence and models the syntactic relations within it. Syntactic analysis uses both top-down and bottom-up approaches. In Uzbek, syntactic analysis can be carried out on the basis of grammatical and morphological information. In this process, the morphological features of the words in a sentence serve as an important source for determining the subsequent syntactic relations.

The semantic analyzer, in turn, is oriented toward determining the meaning-related features of texts and sentences. The quality of semantic analysis depends directly on the completeness of linguistic resources, in particular dictionaries, thesauri, ontologies, and synonymic data. In the paraphrasing process, it is precisely semantic analysis that is important for determining the semantic proximity between sentences, selecting synonymous or equivalent units, and preserving the meaning of the resulting sentence.

2. The Problem of Paraphrasing in Natural Language Processing

Paraphrase is one of the complex problems in computational linguistics and natural language processing. From a practical point of view, a paraphrase can be regarded as an alternative linguistic expression that conveys the same or a similar meaning. The surface lexical composition of the sentences may differ, but their core semantic content is preserved.

The practical significance of paraphrasing manifests itself in several directions. It can be applied in question-answering systems for processing questions of various forms, as well as in information extraction, machine translation, text summarization, natural language generation, search engines, and plagiarism detection. In particular, identifying sentences that differ in lexical composition but have the same or a similar meaning is more complex than methods based on simple string or word matching.

One of the main problems in paraphrase identification is the absence of common lexemes in semantically close sentences. Conversely, sentences with almost identical words may express different meanings. Therefore, relying only on surface lexical matching is not sufficient. Vector-based and machine-learning approaches that take semantic similarity into account are regarded as important tools for solving this problem.

In international practice, various methods have been proposed for paraphrase identification, including logic-based, vector-based, surface-text-similarity-based, syntactic-similarity-based, and symbolic-meaning-representation-based methods, as well as machine-learning and decoding-based methods. With the development of deep-learning methods, models such as LSTM and convolutional neural networks have also begun to be applied to paraphrase identification tasks.

Paraphrase-related tasks can be conditionally divided into three groups: paraphrase extraction, paraphrase identification, and paraphrase generation. Paraphrase extraction aims to find sentences or phrases within a large corpus that may be paraphrases of one another. Paraphrase identification determines whether a given pair of sentences is a paraphrase or not. Paraphrase generation, in turn, is the process of producing a new, semantically equivalent sentence on the basis of a given sentence.

3. Paraphrasing Systems in International Practice

In the field of natural language processing, many paraphrasing tools with different functional capabilities have been developed. Examples include systems such as QuillBot, MyEssaywriter.ai, Ahrefs' Paraphrasing, WriteHuman, Grammarly, CleverSpinner, Paraphraser.io, Spin Rewriter, and the Prepostseo Paraphrasing Tool. Some of them operate in several languages and provide functions such as rewriting text at various levels, replacing synonyms, and changing the stylistic appearance of the text.

However, according to the analysis presented here, the capability of these systems to paraphrase Uzbek-language texts in a complete and specially adapted manner is limited or non-existent. This

situation demonstrates the necessity of developing specialized linguistic resources and automatic paraphrasing mechanisms for the Uzbek language. At the same time, analyzing the functional capabilities of existing systems makes it possible to define the core functions required for a system oriented toward the Uzbek language.

4. Linguistic Features of Paraphrasing in the Uzbek Language

The paraphrasing process in Uzbek is directly connected with the agglutinative structure of the language, its rich synonymic layer, and the wide use of phraseological units. In agglutinative languages, word forms are created by sequentially adding various grammatical affixes to the stem. As a result, the same semantic unit may appear in different morphological forms. This situation requires that morphological features be taken into account in the process of automatic analysis and paraphrasing.

