UDC 8.81.11.81ˈ32
DOI https://doi.org/10.24919/2308-4863/74-1-29

Iryna BASARABA,
orcid.org/0000-0002-3209-9119
PhD in Philology,
Instructor at the Foreign Languages Department
National Academy of the State Border Guard Service of Ukraine
(Khmelnytskyi, Ukraine) irynaborovyk2017@ukr.net

Iryna BETS,
Candidate of Pedagogical Sciences, Associate Professor,
Senior Instructor at the Foreign Languages Department
National Academy of the State Border Guard Service of Ukraine
(Khmelnytskyi, Ukraine) irchikan79@ukr.net

Yurii BETS,
Candidate of Pedagogical Sciences, Associate Professor,
Senior Instructor at the Foreign Languages Department
National Academy of the State Border Guard Service of Ukraine
(Khmelnytskyi, Ukraine) yuzhm75@i.ua

CURRENT TRENDS IN THE RECOGNITION AND DECODING OF PHRASEOLOGICAL UNITS

The article defines that one of the key problems in natural language processing is the recognition of phraseological units. The authors argue that automatic phrase recognition is a rather complex task that requires a combination of linguistic knowledge, machine learning algorithms, and computer technology. Rule-based approaches are one of the most common methods for identifying and categorizing phraseological units in natural language processing. These methods operate on a set of predefined rules to detect and extract phrases from text using syntactic and semantic patterns. The article notes that one of the main advantages of machine learning methods in phrase detection is their ability to effectively cope with the complexity and variability of natural language. According to the authors, rule-based approaches to phrase recognition are mostly implemented in the form of software that can be integrated into various NLP tools and platforms, such as: Natural Language Toolkit (NLTK); Stanford CoreNLP, where CoreNLP is a set of NLP tools; Apache OpenNLP: OpenNLP (an open source NLP library); GATE: General Architecture for Text Engineering (GATE) (an open-source platform for building NLP applications); spaCy: (a Python-based NLP library that includes a rule-based search engine for identifying phrases in text). The authors identify different types of rule-based approaches that can be used for phrase recognition, including pattern matching, rule induction, and decision trees. However, one of the main challenges is the creation of a comprehensive set of rules that can accurately identify and classify all possible phraseological units in a given language, which can be particularly difficult for languages with complex grammatical structures or large vocabularies. The authors argue that a significant advantage of machine learning methods in phrase detection is their ability to effectively cope with the complexity and variability of natural language.

The paper identifies the basic principles that guide how machine learning approaches work for phrase recognition. These include: Unsupervised Learning, neural networks, Deep Learning.

Key words: *phraseological units, phrases, machine learning approaches, rule-based approaches, natural language.*

Ірина БАСАРАБА,
orcid.org/0000-0002-3209-9119
доктор філософії в філології,
викладач кафедри іноземних мов
Національної академії Державної прикордонної служби України
(Хмельницький, Україна) irynaborovyk2017@ukr.net

Ірина БЕЦ,
orcid.org/0000-0001-8241-5493
кандидат педагогічних наук, доцент,
старший викладач кафедри іноземних мов
Національної академії Державної прикордонної служби України
(Хмельницький, Україна) irchikan79@ukr.net

Юрій БЕЦ,
orcid.org/0000-0002-2679-3788
кандидат педагогічних наук, доцент,
старший викладач кафедри іноземних мов
Національної академії Державної прикордонної служби України
(Хмельницький, Україна) yuzhm75@i.ua

