ДОСЛІДЖЕННЯ МЕТОДІВ РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ У НЕСТРУКТУРОВАНОМУ ТЕКСТІ

Люліна К. П., Турута О. П.
Харківський національний університет радіоелектроніки, Харків, Україна

Виникнення технології розпізнавання іменованих сутностей пов'язане з проблематикою автоматичної обробки текстів та їх розуміння у подальшому за допомогою програмних систем. Дане питання було вперше розглянуто в рамках шостої Конференції з розуміння повідомлень (Sixth Message Understanding Conference − MUC-6) у 1995 році [1]. На цій конференції вперше було сформульовано визначення іменованої сутності, що полягало у наступному: іменована сутність - це вставлений у текст тег стандатної узагальненої мови розмітки задля маркування слів.

Методи NER можуть базуватися окремо або одночасно на основі лексики (lexicon-based), правил (rule-based) та машинного навчання (machine learning based).

Метою доповіді є порівняльний аналіз існуючих підходів до проблеми розпізнавання іменованих сутностей, виокремлення суттєвих переваг та недоліків, а також побудова на основі позитивного та негативного досвіду структури системи розпізнавання іменованих сутностей.

В доповіді наводиться аналітичний огляд досліджуваної проблеми.

Було розглянуто застосування кожного типу методів NER на приладі трьох патентів.

Слід зазначити, що найбільш ефективними є методи, базовані на правилах, так як мають змогу визначати іменовані сутності за завчасно визначеними патернами [2]. Їх ефективність обумовлена тим, що природні мови мають достатньо завчасно визначених правил та форматів, за якими можна розпізнавати текст.

Проте методів, базованих на правилах недостатньо для аналізу текстів природною мовою, що у свою чергу обумовлено наявністю контексту в реченнях, багатозначністю слів, а також наявністю зворотів та словосполучень, що мають відхилення від загальноприйнятих правил правопису. Саме тому на практиці алгоритми NER є комбінацією різних методів розпізнавання іменованих сутностей, що працюють у комплексі.

Список літератури

1. Overview of results of the MUC-6 evaluation : веб-сайт. URL: https://aclanthology.org/X96-1048.pdf.

2. Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications: веб-сайт. URL: https://www.sciencedirect.com/topics/computer-science/named-entity-recognition-system#:~:text=7.6%20Named%20Entity%20Recognition&text=There%20are%20three%20major%20approaches,based%2C%20and%20machine%20learning%20based.