In lexical paraphrasing, synonymic substitution is one of the main methods. For example, in the sentence “Ertalab badantarbiya bilan shug‘ullangan bolalar sog‘lom bo‘ladi” (“Children who exercise in the morning stay healthy”), the word “ertalab” (“in the morning”) can be replaced with units such as “tongda” or “sahar payti” (“at dawn”), and the unit “badantarbiya” (“physical training”) with equivalents such as “sport mashg‘ulotlari” (“sports training”) or “jismoniy mashqlar” (“physical exercises”), thereby producing new variants that preserve the meaning. Here it is important that the synonym fit precisely the given context.

In paraphrasing, antonymic units may also be used under certain conditions. For example, the expression “sog‘lom bo‘ladi” (“stays healthy”) may, in some contexts, be replaced with an antonymic-negative form such as “kasal bo‘lmaydi” (“does not get sick”), while preserving the overall meaning of the sentence. Thus, automatic paraphrasing is not limited to direct synonymic substitution alone, but must also take into account the context and the semantic structure of the sentence.

Phraseological units constitute another important layer of paraphrasing in Uzbek. For example, the idiom “bahri dili ochildi” can be expressed through semantically close units such as “kayfiyati ko‘tarildi”, “ko‘ngli yozildi”, “xursand bo‘ldi”, or “yuragi yayradi” (all conveying the sense of “became joyful” or “cheered up”). Some idioms may also be replaced with a free word combination or a single lexeme. This shows that, for an automatic system, it is necessary to form the synonymic and semantic equivalents of phraseological units as a separate linguistic resource.

5. A Taxonomy of Paraphrases

In systematizing paraphrases, it is expedient to use a taxonomic approach based on the lexical, syntactic, and semantic levels. Lexical paraphrases cover changes at the word level, syntactic paraphrases cover changes in sentence structure, and semantic paraphrases cover changes in the meaning and context of a sentence.

At the lexical level, the following cases are observed: simple synonymic substitution, context-dependent synonymy, metaphorical synonymy, antonymic change, change of the part of speech, and the expansion of abbreviations into their full forms. At the syntactic level, it is possible to change word order, simplify a complex sentence, or reorganize the chronological order of the parts of a sentence. At the semantic level, in turn, the addition of detail, the removal of detail, semantic generalization, more precise expression, and certain contrasting semantic changes are considered.

The use of a detailed taxonomy is important for an automatic system, as it makes it possible to describe the paraphrasing process more precisely and to introduce various changes as parameters into machine-learning algorithms. Here, along with the linguistic features of the Uzbek language, it is also necessary to take into account the context and the degree of semantic equivalence.

6. Development Trends of Paraphrasing Methods

Paraphrasing methods initially began with approaches based on the substitution of dictionary entries and synonyms; later they were enriched by statistical and machine-learning methods, and subsequently by models based on deep neural networks and transformer architectures. In the traditional synonym-substitution method, WordNet or other thesauri are used, and an attempt is made to select a synonym appropriate to the context. The study by Bannard and Callison-Burch examined an approach to selecting synonymic equivalents and identifying appropriate variants on the basis of internet statistics. [60; 100]

In neural approaches, models such as Seq2Seq and stacked residual LSTM have been applied. Although these models are effective for natural language generation, limitations related to the lexical diversity of the resulting paraphrases have also been noted. In the next stage, context-based synonym-selection methods were improved and combined with processes such as thesaurus use, tokenization, and part-of-speech (POS) tagging.

In general, the development trend shows a shift from surface lexical matching toward semantic representation, and from rule-based approaches toward data-driven models. From this perspective, in creating a paraphrasing system for the Uzbek language, harmonizing existing linguistic resources with modern machine-learning and semantic-modeling methods is a promising direction.

Discussion

The analysis carried out shows that, in order to organize paraphrasing effectively within a natural language processing system, relying on a single stage of analysis is not sufficient. Morphological analysis serves to determine word forms, syntactic analysis to reveal sentence structure, and semantic analysis to determine the semantic proximity of sentences. Their sequential and interconnected operation forms the linguistic basis of paraphrasing.