СУЧАСНІ ТЕНДЕНЦІЇ РОЗПІЗНАВАННЯ ТА ДЕКОДУВАННЯ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ

В статті визначається, що однією з ключових проблем в обробці природної мови є розпізнавання фразеологічних одиниць. Автори стверджують, що автоматичне розпізнавання фразеологізмів є достатньо складним завданням, що потребує сукупності лінгвістичних знань, алгоритмів машинного навчання та обчислювальної техніки. Підходи, що базуються на правилах передбачені заздалегідь (Rule-based approaches), є одними з найпоширеніших методів визначення та категоризації фразеологічних одиниць у обробці природної мови. Ці методи оперують набором заздалегідь визначених правил для виявлення та виокремлення фраз з тексту, використовуючи синтаксичні та семантичні шаблони. В статті зазначається, що однією з головних переваг методів машинного навчання у визначенні фраз є їхня здатність ефективно впоратися зі складністю та змінністю природної мови. На думку авторів підходи на основі правил для розпізнавання фраз переважно реалізуються у вигляді програмних, що можуть бути інтегровані в різні інструменти та платформи NLP, такі як: Natural Language Toolkit (NLTK); Stanford CoreNLP, де CoreNLP - це набір інструментів NLP; Apache OpenNLP: OpenNLP (бібліотека NLP з відкритим вихідним кодом); GATE: General Architecture for Text Engineering (GATE) (платформа з відкритим вихідним кодом для створення NLP-додатків); spaCy: (бібліотека NLP на основі Python, яка включає пошуковик на основі правил для ідентифікації фраз у тексті). Автори визначають різні типи підходів на основі правил, котрі можна використовувати для розпізнавання фразеологічних одиниць, зокрема зіставлення зі зразком, індукція правил і дерева рішень. Проте однією з головних проблем визначається створення всеосяжного набору правил, які можуть точно ідентифікувати та класифікувати всі можливі фразеологічні одиниці в певній мові, що може бути особливо складно для мов зі складною граматичною структурою або великим словниковим запасом. Автори стверджують, що вагомою перевагою методів машинного навчання у визначенні фраз є їхня здатність ефективно впоратися зі складністю та змінністю природної мови.

В статті визначаються основні принципи, які керують тим, як підходи машинного навчання працюють для розпізнавання фраз. До таких відносяться: Unsupervised Learning , нейронні мережі, Deep Learning.

Ключові слова: *фразеологічні одиниці, фрази, підходи машинного навчання, rule-based approaches, природня мова.*

Problem statement. In recent years, much attention has been paid to the study of natural language processing, as the development of technology and artificial intelligence has made it possible to develop systems capable of analyzing and understanding human speech. One of the key problems in natural language processing is the recognition of phraseological units that are multi-word expressions that have a fixed meaning and are used in a certain context. Phraseological units are an integral part of language, and their recognition is critical for many natural language processing applications, such as text mining, information retrieval, and machine translation.

Automatic phrase recognition is a challenging task that requires a combination of linguistic knowledge, computing, and machine learning algorithms. Many existing automatic phrase recognition systems are rule-based, i.e. they rely on manually created rules to identify and extract these expressions. However, these systems can be limited in their ability to recognize new and context-dependent phraseological units, and their development and maintenance requires significant manual effort.

Analysis of recent research and publications. Linguists pay attention to various aspects of the study of phraseological units, including theoretical, stylistic, lexicographic, methodological and practical. However, there is a lack of attention to the synthesis, processing and improvement of methods and means to prevent the loss of meaning in the array of phraseological units in order to ensure authenticity in interlingual transformation.

The analysis of foreign and domestic scientific works on the problems of translation of phraseological units shows that S. Denysenko, L. Skrypnyk, Yu. Firsova, L. Bulakhovskyi, A. Denysova, M. Kovalchuk, L. Komar, V. Telia, I. Bekhta have been engaged in generalizing the issues of phraseology theory. Among foreign linguists, it is worth noting W. J. Ball, F. Boers, M. Callies, Ch. Bally, H. Colston, A. P. Cowie, A. Cutler, S. Fiedler, Ch. Fernando et al.

Despite considerable research by domestic and foreign experts, the problem of recognizing and identifying phraseological expressions in the text remains unresolved.

Thus, ***the purpose of the article*** is to summarize the existing methods of recognizing and decoding phraseological units using an automation element.

Presentation of the main research material. It is generally accepted that phraseological units, also known as polysyllabic expressions, are groups of words that form a single semantic unit or have a fixed and idiomatic meaning. These units are a fundamental aspect of natural language and are widely used in everyday communication. Phraseological units are important to consider when processing natural language because their meaning cannot be fully understood by simply analyzing the individual words that make up the unit. Instead, it is necessary to understand the meaning and function of the whole unit (Wilson, 2016: 44).

The study of phraseology is an important part of linguistics and natural language processing, as it helps to understand the complex structure of language and the ways it is used in communication (Cowie, 2007: 324). In recent years, there has been a growing interest in the study of phraseological units, and several approaches to their identification, categorization, and analysis have been proposed.

