

РЕФЕРАТ

Актуальність теми. Розвиток сучасного суспільства тісно пов'язаний з автоматизацією процесу обробки інформації, що особливо проявляється в сучасній епосі цифрової трансформації та стрімкого просування інформаційних технологій, коли аналіз великих обсягів текстових даних стає ключовим завданням для багатьох сфер діяльності.

Особливо важливим є розпізнавання іменованих сутностей (NER) в тексті, а саме ідентифікація осіб, місць, організацій та інших важливих категорій, яке знаходить застосування у багатьох галузях, включаючи комп'ютерні науки, лінгвістику, біоінформатику та інші. У бізнесі NER-системи використовуються для аналізу ринків, прогнозування попиту, розподілу ресурсів та ризиків. В сфері медицини вони допомагають в ідентифікації пацієнтів, діагнозів та лікуванні. У правоохоронній діяльності системи NER використовуються для виявлення та прогнозування злочинів, а в галузі науки - для аналізу наукових публікацій та виявлення прихованих або неочевидних знань у працях.

При цьому, з розвитком глибокого навчання та нейронних мереж відкриваються нові можливості для покращення точності та продуктивності систем розпізнавання іменованих сутностей. Інтеграція глибинних нейронних мереж у поєднанні з методами обробки природної мови дозволяє суттєво підвищити ресурсоефективність задач обробки даних у наперед визначені класи завдяки більш ефективному виявленню складних залежностей та шаблонів у текстовій інформації. Отже, розробка систем розпізнавання іменованих сутностей у тексті з використанням глибинних нейронних мереж залишається актуальною задачею, оскільки ця технологія може значно оптимізувати та прискорити роботу зі значними масивами текстових даних.

Мета роботи полягає в аналізі та розробці ефективних моделей для розпізнавання іменованих сутностей на базі класифікаторів обробки природної мови та глибинних нейронних мереж для покращення точності та швидкодії системи з обмеженими вхідними даними. Дослідження спрямоване на порівняння різних підходів розпізнавання для визначення оптимальних

архітектур та методів оптимізації для досягнення максимальної продуктивності.

Об'єктом дослідження є предметно-орієнтована система розпізнавання іменованих сутностей (NER) в тексті, що базується на використанні методів обробки природної мови (NLP) та глибоких нейронних мереж (DNN) зі здатністю до масштабування.

Предметом дослідження є моделі машинного для розпізнавання іменованих об'єктів в тексті, архітектури глибоких нейронних мереж, а також оцінювання продуктивності розроблених систем шляхом емпіричних досліджень і експериментів.

Методи дослідження складаються із загальнонаукових та спеціальних методів пізнання, зокрема:

1. Аналітичний метод: для комплексного вивчення та узагальнення останніх наукових розробок вітчизняних та зарубіжних вчених у сфері розпізнавання іменованих об'єктів (NER) та розробки масштабованих і предметно-орієнтованих систем NER.

2. Метод аналізу та синтезу: для уточнення та конкретизації понятійно-категоріального апарату у сфері NLP, систематизації знань про алгоритми та програмне забезпечення для NER, а також для формулювання теоретичних положень та висновків дослідження.

3. Метод аналогії та порівняння: для виявлення та співставлення особливостей існуючих методів та підходів до NER, а також для обґрунтування вибору та розробки покращених алгоритмів і програмного забезпечення.

4. Систематизація: для розкриття основних тенденцій, закономірностей та принципів розвитку систем NER, а також для формування класифікацій та типологій алгоритмів і програмного забезпечення.

5. Абстрактно-логічний метод: для формулювання гіпотез, теоретичних положень та висновків дослідження, а також для обґрунтування запропонованих новітніх підходів до розробки масштабованих і предметно-орієнтованих NER-систем.

6. Метод дедукції та моделювання: для вироблення пропозицій щодо удосконалення механізмів NER, розробки нових алгоритмів та програмного забезпечення, а також для оцінювання їх ефективності.

7. Експериментальні методи: для проведення досліджень з метою оцінки ефективності розроблених алгоритмів та програмного забезпечення NER на тестових наборах даних.

