Methods of named entities recognition of location type in unstructured Ukrainian texts

Victoria Vysotska[1,†], Kirill Smelyakov[2†], Anastasiya Chupryna[2,†] and Kateryna Liulina[2,*,†]

[1] *Kharkiv National University of Internal Affairs, L. Landau Avenue 27 61080 Kharkiv, Ukraine*

[2] *Kharkiv National University of Radio Electronics, Nauky Ave 14, 61166, Kharkiv, Ukraine*

Abstract

This paper presents a hybrid method for identifying named location entities in unstructured Ukrainian text. The proposed system combines machine learning methods, using the Stanza library for initial named entity detection, with rule-based algorithms that leverage the Universal Dependencies framework to identify prepositions of place and entity expansions through parsing. This enables the recognition of complex location cues, such as relative directions and descriptive elements. Testing on a dataset of 200 Ukrainian texts, including fiction and social media posts, showed that the hybrid approach achieves an F-score of 80.79%, a precision of 82.88%, and a recall of 78.81%, outperforming individual machine learning methods (F-score of 44.31%) or rule-based approaches alone. The proposed named entity recognition scheme for locations performs well on texts taken from news and social media, but highlights the need for greater diversity in training data to handle literary styles. The proposed method can be integrated into modern emergency response systems, delivery services, and geographic information systems.

Keywords

named entity recognition, machine learning, location, rule-based algorithm, Ukranian language

1. Introduction

Natural language processing systems continue to be explored and studied for different tasks, especially named entity recognition for extraction of structured information from unstructured text. While significant progress has been made for languages such as English, Chinese, Spanish, and other European languages, significant challenges remain for other languages. The challenges of extracting named entities from unstructured text include the limited annotated corpora, the morphological complexity of languages, and their linguistic specificity. Ukrainian is a rich language with a well-developed morphology, including a complex case system, extensive inflection, and diverse syntactic constructions. Ukrainian's flexibility comes from its case system, which lets words be arranged freely. This creates specific problems for recognizing named entities in text.

Location recognition is crucial in emergency response systems, delivery services, and geographic information systems. These systems require not only the identification of standard toponyms (cities, countries, streets), but also more complex location details, which may include relative directions, descriptive words, and contextual elements that establish precise georeferencing.

This research examines methods and approaches for recognizing named entities of locations and places in unstructured Ukrainian-language text. We propose a combined approach to entity recognition using statistical machine learning and rule-based methods. Our methodology utilizes the Universal Dependencies framework to identify relationships between words. Analysis of these relationships enables the recognition of complex location descriptions by detecting syntactically related components.

[*] *CIAW-25: Computational Intelligence Application Workshop, September 26-27, 2025, Lviv, Ukraine*

[*] Corresponding author.

[†] These authors contributed equally.

victoria.a.vysotska@lpnu.ua (V. Vysotska); kirillsmelyakov@gmail.com (K. Smelyakov); anastasiya.chupryna@nure.ua (A. Chupryna); kateryna.liulina@gmail.com (K. Liulina)

0000-0001-6417-3689 (V. Vysotska); 0000-0001-9938-5489 (K. Smelyakov); 0000-0003-0394-9900 (A. Chupryna)

The primary contributions of this work include the development of a hybrid system for recognizing named entities of locations and places in Ukrainian-language text and the evaluation of various architectural configurations combining machine learning and rule-based methods.

2. Related Work

2.1. Research in the field of intelligent text data processing

The article "Effectiveness of Modern Text Recognition Solutions and Tools for Common Data Sources" examines the capabilities of optical character recognition (OCR) using tools such as EasyOCR and TesserOCR [1]. The authors evaluate their performance on various data types, including electronic documents, web pages, and advertising banners, and offer recommendations for the use of these tools, taking into account the degree of data corruption. This study is of interest in the context of this work, as it proposes mechanisms for expanding the capabilities of converting various data types into a format suitable for further text processing. For example, in the context of emergency response systems, this solution could be implemented to analyze images provided by witnesses to automatically identify additional information about the scene.

