DOI: https://doi.org/10.15276/ict.01.2024.34
УДК 004.6

Розпізнавання іменованих сутностей та їхня роль при аналізі неструктурованих даних

Стасьо Олег Романович[1)]
Адьюнкт каф. Інформаційних технологій та систем електронних комунікацій
ORCID: https://orcid.org/0009-0005-6049-6161; staso.oleh@gmail.com
Бурак Назарій Євгенович[1)]
Канд. техніч. наук, доцент каф. Інформаційних технологій та систем електронних комунікацій
ORCID: https://orcid.org/0000-0002-3880-4077; n.burak@ldubgd.edu.ua. Scopus Author ID: 57204558265
[1)] Львівський державний університет безпеки життєдіяльності, вул. Клепарівська, 35. Львів, 79007, Україна

АНОТАЦІЯ

У сучасному цифровому світі, де величезні обсяги неструктурованих даних генеруються щодня, здатність ефективно обробляти цю інформацію є ключовою для багатьох галузей. Неструктуровані дані, які включають текстові файли, електронні листи, відео, аудіо, зображення та інші форми медіа, становлять основну частину цифрових даних і вимагають спеціалізованих інструментів для їх аналізу. Обробка природної мови та розпізнавання іменованих сутностей є двома ключовими технологіями, які дозволяють перетворювати неструктуровані дані в структуровану інформацію, що може бути використана для різноманітних застосувань.

Обробка природної мови дозволяє машинам розуміти, інтерпретувати, маніпулювати та генерувати людську мову, відкриваючи можливості для глибокого аналізу текстових даних. Це включає виявлення ключових слів, фраз, тем, а також емоційних нюансів у текстах. Зозпізнавання іменованих сутностей, як важлива складова обробка природної мови, спеціалізується на ідентифікації та класифікації іменованих сутностей у тексті на певні категорії, такі як імена осіб, організацій, локацій, дати, час та інші. Це дозволяє автоматизувати процеси сортування, категоризації та аналізу інформації.

Проте, робота з обробка природної мови та стикається з низкою викликів. Великий обсяг і різноманітність даних ускладнюють їх збір, зберігання та аналіз. Відсутність стандартизації може призвести до проблем з сумісністю та інтеграцією різних джерел даних. Крім того, існують виклики, пов'язані з розпізнаванням іменованих сутностей, зокрема, розрізненням між однаковими іменами, які належать до різних осіб, та розумінням контексту, в якому використовуються імена. Незважаючи на ці виклики, перспективи Обробка природної мови та зозпізнавання іменованих сутностей виглядають оптимістично, з огляду на постійні інновації в галузі штучного інтелекту та машинного навчання, які обіцяють покращення точності та ефективності цих технологій у майбутньому.

Ключові слова: наука про дані; неструктуровані дані; аналіз даних; добування інформації; Data mining; обробка природної мови; розпізнавання іменованих сутностей; розпізнавання іменованих сутностей

Актуальність даної наукової роботи полягає в тому, що вона розглядає важливість і виклики обробки неструктурованих даних у сучасному цифровому світі, де ці дані становлять значну частину всієї інформації, що генерується. Враховуючи, що неструктуровані дані включають різноманітні формати, такі як текст, відео, аудіо та інші, їх аналіз представляє значні технічні виклики, але водночас відкриває величезні можливості для отримання цінних інсайтів.

Особлива увага в роботі приділяється технологіям обробки природної мови (NLP) та розпізнаванню іменованих сутностей (NER), які є ключовими інструментами для ефективної обробки текстових неструктурованих даних. Застосування NLP та NER може значно покращити процеси виявлення, класифікації та аналізу інформації, що сприяє кращому управлінню даними, прийняттю обґрунтованих рішень та розробці нових інноваційних продуктів і послуг.

Проте, існуючі виклики, такі як великий обсяг даних, їх різноманітність, відсутність стандартизації та складнощі з розпізнаванням іменованих сутностей, потребують подальших досліджень та розробки більш ефективних методів і технологій. Вивчення та вдосконалення NLP та NER відіграють критичну роль у подоланні цих викликів, що робить тему дуже актуальною для наукових досліджень і практичного застосування в різних галузях.