8. Методи статистичного аналізу: для обробки результатів експериментів, виявлення статистичних закономірностей та формулювання висновків про ефективність запропонованих методів і підходів.

Наукова новизна одержаних результатів полягає в наступному:

1. Представлення удосконалених архітектур та гібридних моделей, що поєднують різні типи нейронних мереж, додаткові шари та механізми обробки даних з метою досягнення вищої продуктивності в задачах NER порівняно з існуючими моделями глибинного навчання.

2. Імплементация та оцінка ефективності розширених технік попередньої обробки вхідних даних з метою оптимізації продуктивності глибинних ненейронних мереж.

3. Дослідження та застосування методів контекстуального вкладання слів (contextual embeddings) для тонкого налаштування (fine-tuning) предметно-орієнтованих моделей з метою підвищення ефективності розпізнавання іменованих сутностей у різних контекстах.

4. Адаптація моделей нейронних мереж до конкретних доменів з урахуванням унікальних особливостей і термінології цих областей з предметно-орієнтованих баз знань.

5. Проведення комплексного порівняльного аналізу існуючих систем для демонстрації ефективності запропонованих методів.

Практична значення одержаних результатів цієї роботи зводиться до розробки та створення програмної компоненти на основі архітектури глибинних

нейронних мереж в рамках задачі розпізнавання іменованих сутностей. Розроблена NER-система характеризується високою точністю та ефективністю обробки великих обсягів текстових даних, що робить її придатною для використання в різних предметних областях, мінімізуючи потребу в ручному втручанні та зменшуючи ризик людських помилок. Система здатна до адаптації та інтеграції з різними типами баз даних, включаючи популярні СУБД, через API або спеціальні конектори. Завдяки можливості обробки даних у режимі реального часу та автоматизації процесів вилучення та заповнення бази даних, система забезпечує економію ресурсів та підвищує ефективність пошуку та аналізу інформації.

Апробація роботи. Основні положення і результати роботи були представлені та обговорювались на XVI науковій конференції магістрантів та аспірантів «Прикладна математика та комп'ютинг» ПМК-2023 (Київ, 28-23 листопада 2023 р.), а також у 8-ій міжнародній науково-практичній конференції BoScience, Бостон, США. 2023.

Структура та обсяг роботи. Магістерська дисертація складається зі вступу, чотирьох розділів, висновків по кожному розділу та повноцінних висновків загальної роботи, списку використаних літературних джерел (42 найменувань). Повний обсяг дисертації – 204 сторінок, у тому числі 102 сторінок основного тексту, 25 рисунків, 6 таблиць

У вступі подано загальну характеристику роботи, зроблено оцінку сучасного стану проблеми, обґрунтовано актуальність напрямку досліджень, сформульовано мету і задачі досліджень, показано наукову новизну отриманих результатів і практичну цінність роботи, наведено відомості про апробацію результатів і їхнє впровадження.

У першому розділі розглянуто історичний розвиток та сучасні тенденції у галузі розпізнавання іменованих сутностей. Детально розглянуто методи на основі правил, статистичні методи (включаючи приховану марковську модель, умовні випадкові поля та метод опорних векторів), методи на основі машинного навчання, а також застосування глибинних нейронних мереж для розпізнавання

іменованих сутностей. Розділ завершується оглядом сучасних проблем та перспектив розвитку області природної обробки мови у рамках задачі розпізнавання іменованих об'єктів та порівняльним аналізом методів розпізнавання іменованих сутностей.

У другому розділі детально розглянуто методи навчання та адаптації в системах розпізнавання іменованих сутностей. Висвітлено попередній аналіз та підготовку даних, перетворення даних у векторні представлення слів, а також огляд архітектур нейронних мереж для NER (включаючи згорткові нейронні мережі, рекурентні нейронні мережі та трансформаторні моделі). Розглянуто хід навчання та оцінку метрик ефективності моделей.