The paper "Methods of Multilingual Question Answering" examined question-answering systems for various languages, with a particular focus on the Ukrainian language, for which machine learning data is typically lacking [2]. They explore various approaches to creating such systems using BERT-based language models. It was found that tuning the model to English improves its performance with the Ukrainian language. Together, these studies demonstrate how text information is processed in intelligent systems today. First, text is extracted from visual sources. The system then recognizes the text, understands its meaning, analyzes questions, and finds appropriate answers.

Both studies are important in that they offer practical solutions that expand the capabilities of obtaining information about incident locations and location markers and help convert user-provided information into an acceptable format for further text analysis and extraction of useful information, particularly named location entities.

Modern researchers also point out that the development of new Ukrainian-language datasets is a pressing issue [2, 3]. This problem stems from the fact that some machine learning algorithms require significant datasets for research, especially when applied to a specific application domain. These considerations also suggest that text processing methods that are less dependent on the comprehensiveness of the provided dataset may be of particular interest.

It is worth noting that research in natural language processing, as applied to the Ukrainian language, remains relevant. Contemporary researchers are developing approaches to improving the quality of Ukrainian source text for further processing, specifically proposing solutions for eliminating errors in Ukrainian text [4].

2.2. Common approaches to Named Entity Recognition

Named Entity Recognition methods can be based individually or simultaneously on lexicon-based, rule-based, and machine learning-based approaches. Lexicon-based methods can be used in settings where a corpus is not available. These models combine the results of morphological analysis, a set of lexicons, and stemming and lemmatization methods. Rule-based methods can be used for entities that have a specific structure that can be constructed according to predefined patterns and rules. They can be used to recognize entities such as phone numbers, email addresses, personal data (ID numbers, bank card numbers), dates, and other data in a predefined format [5]. Machine learning-based methods are more adaptive and require the creation of training data for predictive text labeling based on predefined named entities and their categories.

2.3. Applied research in Named Entity Recognition

Studies in named entity recognition explore different ways of dealing with data limits and language problems. The paper "Named Entity Recognition for Sensitive Data Discovery in Portuguese" examined and analyzed methods for recognizing named entities in Portuguese source texts. The authors proposed a comprehensive solution for named entity recognition that involves rule-based and grammatical pattern-based methods; dictionary-based methods; and machine learning methods. It was found that the most effective machine learning methods for identifying named entities of locations in Portuguese are Bi-LSTM (bidirectional LSTM) and CRF (Conditional Random Fields) models [5]. Similar methods have been used to recognize named location entities in Chinese text and have demonstrated high efficiency in quickly identifying locations in mission-critical systems [6, 7].

The paper "Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model" demonstrated that the developed model can effectively identify fine-grained categories of geographic entities in Chinese [8]. This model incorporates components such as ALBERT (A Lite BERT), a lightweight version of BERT for efficient feature extraction, a bidirectional LSTM for capturing contextual dependencies, and a CRF layer for optimal sequential labeling.

The paper "Location Named-Entity Recognition using a Rule-Based Approach for Balinese Texts" investigated the effectiveness of rule-based methods for identifying named location entities in Balinese source texts. It was found that rule-based methods are effective in identifying specific locations based on given rules and patterns. It was found that location entities that do not follow expected patterns are better recognized using machine learning methods [9].

Researchers often note that hybrid methods are effective for recognizing named location entities when processing text in different natural languages. For example, the paper "Location Reference Recognition from Texts: A Survey and Comparison" found that combining regular expressions and grammar rules with machine learning classifiers enables higher F1 scores for location extraction tasks [10].

The patent "Efficient and Accurate Method and Apparatus for Recognizing Named Entities" describes an algorithm for extracting and recognizing named entities in digital documents and text. This patent describes an approach to recognizing named entities using machine learning algorithms in combination with rule-based algorithms. The results of both algorithms are analyzed separately and combined to make a final decision on whether the recognized entity is the desired one. This patent proposes the use of two sets of rules for identifying named entities: a general set and a specific set of rules. The first set includes statistical information about how frequently used words typically refer to certain named entities or are used as a specific part of speech. The specific set of rules defines patterns and regular expressions that typically describe specific types of named entities [11].