In English, there are different types of phrases, each of which has its own unique characteristics and functions. Here are some of the most common types:

1. Idioms. An idiom is a phrase whose meaning cannot be derived from the literal meaning of its constituent words. Idioms are widely used in everyday speech, and they can be either fixed or flexible. Fixed idioms are those that have a certain structure and cannot be changed, such as “kick the bucket” or “break a leg”, while flexible idioms are those that can have some variations in structure, such as “spill the beans” or “let the cat out of the bag”.

2. Lexical Collocations. A collocation is a pair or group of words that often occur together and form a natural combination. Unlike idioms, collocations have a more transparent meaning that can be understood from the individual meanings of the words. Examples of phrases: “heavy rain”, “make a decision” or “take a shower” (Hoey, 2005: 247).

3. Proverbs. Proverbs are short, memorable statements that express a general truth or advice. They are often used to provide guidance or to reinforce social norms. Proverbs tend to have a fixed form and often have a rhyme or rhythm that makes them easier to remember. Examples of proverbs: “Actions speak louder than words”, “When in Roma, do as the Romans do” or “Don’t judge a book by its cover”.

4. Fixed Expressions. Fixed expressions are phrases that have a certain structure and meaning and cannot be changed. They are often used to convey a certain idea or achieve a certain effect. Examples of stable expressions are “How do you do?”, “Goodbye” and “Thank you very much”.

5. Formulaic Language. Formulaic language is a variety of wordy expressions that are commonly used in certain contexts or situations. Formulaic language includes routine expressions such as greetings and farewells, as well as speech acts such as apologies, requests, and compliments. Formulaic language can also include ready-made patterns, such as “If you are X, then you are Y” or “X is not only Y, but also Z”. (Williams, 2018: 68)

6. Phrasal Verbs. Phrasal verbs are verbal phrases that consist of a verb and one or more particles (prepositions or adverbs). They often have a figurative meaning that cannot be derived from individual words. Examples of phrasal verbs: “take off”, “put up with”, and “run into”.

7. Binomials. Binomials are pairs of words that are connected by a conjunction and function as a whole. They are often used to express contrast or emphasis. Examples of binomials: “black and white”, “thick and thin”, “odds and ends”.

8. Discourse Markers are words or phrases used to indicate the relationship between sentences or clauses. They are often used to indicate a change of topic, to show contrast, or to express the speaker’s attitude or position. Examples of discourse markers include “however”, “moreover”, “on the other hand” and “in my opinion” (Granger, Meunier, 2008: 705).

One of the most common approaches to identifying phraseological units is corpus linguistics, which involves analyzing large collections of texts to identify common patterns and structures (Clark, Fox, Lapin, 2008: 82). Other approaches include rule-based methods, statistical methods, and hybrid methods that combine different approaches. However, to summarize the role of phraseological units in a text, we can identify the following:

1. Improving communication: Phraseological units play an important role in improving communication and making language more effective. They allow speakers and writers to convey complex ideas and meanings in a concise and more accurate way, reducing the need for long explanations.

2. Cultural relevance: Many phraseological units have cultural significance and are often used in literature, art, and other media. Understanding these units is important for gaining insight into the culture and history of a language.

3. Humor and irony: Idiomatic phrases are often used for humor and irony, adding depth and complexity to language. For example, the phrase "It's raining cats and dogs" is a phraseological unit used to describe heavy rain, but it can also be used to create a somewhat comical effect.

4. Expression of emotions: Phraseological units are often used to express emotions, conveying the mood and tone of the speaker. For example, the phrase "It's a piece of cake" is used to create an atmosphere of lightness and simplicity, while the phrase "It's the end of the world" expresses disappointment.

5. Marketing and advertising: phraseological units are often used in advertising and marketing to create slogans and catchphrases that are easy to remember. For example, the phrase "Just do it" is a world-famous slogan used by Nike to promote its products.

6. Linguistic variation: The use of phraseological units varies across social groups and regions, creating linguistic diversity and variation. Understanding these differences is crucial for effective communication and building strong relationships with people from different backgrounds.

In summary, phraseological units play a vital role in language and communication, increasing efficiency, adding depth and complexity, while conveying cultural meaning and emotion. Understanding them can help improve language understanding and production, as well as contribute to the development of natural language processing tools and techniques (Clark, 2019: 57).