У третьому розділі розглянуто програмне середовище для розробки та вибір програмного забезпечення. Описано вибір та попередню обробку набору даних, включаючи формування та характеристику корпусу текстів, методи попередньої обробки даних та збір даних. Детально розглянуто проектування та архітектуру системи розпізнавання іменованих сутностей, інтеграцію контекстуальних представлень для подальшого використання у моделі машинного навчання, а також реалізацію архітектур моделей глибинних нейронних мереж. Розглянуто налаштування параметрів та оптимізацію роботи системи. Описано наукову новизну дослідження..

У четвертому розділі описано набори даних для проведення дослідження. Виконано тестування точності та швидкодії системи на реальних даних, а також порівняння запропонованої системи з існуючими методами. Експериментально підтверджено покращення моделей та їх доцільність.

У висновках представлені результати проведеної роботи, висновки щодо ефективності розробленої системи розпізнавання іменованих сутностей на основі глибинних нейронних мереж, а також рекомендації для подальших досліджень у цій галузі.

Ключові слова: розпізнавання іменованих сутностей, глибинні нейронні мережі, обробка природної мови, машинне навчання, аналіз тексту, вилучення сутностей, контекстно-залежні векторні представлення, вилучення інформації.

ABSTRACT

Relevance of the topic. The development of modern society is closely related to the automation of the information processing process, which is especially evident in the modern era of digital transformation and rapid advancement of information technologies, when the analysis of large volumes of text data becomes a key task for many fields of activity.

Particularly important is named entity recognition (NER) in text, namely the identification of persons, places, organizations, and other important categories, which finds application in many fields, including computer science, linguistics, bioinformatics, and others. In business, NER systems are used for market analysis, demand forecasting, resource and risk allocation. In the field of medicine, they help in patient identification, diagnosis and treatment. In law enforcement, NER systems are used to detect and predict crimes, and in science - to analyze scientific publications and identify hidden or non-obvious knowledge in works.

At the same time, with the development of deep learning and neural networks, new opportunities are opening up to improve the accuracy and performance of named entity recognition systems. The integration of deep neural networks in combination with natural language processing methods makes it possible to significantly increase the resource efficiency of data processing tasks into predefined classes due to more effective detection of complex dependencies and patterns in textual information. Therefore, the development of systems for recognizing named entities in text using deep neural networks remains an urgent task, since this technology can significantly optimize and speed up work with large arrays of text data.

The purpose of the work is to analyze and develop more effective models for recognizing named entities based on natural language processing classifiers and deep neural networks to improve the accuracy and speed of the system. The research is aimed at comparing different recognition approaches to determine optimal architectures and optimization methods to achieve maximum accuracy and system speed.

The object of the research is the named entity recognition (NER) system in the text, which is based on the use of natural language processing (NLP) and deep neural networks (DNN).

The subject of the research is the development of algorithms and software for the recognition system of named entities in the text based on deep neural networks (DNN).

Research methods consist of general scientific and special methods of cognition, in particular:

1. Analytical method: for the comprehensive study and generalization of the latest scientific developments of domestic and foreign scientists in the field of named object recognition (NER) and the development of scalable and subject-oriented NER systems.

2. Method of analysis and synthesis: to clarify and specify the conceptual and categorical apparatus in the field of NLP, to systematize knowledge about algorithms and software for NER, as well as to formulate theoretical propositions and research conclusions.

3. Method of analogy and comparison: to identify and compare the features of existing methods and approaches to NER, as well as to justify the selection and development of improved algorithms and software.

4. Systematization: to reveal the main trends, patterns and principles of development of NER systems, as well as to form classifications and typologies of algorithms and software.

5. Abstract-logical method: for the formulation of hypotheses, theoretical propositions and research conclusions, as well as for the justification of the proposed latest approaches to the development of scalable and subject-oriented NER systems.

6. The method of deduction and modeling: to develop proposals for improving NER mechanisms, developing new algorithms and software, as well as for evaluating their effectiveness.

7. Experimental methods: to conduct research to evaluate the effectiveness of the developed algorithms and NER software on test data sets.

8. Methods of statistical analysis: for processing the results of experiments, identifying statistical regularities and formulating conclusions about the effectiveness of the proposed methods and approaches.