Метою дослідження є детальний аналіз ролі та перспектив технологій обробки природної мови (NLP) і розпізнавання іменованих сутностей (NER) у контексті обробки неструктурованих даних. Дослідження зосереджується на вивченні того, як ці технології можуть сприяти ефективному виявленню, класифікації та аналізу інформації, що відкриває нові можливості для глибокого розуміння та використання великих обсягів неструктурованих даних у різних галузях. Завданнями аналізу є визначення ключових викликів, з якими стикаються NLP та NER, розробка стратегій для подолання цих викликів, а також оцінка потенційних шляхів покращення точності та ефективності цих технологій. Особлива увага приділяється викликам та потенційним шляхам подолання обмежень існуючих методів NLP та NER для підвищення їхньої ефективності та точності.

На даний час світовим співтовариством вже усвідомлений головний напрямок у боротьбі з інформаційним вибухом – перехід від збереження й обробки даних до накопичення й обробки знань. Тому виникає потреба у засобах та методах здобуття знань з тих даних, що генеруються в процесі діяльності людства та можуть бути корисними для подальшого використання. І тут виникає проблема в аналізі цих даних, тому що за інформацією експертів більше 85 % даних зберігається в неструктурованій формі [1].

Неструктуровані дані відносяться до інформації, яка не має попередньо визначеної моделі або не організована у вигляді традиційних баз даних. Ці дані можуть включати текстові файли, електронні листи, відео, аудіо, зображення, веб-сторінки та інші форми медіа [3]. Відсутність чіткої структури ускладнює їх зберігання, обробку та аналіз за допомогою стандартних інструментів і методів.

В сучасному світі неструктуровані дані становлять більшу частину всіх цифрових даних, що генеруються в різних галузях, включаючи охорону здоров'я, фінанси, медіа та розваги. В охороні здоров'я, наприклад, неструктуровані медичні записи та зображення можуть бути аналізовані для виявлення тенденцій та покращення діагностики. Важливість обробки неструктурованих даних полягає в можливості отримання глибоких інсайтів, які можуть сприяти прийняттю обґрунтованих рішень та інноваційним розробкам.