One of the main problems for the Uzbek language is connected with the insufficient development of resources. The lack of a comprehensive WordNet or a large-scale paraphrase corpus complicates the automatic identification of synonymic, phraseological, and semantic relations. Therefore, forming a linguistic database appropriate for the Uzbek language, systematizing synonymic and phraseological units, and combining them with modern semantic models is an important scientific task.

According to the results of the analysis, paraphrasing in Uzbek can be viewed in three main functional directions: extracting paraphrases from an existing text, determining whether a given pair of sentences is a paraphrase, and generating a new paraphrase variant for a given sentence. Combining these functions within a single system expands the possibilities for the semantic processing of Uzbek-language texts.

Conclusion

The development of systems for the automatic analysis of sentences in natural languages is closely connected with the consistent improvement of morphological, syntactic, and semantic analyzers. On the basis of the analyzed sources, it has been established that linguistic analyzers are an important tool for the computer-assisted processing of natural language. In particular, in the paraphrasing process, semantic analysis, synonymic relations, and the determination of contextual correspondence acquire special importance.

The modern development trend of paraphrasing shows that it is moving from rule-based synonym-substitution methods toward data-driven, machine-learning, and deep neural models. For the Uzbek language, the agglutinative nature of the language, its synonymic and phraseological richness, and the limited availability of specialized resources must be taken into account in particular.

Thus, in solving the problem of paraphrasing in the Uzbek language, it is expedient to combine, within a single methodological approach, the formation of linguistic resources, the classification of

lexical, syntactic, and semantic paraphrases, the determination of semantic similarity, and the use of modern machine-learning models. This direction is promising for Uzbek computational linguistics from both a scientific and a practical point of view.

References

1. Hamroyeva Sh. Linguistic Support of the Uzbek Morphological Analyzer [in Uzbek]. Monograph. — GlobeEdit, 2020.
2. Hamroyeva Sh. Linguistic Foundations of Compiling an Author Corpus of the Uzbek Language [in Uzbek]: PhD Dissertation in Philology. — Karshi, 2018. — 250 p.
3. Abdurakhmonov G., Sulaymonov A., Kholiyorov Kh., Omonturdiyev J. Modern Uzbek Literary Language. Syntax [in Uzbek]. — Tashkent: O'qituvchi, 1979.
4. G'ulomov A., Asqarov M. Modern Uzbek Literary Language. Syntax [in Uzbek]. — Tashkent: O'qituvchi, 1965.
5. Makhmudov N., Nurmonov A. Theoretical Grammar of the Uzbek Language [in Uzbek]. — Tashkent, 1995.
6. Hamroyeva Sh. Linguistic Support of the Uzbek Morphological Analyzer [in Uzbek]: Author's Abstract of the Doctor of Sciences (DSc) Dissertation in Philology. — Fergana, 2021. — 34 p.
7. Rakhmatullayev Sh., Makhmudov N., Kholmanova Z., O'razova I., Rikhsiyeva K. Phraseological Dictionary of the Uzbek Language [in Uzbek]. — Tashkent: Gafur Gulom Publishing and Printing Creative House, 2022. — 636 p.
8. Androutsopoulos I., Malakasiotis P. A Survey of Paraphrasing and Textual Entailment Methods.
9. Bannard C., Callison-Burch C. Paraphrasing with Bilingual Parallel Corpora. In: Proceedings of the 43rd Annual Meeting of the ACL. — 2005. — P. 597–604.
10. Bolshakov I.A., Gelbukh A. Synonymous Paraphrasing Using WordNet and Internet. In: Lecture Notes in Computer Science. — P. 312–323.
11. Prakash A., Hasan S.A., Lee K., Datla V.V., Qadir A., Liu J., Farri O. Neural Paraphrase Generation with Stacked Residual LSTM Networks. — 2016. — P. 2923–2934.
12. Barmawi A.M., Muhammad A. Paraphrasing Method Based on Contextual Synonym Substitution. Journal of ICT Research and Applications. — 2019. — Vol. 13(3). — P. 257–282.