УДК 004.85:004.934

CHALLENGES IN NAMED ENTITY RECOGNITION

Serhiienko D.V., Ryabova N.V.
e-mail: dmytro.serhiienko1@nure.ua, nataliya.ryabova@nure.ua
National University of Radio Electronics, AI department
Kharkiv, Ukraine

This work explores the challenges and advancements in Named Entity Recognition (NER) within natural language processing. The study provides a comprehensive review of the key difficulties faced in NER. Additionally, it examines contemporary methods and models designed to address these challenges, highlighting recent innovations in deep learning and transformer-based architectures. The paper also investigates the impact of linguistic diversity, noisy data, and evolving language on NER performance, offering insights into optimization strategies and future research directions in entity recognition.

NER is one of the key tasks in the field of Natural Language Processing (NLP). Its main goal is the automatic identification and classification of entities in text, such as person names, organizations, locations, dates, and other categories [1]. This technology is widely used in search engines, text analytics, machine translation, chatbots, and many other applications.

Let’s consider the main challenges for NER and their specific features:

1) ambiguity. Example of query: «Apple». The word «Apple» can be interpreted as either the company (Apple Inc.) or the fruit. The NER model needs to correctly determine whether the user refers to the Apple product or the fruit;

2) domain adaptation. Example of query: «accounting software». Models trained on general texts may struggle to recognize terms specific to accounting software, such as «BAS» or «QuickBooks»;

3) data scarcity. Example of query: «best electric scooters for Ukraine». For low-resource languages like Ukrainian, there may be limited annotated datasets, making it harder for models to recognize local brands or specific terms;

4) entity variability. Example of query: «iPhone 13 pro max» vs. «iPhone 13 Pro Max». Product names can have different forms, abbreviations, or spelling errors, making it difficult to accurately recognize the same product;

5) complex entity types. Example of query: «L'Oréal Paris men's shampoo 400 ml». The model needs to correctly identify multiple entities: the brand (L'Oréal Paris), product type (shampoo), category (men's), and volume (400 ml);

6) multilinguality. Example of query: «best smartphones for gaming» vs. «кращі смартфони для ігор». Handling queries in different languages (English vs. Ukrainian) requires additional training and adaptation of models to different language structures and terms;

7) out-of-vocabulary (OOV) Words. Example of query: «Xiaomi Mi 13». New or rare products that were not included in the training data can be challenging to recognize, especially if the product name or brand is new;

8) noisy data. Example of query: «samsung galaxy s20+ on sale!!!». Noisy data, such as informal language, capitalization, or exclamation marks, can interfere with accurate entity recognition;

9) evolving language. Example of query: «best gadgets of 2025». New brands, products, or technologies, such as upcoming gadget models, may emerge quickly, and the NER model needs to be updated to detect these new entities.

These are the main types of problems one may encounter at present. All of them need to be addressed, as solving them will help manage resources more effectively and meet the requirements of relevant business processes [2].

For effectively addressing the challenges related to NER, it is necessary to have high-quality and diverse training data that covers different domains, languages, and variations in entity names. Clearly, things are a bit simpler when the data is high-quality and abundant, but this is not always the case, so alternative approaches need to be sought for working with limited or noisy data, but those are different problems [3]. Models must be adapted to each specific context, such as medical, legal, or commercial text.

The use of contextual language models like BERT or GPT can significantly improve entity recognition by considering the surrounding context. It is also important to develop methods for handling noisy data, as informal language or low-quality texts, such as those from social media, can complicate the process.

To effectively work with texts in different languages, multilingual models need to be developed that can adapt to various lexical and grammatical structures. Additionally, using data augmentation techniques will allow the creation of new training datasets, which helps address the resource deficit, especially for low-resource languages. Moreover, for recognizing rare or new entities, active learning methods and regular database updates should be implemented.

This work is devoted to solving these challenges with deep learning methods. It will significantly improve the accuracy and effectiveness of NER systems, making them more reliable and adaptable for various business processes and applications. Continuous advancements and adaptation to new data and contexts are essential for maximizing their potential.

List of sources used:

1. O. Shatalov, N. Ryabova, "Named Entity Recognition Problem for Long Entities in English Texts," 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), 2021, Volume 1, pp. 76-79, DOI: 10.1109/CSIT52700.2021.9648768 (дата звернення: 01.03.2025).

2. Named Entity Recognition: A Practical Guide https://labelyourdata.com/articles/named-entity-recognition (дата звернення: 01.03.2025).

3. Exploring named-entity recognition techniques for academic books https://onlinelibrary.wiley.com/doi/epdf/10.1002/leap.1610 (дата звернення: 01.03.2025).