Thus, it can be concluded that to solve the problem of recognizing specific types of named entities, it is worthwhile to conduct research on both named entity recognition methods based on machine learning and to analyze the features of the language to determine a set of specific rules that help recognize concrete types of named entities.

3. Methodology

To solve the problem of recognizing named location entities in Ukrainian text, we propose a hybrid system consisting of several modules: a module for preprocessing the incoming text; a module for recognizing named location entities using machine learning; a module for recognizing prepositions of place using a rule-based algorithm; a module for expanding recognized named entities using a rule-based algorithm; and a module for outputting the text processing results. The schematic of the proposed solution is shown in Figure 1.

At the initial stage, the text preprocessing module splits the text into sentences, and the sentences into tokens; determines parts of speech; analyzes morphological features of words; and identifies relationships between words in a sentence. To implement this module, we propose using the Stanza library, which supports the Ukrainian language and contains ready-made trained models using the uk-lang corpus [12, 13].

The second stage involves recognizing keywords in named location entities. We propose using a module for recognizing named location entities using machine learning. This module uses the Stanza library, which provides recognition capabilities for named location entities in text. As a result, keywords indicating locations are recognized. To detect dependent words, an additional module recognizing prepositions of place is proposed. This module analyzes the morphological properties of a word and, if a spatial preposition is detected, marks it as a separately recognized named location entity. Spatial prepositions in Ukrainian include words such as "у» (in), "на" (on), "біля" (near), "за" (behind), and others. They signal potential location references.

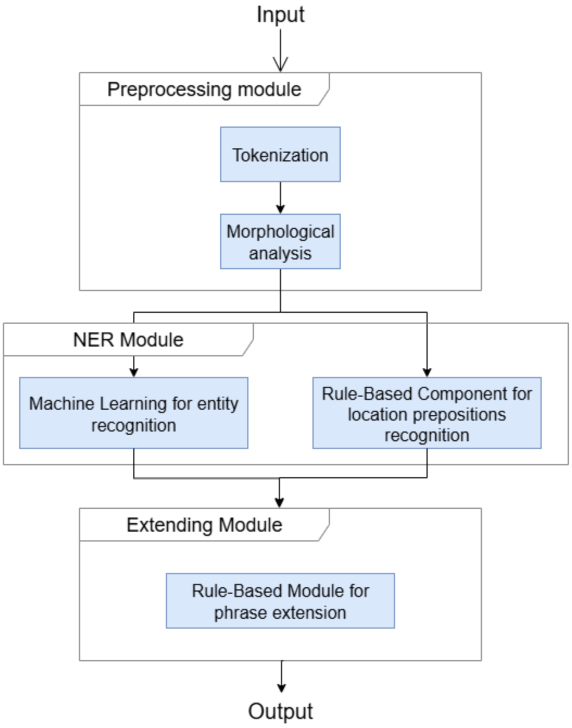
Input
Preprocessing module
Tokenization
Morphological analysis
NER Module
Machine Learning for entity recognition
Rule-Based Component for location prepositions recognition
Extending Module
Rule-Based Module for phrase extension
Output

Figure 1: Scheme of a system for recognizing named entities.

The third stage uses a module to expand recognized named entities using a rule-based algorithm. This stage is necessary for identifying dependent words that can clarify or complement the keywords of named location entities. Clarifying words include adjectives, conjunctions, relative nouns, and numerals. Additional location information is extracted by analyzing the relationships between words. This is done using the Universal Dependencies framework, which describes the relationships between words.

3.1. A module for recognizing prepositions of place using a rule-based algorithm

The Universal Dependencies framework describes and systematizes the annotation of grammars for various natural languages [14, 15]. This instruction is intended to describe the parts of languages, the morphological features of words, and the syntactic relationships between them [16-21]. An analysis of the grammar of the Ukrainian language reveals that spatial prepositions are used in most cases to indicate place and location in sentences. Some prepositions and examples of their use are listed in Table 1.