Rule-based approaches for phrase recognition are among the most common methods for identifying and categorizing phraseological units in natural language processing. These methods rely on a set of predefined rules to identify and extract phrases from text based on syntactic and semantic patterns. Rule-based approaches are widely used in various natural language processing tasks, including machine translation, text classification, and information retrieval (Jurafsky, Martin, 2019: 178).

There are several different types of rule-based approaches that can be used for phrase recognition, including pattern matching, rule induction, and decision trees. Pattern matching involves searching for certain syntactic or semantic patterns in the text and identifying phrases based on these patterns. Rule induction involves the automatic generation of rules based on a set of training data, while decision trees use a set of rules organized in a hierarchical structure to classify text (Grishman, 2008: 583).

Rule-based approaches to phrase recognition are mostly implemented as software rather than hardware. These programs can be integrated into various NLP tools and platforms, such as:

1. Natural Language Toolkit (NLTK): NLTK is a widely used platform for building NLP applications in Python. It includes modules for rule-based phrase recognition, such as the RegexpParser module.

2. Stanford CoreNLP: CoreNLP is a suite of NLP tools developed by Stanford University. It includes a rule-based parser for identifying phrases in text that can be used for tasks such as named entity recognition and sentiment analysis.

3. Apache OpenNLP: OpenNLP is an open source NLP library that includes a rule-based parser for identifying phrases in text. It can be used for tasks such as recognizing named objects, tagging parts of speech, and identifying main links.

4. GATE: General Architecture for Text Engineering (GATE) is an open source platform for building NLP applications. It includes a rule-based Gazetteer tool for identifying specific phrases and entities in a text.

5. spaCy: spaCy is a Python-based NLP library that includes a rule-based search engine for identifying phrases in text. It can be used for tasks such as named entity recognition, dependency parsing, and sentence segmentation (Struhl, 2010: 290).

In general, rule-based approaches to phrase recognition remain an important area of research in the field of natural language processing. Despite their limitations, they offer a flexible and powerful tool for automating language processing tasks and can be highly effective if adapted to a specific task or subject area (Johnson, 2019: 113).

In recent years, the development of deep learning algorithms, such as neural networks, has led to significant progress in machine learning approaches for phrase recognition. Deep learning algorithms are based on artificial neural networks that are designed to mimic the structure and functions of the human brain. These algorithms are capable of learning from very large data sets and can automatically identify and extract features from raw data, making them highly effective for natural language processing tasks. In addition, machine learning approaches for phrase recognition are increasingly being used in social media data analysis, where the ability to recognize and isolate phrases and idiomatic expressions is essential for understanding and analyzing the mood and tone of user-generated content.

Machine learning approaches for phrase recognition are a type of natural language processing technique that uses algorithms to identify and extract phrases or idiomatic expressions from text. These algorithms are based on statistical models and use large data sets to train the system to recognize patterns and identify phrases with a high degree of accuracy. Using machine learning for phrase recognition has become increasingly popular in recent years due to the growing demand for accurate and efficient natural language processing tools.

One of the key advantages of machine learning approaches to phrase recognition is their ability to cope with the complexity and variability of natural language. Traditional rule-based approaches require a large set of manually created rules and templates to identify phrases, which can be time-consuming and difficult to maintain. Machine learning algorithms, on the other hand, can automatically learn from data and identify patterns that may not be immediately obvious to human analysts. A significant advantage of machine learning methods in phrase detection is their ability to constantly adapt and improve over time. Gradually adding more data allows the system to learn to recognize new phrases and adapt to changes in speech. This makes machine learning approaches to phrase detection very scalable and efficient in handling large amounts of text.

There are various principles that guide how machine learning approaches work to recognize phrases. Here are some examples:

1. Supervised Learning. In supervised learning, the algorithm is trained on a labeled dataset where each phrase is labeled with a corresponding category. The algorithm learns to recognize patterns in the data and creates a model. This model can then be used to predict the category of new phrases.

2. Unsupervised Learning. In unsupervised learning, the algorithm is provided with an unlabeled data set and is tasked with finding patterns or similarities in the data. The algorithm groups similar phrases based on their features and creates a model. This model can then be used to identify new phrases that belong to the same category.