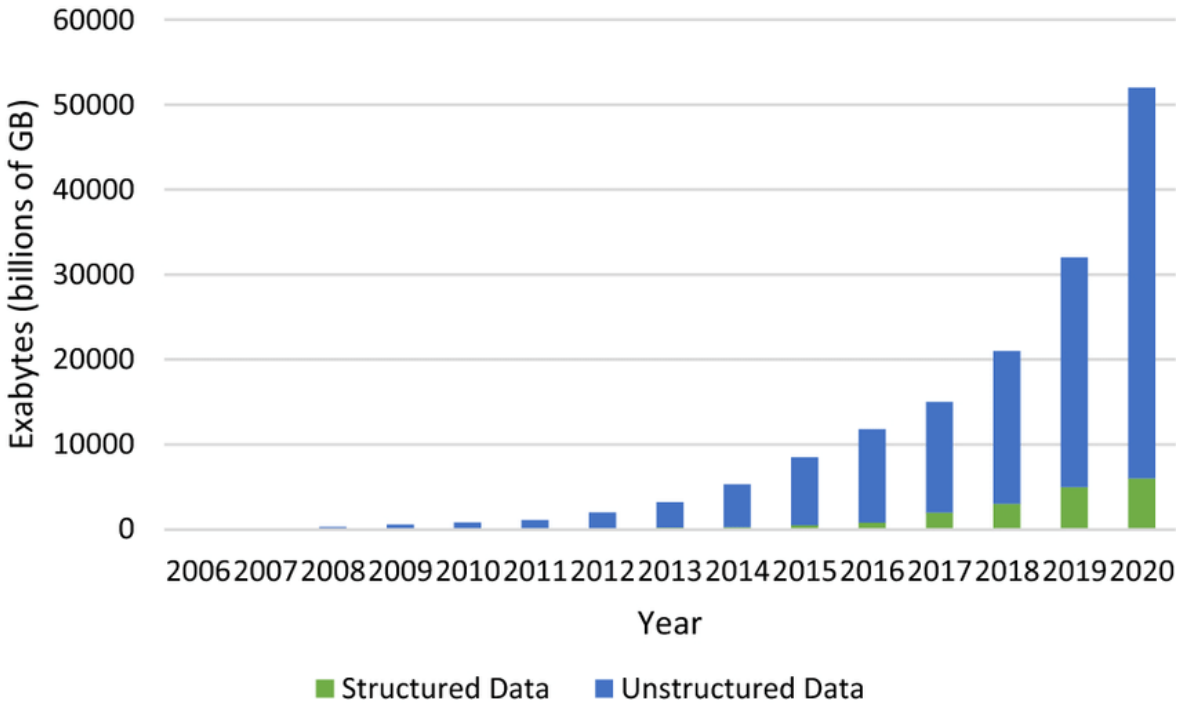

Рис. 1. **Ріст кількості неструктурованих даних відносно структурованих [2]**

Однак робота з неструктурованими даними стикається з низкою проблем і викликів. Однією з основних проблем є великий обсяг і різноманітність даних, що ускладнює їх збір,

зберігання та аналіз. Крім того, відсутність стандартизації може призвести до проблем з сумісністю та інтеграцією різних джерел даних. Також існує виклик забезпечення конфіденційності та безпеки цих даних, особливо коли вони містять чутливу інформацію [4].

Одним з поширених способів опрацювання неструктурованої інформації є обробка природної мови. Обробка природної мови (NLP) є галуззю штучного інтелекту, яка зосереджена на взаємодії між комп'ютерами та людською мовою [5]. Вона дозволяє машинам розуміти, інтерпретувати, маніпулювати та генерувати людську мову, що робить її особливо корисною для роботи з неструктурованими даними. Завдяки NLP комп'ютери можуть аналізувати великі обсяги текстових даних, виявляючи ключові слова, фрази, теми та навіть емоційні нюанси, що відкриває нові можливості для глибокого аналізу інформації [6].

Основні підходи в NLP включають статистичні методи, машинне навчання та глибоке навчання. Ці методи дозволяють вирішувати різноманітні задачі, такі як семантичний аналіз, розпізнавання мови, переклад, автоматичне резюмування та генерація тексту. Наприклад, семантичний аналіз може допомогти визначити настрій тексту, в той час як розпізнавання мови дозволяє системам взаємодіяти з користувачами за допомогою голосових команд.

Однак, попри значний прогрес у галузі NLP, існують складнощі, зокрема у розпізнаванні іменованих сутностей та власних назв. Ці виклики включають розрізнення між однаковими іменами, які належать до різних осіб, а також розуміння контексту, в якому використовуються імена. Крім того, існує проблема розпізнавання імен, які можуть мати різні написання в різних культурах або мовах. Ці виклики вимагають розробки більш складних алгоритмів та використання більш обширних наборів даних для тренування систем.

Для вирішення проблем з розпізнаванням імен і власних назв при обробці природної мови застосовують розпізнавання іменованих сутностей. Зозпізнавання іменованих сутностей (NER), або розпізнавання іменованих сутностей, є ключовим інструментом у галузі обробки природної мови, який спрямований на ідентифікацію та класифікацію іменованих сутностей у тексті на певні категорії, такі як імена осіб, організацій, локацій, дат, часу, кількостей тощо [7]. Цей процес допомагає у структуруванні неструктурованих даних, забезпечуючи можливість глибшого аналізу та розуміння контенту.

Зозпізнавання іменованих сутностей (NER), виконує кілька основних задач, які включають виявлення іменованих сутностей та їх класифікацію за визначеними категоріями. Наприклад, у фразі «Джордж Вашингтон був першим президентом США» NER ідентифікує «Джордж Вашингтон» як ім'я особи та «США» як локацію. Це дозволяє системам зберігати та обробляти інформацію більш ефективно, використовуючи її для різних застосувань, таких як автоматичне резюмування текстів, пошук інформації та інші.