Table 1
Ukranian prepositions of location

Preposition (UA)	Preposition (EN)	Example (UA)	Example (EN)
у, в	in	У місті, в Ужгороді, у саду, в театрі	In the city, in Uzhhorod, in the garden, in the theater
на	on	На дорозі, на столі	On the road, on the table
від	from	Від дороги, від школи	From the road, from the school
під	under	Під столом, під деревом	Under the table, under the tree
за	behind	За будинком, за лікарнею	Behind the house, behind the hospital
перед	in front of	Перед будинком, перед машиною	In front of the house, in front of the car
по	on	По вулиці, по воді	On the street, on the water
до	to	До школи, до зупинки	To school, to the bus stop

When forming phrases with prepositions, a dependency is formed between words, where the preposition of place is the dependent word, and the noun or pronoun forming the phrase acts as the main word. The Universal Dependencies framework annotates such dependencies as “Case”. A “Case” dependency is used to indicate relationships between nouns and other parts of speech (adjectives, numerals, prepositions). There are also subtypes of Case dependencies, one of which is “Loc”. The “Loc” identifier is used to indicate relationships between words indicating place or time. For example, the sentence “In September I was in Greece” contains two “Loc”-type case dependencies: the first is “In September”, which denotes time; the second is “in Greece”, which denotes place. By identifying such dependencies, it is possible to discover keywords that indicate either place or time.

“Loc”-type case dependency analysis can be used to develop a rule-based algorithm for identifying locations in unstructured text. This approach allows to recognize additional named location entities that would otherwise be missed by machine-learning-based algorithms.

3.2. A module for expanding named entities using a rule-based algorithm

The entity expansion algorithm uses recursive analysis of syntactic relationships to identify words related to location words. It distinguishes between the expansion of head words by dependent words and vice versa.

For nominal relationships (nmod, amod, nummod, det), bidirectional expansion is used. That is, when a head word is detected, all dependent words are included, and when a dependent word is detected, the head word and its other dependent words are included. This helps accurately define complex location descriptions while avoiding redundancy.

For structural relationships (appos, conj), related elements are considered equivalent, and expansion involves identifying dependent words for each equivalent component. For acl (adjective clause) relationships, the algorithm analyzes the structure of the dependent clause to capture detailed descriptive content.

4. Experimental Setup

4.1. Dataset Construction

The Ukrainian corpus lang-uk, comprising 262 texts, was used as the Ukrainian-language corpus for training named entity recognition models. The primary source is the open corpus of Ukrainian texts, the Brown Corpus of the Ukrainian Language.

Ukrainian fiction and a set of messages from Telegram news channels were used as additional data for experiments on the effectiveness of location entity recognition. These data include sentences specifying locations using complex literary expressions, as well as live messages in Ukrainian from modern native speakers.

The following text sources were used for analyzing location entities from fiction: the social and everyday novella "Kaidasheva Family" by Ivan Nechuy-Levytsky, the autobiographical novella "Ocharovannaya Desna" by Oleksandr Dovzhenko, and the novel "Do Oxen Roar as if the Manger is Full?" by Panas Myrnyi.

When forming the test data sample, 20% of the data was taken from fiction and 80% from news channels. This distribution is primarily due to the fact that the channels contain texts in the format in which system users might specify and describe locations in third-party service offers. For example, Telegram messages include messages such as "If anyone can, please check if anyone lives on Solnechnaya Street and Molochnaya Street, 5 (opposite the Department Store)" or "Does anyone know where Alex and his family are now? They lived on Nizhnyaya Street and were in the house near the clinic." These offers include both partial address information and additional details that could help pinpoint a more precise location.

To conduct the experiments, a test dataset in Ukrainian was created containing 200 records. It was taken 40 records from fiction and 160 records from Telegram news channels.

4.2. Evaluation Metrics

Standard information retrieval metrics are used to evaluate the system's performance: precision, recall, and F-score. These metrics are calculated at the level of individual words, rather than entire named entities. This is necessary to account for complex, wordy location descriptions, which are the primary focus of this paper.