3. Neural networks are a type of machine learning algorithm that is modeled after the structure and function of the human brain. They consist of layers of interconnected nodes that process and analyze data. In phrase recognition, neural networks can be trained to identify patterns in text and classify phrases accordingly. 4. Deep Learning. Deep learning is a type of machine learning that uses neural networks with many layers to analyze and process data. Deep learning algorithms can be trained to recognize complex patterns in text and detect subtle nuances in language that may not be obvious to other machine learning approaches.

Conclusions. Thus, machine learning approaches to phrase recognition have become an important tool for natural language processing tasks. They have significant advantages over traditional rule-based approaches, including the ability to cope with the complexity and variability of natural language, adapt and improve over time, and process large amounts of text. As the demand for accurate and efficient natural language processing tools continues to grow, the use of machine learning approaches for phrase recognition is likely to become even more widespread and important in the coming years.

BIBLIOGRAPHY

1. Clark A., Fox C., Lappin S. The Handbook of Computational Linguistics and Natural Language Processing. Wiley-Blackwel, 2008. P. 82.

2. Clark A. Role of Phraseology in Second Language Acquisition (Part of the publication: Conference Materials): *International Conference on Language Teaching and Learning,* 2019. Pp. 55-68.

3. Cowie A. P. Phraseology and Culture in English. Oxford: Clarendon Press, 2007. P. 324.

4. Dan Jurafsky, James H. Martin Speech and Language Processing: An Introduction to Natural Language Processing. *Computational Linguistics and Speech Recognition*, 2019. Pp. 175-195.

5. Granger S. Meunier F. Phraseology: An Interdisciplinary Perspective. John Benjamins, 2008. P. 705.

6. Grishman R. Computational Linguistics: An Introduction. Coling, 2008. P. 583.

7. Johnson R. Deep Learning Approaches for Sentiment Analysis. (Part of the publication: Conference Materials). *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Dublin, 2019. Pp. 102-117.

8. M. Hoey Lexical Priming: A New Theory of Words and Language. Routledge, 2005. P. 247.

9. Struhl S. Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence. *Techniques, tools, and methodologies for extracting meaningful insights from text data.* Vol. 4, No. 3, 2010. Pp. 285-294.

10. Williams A. Phraseology in Translation: Challenges and Strategies. (Part of the publication: Conference Materials). *Proceedings of the International Conference on Translation Studies*. Barcelona, 2008. Pp. 67-80.

11. Wilson K. Lexical Semantics and Word Embeddings. (Part of the publication: Conference Materials): *Empirical Methods in Natural Language Processing*. Hong Kong, 2016. Pp. 42-55.

REFERENCES

1. Clark A., Fox C., Lappin S. (2008) The Handbook of Computational Linguistics and Natural Language Processing. Wiley-Blackwel. 82.

2. Clark A. (2019) Role of Phraseology in Second Language Acquisition (Part of the publication: Conference Materials): *International Conference on Language Teaching and Learning.* 55-68.

3. Cowie A. P. (2007) Phraseology and Culture in English. Oxford: Clarendon Press. 324.

4. Dan Jurafsky, James H. Martin. (2019) Speech and Language Processing: An Introduction to Natural Language Processing *Computational Linguistics and Speech Recognition*. 175-195.

5. Granger S. (2008) Meunier F. Phraseology: An Interdisciplinary Perspective. John Benjamins. 705.

6. Grishman R. (2008) Computational Linguistics: An Introduction. Coling. 583.

7. Johnson R. (2019) Deep Learning Approaches for Sentiment Analysis. (Part of the publication: Conference Materials). *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases,* Dublin. 102-117.

8. M. Hoey (2005) Lexical Priming: A New Theory of Words and Language. Routledge. 247.

9. Struhl S. (2010) Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence. *Techniques, tools, and methodologies for extracting meaningful insights from text data.* Vol. 4, No. 3. 285-294

10. Williams A. (2018) Phraseology in Translation: Challenges and Strategies. (Part of the publication: Conference Materials). *Proceedings of the International Conference on Translation Studies*, Barcelona. 67-80.

11. Wilson K. (2016) Lexical Semantics and Word Embeddings. (Part of the publication: Conference Materials): *Empirical Methods in Natural Language Processing,* Hong Kong. 42-55.