Принцип роботи NER полягає у використанні алгоритмів машинного навчання або глибокого навчання для аналізу тексту та виявлення зазначених сутностей. Спочатку система NER навчається на великих обсягах анотованих текстових даних, де іменовані сутності вже позначені. Це дозволяє моделі вивчити контекстуальні шаблони та лінгвістичні особливості, які характеризують різні категорії сутностей. Після тренування модель може застосовувати набуті знання для ідентифікації та класифікації сутностей у нових, нерозмічених текстах. Таким чином, NER сприяє автоматизації обробки текстових даних, підвищуючи ефективність та точність аналітичних додатків.

У сфері розпізнавання іменованих сутностей (NER) існує кілька популярних моделей, які використовуються для аналізу та класифікації текстових даних. Однією з таких моделей є CRF (Conditional Random Fields), яка є статистичним методом для передбачення послідовностей міток, заснованим на контексті. Інша важлива модель - це LSTM (Long Short-Term Memory), тип рекурентної нейронної мережі, який ефективно обробляє послідовності

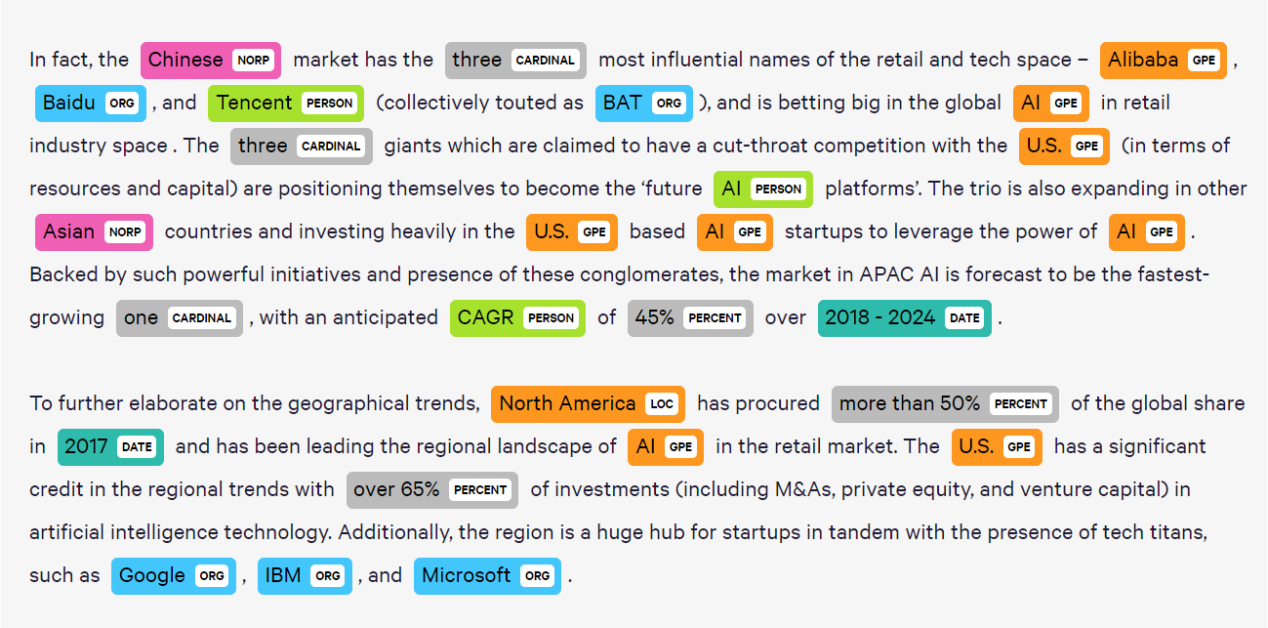

Рис. 2. **Зразок тексту з тегами іменованих сутностей [8]**

даних з врахуванням довгострокових залежностей. Нещодавно велику популярність набули моделі на основі трансформерів, такі як BERT (Bidirectional Encoder Representations from Transformers) та її модифікації (наприклад, RoBERTa, ALBERT), які використовують механізми уваги для кращого розуміння контексту слова в тексті. Ці моделі демонструють високу точність у виявленні іменованих сутностей завдяки своїй здатності аналізувати великі обсяги даних та вивчати складні мовні патерни. Використання цих передових технологій відкриває нові можливості для покращення процесів NER і розширення їх застосування в різних областях.