The Precision metric calculates the percentage of correctly identified entities among all recognized entities (see Formula 1).

$$Precision=\frac{TP}{TP+FP}, \tag{1}$$

where TP – a number of true positive results, FP – a number of false positive results.

The Recall metric reflects the percentage of correctly recognized named entities among those that should have been recognized (see Formula 2).

$$Recall=\frac{TP}{TP+FN}, \tag{2}$$

where TP – a number of true positive results, FN – a number of false negative results.

The F-score is a balanced metric of the method's performance and is calculated using the following formula (see Formula 3).

$$F=2\frac{Precision\cdot Recall}{Precision+Recall}, \tag{3}$$

4.3. System Configuration

During the experiment, the effectiveness of each system module used to solve the problem of recognizing named entities of locations is assessed. The following system configurations are proposed:

- Using only the named entity recognition module based on machine learning (Configuration 1 – ML module).
- Using the named entity recognition module based on machine learning and a module for expanding recognized named entities using a rule-based algorithm (Configuration 2 – ML module + Extending module).
- Using a module for recognizing prepositions of place using a rule-based algorithm and a module for expanding recognized named entities using a rule-based algorithm (Configuration 3 – Rule-based module + Extending module)
- Using the named entity recognition module based on machine learning, a module for recognizing prepositions of place using a rule-based algorithm, and a module for recognizing prepositions of place using a rule-based algorithm (Configuration 4 – ML module + Rule-based module + Extending module).

5. Results

5.1. Overall Performance

Table 2 provides detailed performance evaluation results for all system configurations tested on the full dataset, consisting of 200 samples. The results demonstrate a clear performance improvement using hybrid architectural solutions.

Table 2
Experimental results on full test dataset

Configuration	F-score	Precision	Recall
ML module	44.31%	97.41%	28.68%
ML module + Extending module	65.22%	81.46%	54.38%
Rule-based module + Extending module	71.19%	83.10%	62.27%
ML module + Rule-based module + Extending module	80.79%	82.88%	78.81%

Although pure machine learning module (Configuration 1) shows very high precision (97.41%), its recall is quite low (28.68%), resulting in a mediocre F-score (44.31%). This suggests that this method is good at identifying frequently occurring, standard place-name descriptors but misses non-standard location descriptions.

Combination of the machine learning module with the module for extracting additional words using rule-based algorithm (Configuration 2) significantly increases recall to 54.38% while maintaining acceptable precision (81.46%). As a result, the F-score increases by 20.91% to 65.22%. This demonstrates that syntactic relationship analysis helps more accurately identify named location entities, rather than simply identifying individual words.

The combination of rule-based modules for location prepositions detection and their extension (Configuration 3) performs better than the Configuration 2, with an F-score of 71.19%. Thus, using linguistics to identify spatial prepositions improves the retrieval of named location entities compared to machine learning methods.

The hybrid system (Configuration 4) demonstrates the best results, achieving an F-score of 80.79% and allows for the most accurate detection of named location entity boundaries in text.

5.2. Domain-Specific Analysis

Tables 3 and 4 present a performance analysis showing significant differences depending on the type of text: literary, news, or social media content.

Table 3
Experimental results on test dataset, which consists of fiction texts

Configuration	F-score	Precision	Recall
ML module	6.89%	100%	3.67%
ML module + Extending module	11.23%	100%	5.95%
Rule-based module + Extending module	56%	83.36%	41.66%
ML module + Rule-based module + Extending module	59.37%	86.36%	45.23%

Machine learning methods, even with the recognized named entity extension module enabled, face challenges when working with literary texts, resulting in a significant performance drop in all configurations. Specifically, the machine learning component demonstrates low efficiency in processing literary material, achieving minimal recall rates (3.57% and 5.95%), even with high accuracy. This suggests that the training set lacks literary texts, limiting the applicability of machine learning methods and requiring the development of a specialized dataset for a specific application. Approaches based on linguistic rules demonstrate superior performance in analyzing literary content, achieving an F-score of 56.00% compared to 11.23% for Configuration 2. This superiority indicates that spatial prepositions and syntactic structures maintain their consistency across different text types, while statistical patterns identified based on news texts cannot be generalized to literary language.