The scientific novelty of the obtained results is as follows:

1. Presentation of improved architectures and hybrid models combining different types of neural networks, additional layers and data processing mechanisms to achieve higher performance in NER tasks compared to existing deep learning models.

2. Implementation and evaluation of the effectiveness of advanced techniques of preprocessing input data in order to optimize the performance of deep neural networks.

3. Research and application of methods of contextual embeddings for fine-tuning object-oriented models in order to improve the recognition efficiency of named entities in different contexts.

4. Adaptation of neural network models to specific domains, taking into account the unique features and terminology of these areas, potentially including subject-oriented knowledge bases.

5. Development and evaluation of new methods of data augmentation (data augmentation) to increase the robustness and generalization ability of models, especially in scenarios with a low resource of training data. In addition, research on the possibilities of using transfer learning to transfer knowledge from related tasks or domains in order to improve the performance of NER systems.

6. Carrying out a comprehensive comparative analysis of existing systems to demonstrate the effectiveness of the proposed methods.

The practical significance of the obtained results of this work is reduced to the development and creation of a software component based on the architecture of deep neural networks within the framework of the task of recognizing named entities. The developed NER system is characterized by high accuracy and efficiency of processing large volumes of textual data, which makes it suitable for use in various subject areas, minimizing the need for manual intervention and reducing the risk of human errors. The system is capable of adaptation and integration with various types of databases, including popular DBMS, through APIs or special connectors. Due to the possibility

of real-time data processing and automation of database extraction and filling processes, the system saves resources and increases the efficiency of information search and analysis.

Approbation of work. The main provisions and results of the work were presented and discussed at the XVI Scientific Conference of Master's and Postgraduate Students "Applied Mathematics and Computing" PMK-2023 (Kyiv, November 28-23, 2023). Also, in a scientific article, on the pages of the journal "Visnyk KhPI National Technical University. Series: System analysis of management and information technologies".

Structure and scope of work. The master's thesis consists of an introduction, four chapters, conclusions on each chapter and full conclusions of the general work, a list of used literary sources (42 names). The full volume of the dissertation is 204 pages, including 102 pages of the main text, 25 figures, 6 tables

The introduction provides a general description of the work, assesses the current state of the problem, substantiates the relevance of the research direction, formulates the purpose and tasks of the research, shows the scientific novelty of the obtained results and the practical value of the work, provides information about the approbation of the results and their implementation.

The first chapter examines the historical development and modern trends in the field of recognition of named entities. Rule-based methods, statistical methods (including hidden Markov model, conditional random fields, and support vector method), machine learning-based methods, and the application of deep neural networks for named entity recognition are discussed in detail. The chapter concludes with an overview of modern problems and prospects for the development of natural language processing within the framework of the task of recognizing named objects and a comparative analysis of methods for recognizing named entities.

In the second chapter, the methods of training and adaptation in systems of recognition of named entities are considered in detail. Preliminary data analysis and preparation, transformation of data into vector representations of words, and an overview of neural network architectures for NER (including convolutional neural

networks, recurrent neural networks, and transformer models) are covered. The course of training and assessment of model performance metrics are considered.

The third chapter discusses the software development environment and software selection. Data set selection and pre-processing are described, including corpus formation and characterization, data pre-processing methods, and data collection. The design and architecture of the named entity recognition system, the integration of contextual representations for further use in machine learning models, as well as the implementation of deep neural network model architectures are considered in detail. Parameter settings and optimization of system operation are considered. The scientific novelty of the study is described.

The fourth chapter describes the data sets for conducting the study. The accuracy and speed of the system were tested on real data, as well as a comparison of the proposed system with existing methods. The improvement of the models and their expediency have been confirmed experimentally.

The conclusions present the results of the work carried out, conclusions regarding the effectiveness of the developed named entity recognition system based on deep neural networks, as well as recommendations for further research in this field.

Keywords: Named entity recognition (NLP), deep neural networks (DNN), natural language processing (NLP), machine learning, text analysis, entity extraction, context-sensitive vector representations, information extraction