Ефективне розпізнавання іменованих сутностей в неструктурованих даних вимагає використання комплексних підходів, які можуть адаптуватися до різноманітності та складності мовних структур. Однією з базових рис таких підходів є здатність до глибокого семантичного аналізу, що дозволяє системі розуміти контекст і відтінки значення слів у великих текстових масивах. Інша важлива риса - це використання навчальних даних великого обсягу для тренування моделей, що забезпечує високу точність і адаптивність системи до нових даних та сценаріїв використання. Також критично важливим є впровадження механізмів машинного навчання, які можуть ефективно обробляти великі обсяги даних в реальному часі, що є ключовим для застосувань, де швидкість реакції є критичною. Використання моделей, заснованих на трансформерах, таких як BERT, які вже показали свою ефективність у розумінні мовних нюансів, також є перспективним напрямком. Нарешті, інтеграція з іншими NLP інструментами, такими як синтаксичний аналіз та сентимент-аналіз, може значно покращити здатність системи до всебічного аналізу тексту.

Сучасні проблеми і виклики при роботі з NER включають обмежену кількість якісних анотованих даних для тренування моделей, складність розпізнавання сутностей у неструктурованих або неформальних текстах, таких як соціальні медіа, а також виклики, пов'язані з розпізнаванням іронії та жартів, що можуть ввести систему в оману. Крім того, існує проблема з визначенням і розрізненням сутностей, які мають однакові імена але належать до різних категорій.

Незважаючи на виклики при роботі з іменованими сутностями, перспективи і майбутнє NER виглядають обнадійливо з огляду на постійні інновації в галузі штучного інтелекту та машинного навчання. Зокрема, розвиток технологій глибокого навчання та поява нових моделей, які можуть краще розуміти контекст та нюанси людської мови, обіцяють значне покращення точності та ефективності систем NER. Це, у свою чергу, може привести до

більш широкого застосування NER у різних областях, від автоматичного контент-аналізу до розуміння і взаємодії з користувачами в реальному часі.

Висновки. У сучасному світі обробки великих обсягів неструктурованих даних, важливість технологій, таких як обробка природної мови (NLP) та розпізнавання іменованих сутностей (NER), не може бути переоцінена. Обробка природної мови дозволяє машинам розуміти і аналізувати людську мову, перетворюючи неструктуровані дані в структуровану інформацію, що може бути використана для різноманітних застосувань. NER, як важлива складова NLP, спеціалізується на ідентифікації і класифікації іменованих сутностей, що значно підвищує можливості аналізу тексту.

Завдяки NER, системи можуть автоматично виявляти і категоризувати ключові елементи в текстах, такі як імена людей, організацій, географічні назви та інші специфічні дані. Це дозволяє не тільки ефективніше управляти інформацією, але й використовувати її для покращення прийняття рішень, автоматизації процесів і створення нових сервісів. Наприклад, в медіа і журналістиці, NER може допомогти автоматично сортувати статті за темами або ключовими фігурами, а в фінансовому секторі - аналізувати новини для прогнозування ринкових тенденцій.

Проте, попри значний прогрес у розвитку NER, існують виклики, які потребують подальших досліджень і вдосконалення. Складність розпізнавання іменованих сутностей в неформальних або неструктурованих текстах, а також необхідність розрізняти сутності з однаковими назвами в різних контекстах, залишаються ключовими проблемами. Однак, з розвитком технологій машинного навчання та глибокого навчання, перспективи NER виглядають оптимістично, обіцяючи ще більшу точність і ширше застосування в майбутньому.