Table 4
Experimental results on test dataset, which consists of social media content

Configuration	F-score	Precision	Recall
ML module	52.01%	97.34%	35.48%
ML module + Extending module	73.83%	81.1%	67.76%
Rule-based module + Extending module	74.63%	82.73%	67.98%
ML module + Rule-based module + Extending module	85.16%	82.41%	88.11%

The machine learning component performs better on news texts (F-score 52.01%) than on literary texts (F-score 6.89%). The hybrid system shows significantly better results on news and social media texts (F-score 85.16%) than on literary texts (F-score 59.37%). This is due to the lang-uk training corpus primarily consisting of news and web texts, which matches the characteristics of the target text.

Rule-based approaches to named location entity recognition demonstrate consistent performance across domains (74.63% versus 56.00%). This suggests that syntactic patterns transfer better across domains than statistical models.

5.3. Modules Usage Analisys

An analysis of the contributions of the individual components reveals different trends in the precision-recall tradeoff, as well as complementary strengths. The machine learning component is better at identifying common place names and well-represented feature types, but struggles with morphological inflections and descriptive place names.

The rule-based components demonstrate more consistent results in precision and recall by consistently identifying spatial prepositions and associated noun phrases. This approach allows for the detection of place descriptions that lack explicit names but contain location markers.

The analysis of related words and their identification consistently improves results for all major configurations, especially for machine learning. The 20.91% increase in F-score from usage of machine learning component only to machine learning component usage in combination with relative words extraction module demonstrates the importance of expanding the syntax for fully recognizing named location entities, which may include descriptive elements.

6. Discussion

6.1. Effectiveness of Hybrid Architecture

Analysis shows that different parts of the system have distinct characteristics in terms of precision and recall, as well as how they complement each other. Machine learning is good at recognizing common place names and known object types, but struggles with word inflections and descriptive names.

The parts of the system based on linguistic rules perform more consistently in terms of precision and recall. When recognizing named location entities in unstructured text, identifying spatial prepositions and related word groups is an effective solution. Algorithms based on linguistic text analysis can accurately identify locations that are not explicitly named but contain location references.

Analyzing the dependencies of recognized keywords in named location entities always improves system performance, especially for machine learning. The 20.91% improvement in F-score when incorporating the module for expanding named location entities with dependent qualifying or descriptive words into machine learning suggests that syntactic analysis is essential for accurately identifying place names.

6.2. Domain Transfer Challenges

The significant difference in performance between literary and news texts using the hybrid model (F-score 85.16% versus 59.37%) indicates that there is still potential for developing systems for recognizing named entities in Ukrainian text in specific domains.

Literary texts include archaic place names, poetic descriptions, and narrative language structures, which differ markedly from the modern news content used to train the model. The complete failure of machine learning methods on literary material (F-score 6.89%) suggests the need to retrain the neural network model using a more balanced dataset. This means that robust Ukrainian NER systems require diverse training materials, including different text types and historical language variants.

Recognition methods based on linguistic rules demonstrate good transferability across domains, maintaining acceptable performance on literary material (F-score of 56.00%) compared to their results on news text (F-score of 74.63%). This stability reflects the robustness of syntactic structures and the use of prepositions of place across different text types and historical periods.

6.3. Practical Application

The proposed system can be implemented in the following domains:

- Food and service delivery services;
- Emergency services, including fire departments, medical services, police services, etc.;
- Location analysis in news and chats;
- Taxi service systems.

Rapid response systems can benefit from an 80.79% F-score when extracting location data from user reports and emergency messages.

Delivery services can use this system to extract addresses and identify geographic landmarks in customer correspondence. The balance between precision and recall (82.88% and 78.81%, respectively) enables automated processing of unstructured text with the option of human review.