СПИСОК ЛІТЕРАТУРИ

1. Захарчишин Н. Г., Захарчишин Н. Р., «Ріст структурованих та неструктурованих даних та управління ними: загальні аспекти». *Вчені записки ТНУ ім. В. І. Вернадського, Серія: Технічні науки*. 2021; 32 (71) (5): 83–87. DOI: https://doi.org/10.32838/2663-5941/2021.5/13.

2. Azad, P., Navimipour, N.J., Rahmani, A.M. et al. “The role of structured and unstructured data managing mechanisms in the Internet of things”. *Cluster Computing*. 2020; 23 (2): 1185–1198. DOI: https://doi.org/10.1007/s10586-019-02986-2.

3. “Unstructured data”. *Wikipedia*. – Available from: https://en.wikipedia.org/wiki/Unstructured_data.

4. “The challenges of analysing unstructured data”. *Selerity*. – Available from: https://seleritysas.com/blog/2019/08/27/the-challenges-of-analysing-unstructured-data/

5. Egger R., Gokce E. “Natural Language Processing (NLP): An Introduction: Making Sense of Textual Data”. 2022. p 307–334. DOI: https://doi.org/10.1007/978-3-030-88389-8_15.

6. Murugan, M. “Natural Language Processing (NLP)”. 2024. DOI: http://dx.doi.org/10.13140/RG.2.2.13534.04169.

7. Mohit, B., Zitouni, I. “Named Entity Recognition”. 2014. p. 221–245. DOI: http://dx.doi.org/10.1007/978-3-642-45358-8_7.

8. “SpaCy – named entity and dependency parsing visualizers”. – Available from: https://meenavyas.wordpress.com/2018/06/10/spacy-named-entity-and-dependency-parsing-visualizers.

DOI: https://doi.org/10.15276/ict.01.2024.34
UDC 004.6

Named entity recognition and its role in unstructured data analysis

Oleh R. Staso[1)]
Postgraduate student, Department of Information technologies and Electronic communication systems
ORCID: https://orcid.org/0009-0005-6049-6161; staso.oleh@gmail.com
Nazarii Ye. Burak[1)]
PhD, Associate Professor, Department of Information technologies and Electronic communication systems
ORCID: https://orcid.org/0000-0002-3880-4077; n.burak@ldubgd.edu.ua. Scopus Author ID: 57204558265
[1)] Lviv State University of Life Safety, 35, Kleparivska Str. Lviv, 79007, Ukraine

ABSTRACT

In today's digital world, where vast amounts of unstructured data are generated every day, the ability to efficiently process this information is key for many industries. Unstructured data, which includes text files, emails, video, audio, images, and other forms of media, is the bulk of digital data and requires specialized tools to analyze it. Natural Language Processing (NLP) and Named Entity Recognition (NER) are two key technologies that enable the transformation of unstructured data into structured information that can be used for a variety of applications.

Natural Language Processing enables machines to understand, interpret, manipulate and generate human language, opening up possibilities for deep analysis of textual data. This includes identifying key words, phrases, themes, and emotional nuances in texts. NER, as an important component of Natural Language Processing, specializes in identifying and classifying named entities in the text into certain categories, such as names of persons, organizations, locations, dates, times, and others. This allows you to automate the processes of sorting, categorizing and analyzing information.

However, working with Natural Language Processing and Named Entity Recognition faces a number of challenges. The large volume and variety of data make it difficult to collect, store and analyze it. Lack of standardization can lead to problems with interoperability and integration of different data sources. In addition, there are challenges related to the recognition of named entities, in particular, distinguishing between the same names belonging to different persons and understanding the context in which the names are used. Despite these challenges, the outlook for Natural Language Processing and Named Entity Recognition looks bright, with continued innovations in artificial intelligence and machine learning promising to improve the accuracy and efficiency of these technologies in the future.

Keywords: Data science; unstructured data; data analysis; data mining; data analysis; Natural Language Processing; Named Entity Recognition