News analysis and social media monitoring can leverage the high performance of modern text (F-score 85.16%) to determine the geographic location of events and analyze location-based content. This is justified by the fact that the implementation of the proposed scheme for named location entity recognition will enable the processing of complex location descriptions specified in the Ukrainian language, ensuring comprehensive extraction of geographic information.

6.4. Limitations and Future Work

The current system has several limitations that point to avenues for future research. Problems with transfer across domains indicate the need for more training data, including literary works, historical texts, and specialized materials.

It is worth noting that methods for recognizing named location entities based on linguistic rules rely on knowledge of a specific language. Knowledge transfer across languages is an interesting avenue for research. The rules partially described in this paper can be migrated to other languages, but their application still requires reconsideration of the applicability of specific rules to the linguistics of a specific language. Therefore, to make this model applicable to recognizing named location entities in texts of other languages, it would be necessary to introduce a separate module that could automatically generate morphologically correct forms and rules within the target language. It's worth noting that this named entity recognition scheme can still be applied to texts in other languages, including Slavic, thanks to the use of standardized morphological and syntactic markup provided by the Universal Dependencies framework.

Furthermore, when recognizing named location entities in Ukrainian text, the set of rules can be expanded based on the specific features of this language, thereby improving the module's ability to recognize named location entities using rule-based algorithms.

7. Conclusions

This paper proposed a hybrid model for recognizing named location entities in unstructured Ukrainian text. Three main components were proposed for solving this problem. The first component processes the input text, extracts information about words in sentences and the relationships between them, and recognizes named location entities using a trained neural network model. The Stanza library was used to implement this component. To improve the efficiency of named entity recognition, a module for recognizing keywords in named location entities based on morphological analysis of the text using rule-based algorithms was proposed. This module is independent of the dataset configuration and relies solely on the specifics of a particular natural language. To more fully and thoroughly recognize described locations in the text, a module for expanding named location entities through analysis of dependent words was introduced.

This study demonstrates that the combined use of different named entity recognition methods in unstructured text is well suited for identifying place names in the Ukrainian language. An accuracy of 80.79% was achieved by combining machine learning-based and rule-based algorithms for recognizing location names. Experimental studies revealed that the different approaches complement each other. To obtain complete and comprehensive information about given location indicators in text, it is important to analyze the morphological relationships between words, which ensures accurate identification of location names. The significant increase in performance using the combined approaches (accuracy increased by 36.48% compared to standard machine learning) demonstrates that the proposed framework effectively identifies and extracts location information. Based on this, the proposed model can be recommended for use in software systems developed for emergency response services, logistics, and data analysis.

The study revealed challenges in recognizing named location entities in texts of varying styles. This suggests the need to create more diverse and comprehensive Ukrainian-language datasets for training neural networks when developing systems focused on named entity recognition.

Acknowledgements

The research of related work was conducted by Kirill Smelyakov, Kateryna Liulina and Victoria Vysotska. The hybrid schema for location entity recognition in Ukrainian texts was proposed and implemented by Kateryna Liulina, Kirill Smelyalov and Anastasiia Chupryna. The experimental plan was prepared by Victoria Vysotska and Anastasiia Chupryna. The experiments were conducted and analyzed by Kateryna Liulina, Kirill Smelyalov, Anastasiia Chupryna and Victoria Vysotska.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] K. Smelyakov, A. Chupryna, D. Darahan, S. Midina, Effectiveness of modern text recognition solutions and tools for common data sources, in: Proc. 5th Int. Conf. on Computational Linguistics and Intelligent Systems (COLINS 2021), CEUR-WS, vol. 2870, 2021, pp. 154–165.

[2] D. Dashenkov, K. Smelyakov, O. Turuta, Methods of multilanguage question answering, in: Proc. 2021 IEEE 8th Int. Conf. on Problems of Infocommunications, Science and Technology (PIC S&T), IEEE, Kharkiv, Ukraine, 2021, pp. 251–255. doi:10.1109/PICST54195.2021.9772145.

[3] D. Panchenko, D. Maksymenko, O. Turuta, M. Luzan, S. Tytarenko, O. Turuta, Ukrainian News Corpus as text classification benchmark, in: Proc. 17th Int. Conf. on ICT in Education, Research and Industrial Applications. Volume II: Workshops, 2021, pp. 717–726.

[4] R. Fedchuk, V. Vysotska, Mathematical model of a decision support system for identification and correction of errors in Ukrainian texts based on machine learning, CEUR Workshop Proceedings, vol. 4005, 2025. Available at: https://ceur-ws.org/Vol-4005/paper3.pdf.

[5] M. Dias, et al., Named entity recognition for sensitive data discovery in Portuguese, Applied Sciences (2020) 1–15.

[6] Y. Li, L. Luo, X. Zeng, Fine-tuned BERT-BiLSTM-CRF approach for named entity recognition in geological disaster texts, Earth Sci. Informatics 18(2) (2025) 123–135. doi:10.1007/s12145-025-01870-5.

[7] K. Ma, Y. Tan, Z. Xie, Q. Qiu, S. Chen, Chinese toponym recognition with variant neural structures from social media messages based on BERT methods, J. Geogr. Syst. 24(2) (2022) 143–169. doi:10.1007/s10109-022-00375-9.

[8] L. Tao, Z. Xie, D. Xu, K. Ma, Q. Qiu, S. Pan, B. Huang, Geographic named entity recognition using NLP and an improved BERT model, ISPRS Int. J. Geo-Information 11(12) (2022) 598. doi:10.3390/ijgi11120598.

[9] N.A.S. ER, Location named-entity recognition using rule-based approach for Balinese texts, 2021. Available at: https://www.researchgate.net/publication/349518820.

[10] X. Hu, Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, F. Klan, Location reference recognition from texts: a survey and comparison, ACM Comput. Surveys 56(5) (2023) Article 112, 1–37. doi:10.1145/3625819.

[11] World Intellectual Property Organization, Efficient and accurate named entity recognition method and apparatus, WO2020118741, 2020. Available at: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2020118741.

[12] Stanza. Named entity recognition, n.d. Available at: https://stanfordnlp.github.io/stanza/ner.html

[13] Lang-uk. NER annotation of the Ukrainian corpus, n.d. Available at: https://github.com/lang-uk/ner-uk.

[14] Universal Dependencies. Universal Dependencies project, n.d. Available at: https://universaldependencies.org/

[15] M. de Marneffe, C.D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47(2) (2021) 255–308. doi:10.1162/coli_a_00402.

[16] V. Vysotska, Computer linguistic system architecture for Ukrainian language content processing based on machine learning, in: CEUR Workshop Proceedings, vol. 3723, 2024, pp. 133–181.

[17] V. Vysotska, Computer linguistic system modelling for Ukrainian language processing, in: CEUR Workshop Proceedings, vol. 3722, 2024, pp. 288–342.

[18] V. Vysotska, Computer linguistic systems design and development features for Ukrainian language content processing, in: CEUR Workshop Proceedings, vol. 3688, 2024, pp. 229–271.

[19] V. Vysotska, Linguistic intellectual analysis methods for Ukrainian textual content processing, in: CEUR Workshop Proceedings, vol. 3722, 2024, pp. 490–552.

[20] V. Vysotska, K. Przystupa, Y. Kulikov, S. Chyrun, Y. Ushenko, Z. Hu, D. Uhryn, Recognizing fakes, propaganda and disinformation in Ukrainian content based on NLP and machine-learning technology, Int. J. Comput. Netw. Inf. Secur. 17(1) (2025) 92–127.

[21] D. Levkivskyi, V. Vysotska, L. Chyrun, Y. Ushenko, D. Uhryn, C. Hu, Agile methodology of information engineering for semantic annotations categorization and creation in scientific articles based on NLP and machine learning methods, Int. J. Inf. Eng. Electron. Business 17(2) (2025) 1–50. doi:10.5815/ijieeb.2025.02.01.