МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Харківський національний університет радіоелектроніки
Факультет Комп'ютерних наук
Кафедра Програмної інженерії

КВАЛІФАЦІКАЦІЙНА РОБОТА

Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

Дослідження методів розпізнавання іменованих сутностей в неструктурованому тексті

Виконала: студентка	2 курсу групи ІПЗм-20-2
Люліна К.П. (прізвище, ініціали)	
Спеціальність	121 – Інженерія програмного забезпечення
Тип програми	Освітньо-наукова
Керівник	доц. Турута О.П. (посада, прізвище, ініціали)

Допускається до захисту

Зав. Кафедри _____ З.В. Дудар

2022

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Програмної інженерії

Рівень вищої освіти другий (магістреський)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва)

Тип програми освітньо-наукова програма

Освітня програма Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:

Зав. кафедри _____

(підпис)

«__»_____202_ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентки Люліній Катерині Петрівні
(прізвище ім'я по батькові студента)

1. Тема роботи «Дослідження методів розпізнавання іменованих сутностей в структурованому тексті»

затверджена наказом університету від «__»_____202_ р. №___

2. Термін подання студентом роботи до екзаменаційної комісії «__»_____202_ р.

3. Вихідні дані до роботи електронні ресурси за обраною тематикою, алгоритми машинного навчання для розпізнавання іменованих сутностей, методичні вказівки до виконання кваліфікаційної роботи магістра, пояснювальна записка.

4. Перелік питань, що потрібно опрацювати в роботі аналіз предметної області, постановка задачі, огляд існуючих методів NER, опис розробленої системи NER.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Огляд наукової та патентної літератури	15.02.2022	виконано
2	Постановка задачі	20.02.2022	виконано
3	Розробка моделі NER	04.03.2022	виконано
4	Збір тестових даних	10.03.2022	виконано
5	Проектування та реалізація програмної системи	01.04.2022	виконано
6	Розробка плану проведення експериментальних досліджень	06.04.2022	виконано
7	Проведення експериментальних досліджень	12.04.2022	виконано
8	Аналіз отриманих результатів	16.04.2022	виконано
9	Підготовка пояснювальної записки	06.05.2022	виконано
10	Підготовка презентації та доповіді	15.05.2022	виконано
11	Перевірка на плагіат	15.05.2022	виконано
12	Архівування	20.05.2022	виконано
13	Нормконтроль та рецензування	20.05.2022	виконано
14	Попередній захист	22.05.2022	виконано
15	Допуск до захисту у зав. кафедри	22.05.2022	виконано

Дата видачі завдання 17 січня 2022 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Турута О. П.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Кваліфікаційна робота магістра містить: 65 с., 26 рис., 5 табл., 26 джер.

ІМЕНОВАНІ СУТНОСТІ, РОЗПІЗНАВАННЯ, НЕСТРУКТУРОВАНИЙ ТЕКСТ, ПОШУК ЛОКАЦІЙ

Об’єктом дослідження є розпізнавання іменованих сутностей.

Предметом дослідження є методи розпізнавання іменованих сутностей у неструктурованому тексті.

Метою роботи є дослідження та аналіз методів розпізнавання іменованих сутностей у неструктурованому тексті та їх подальша реалізація.

Методи дослідження, що були використані: аналіз та синтез, методи порівняння, математичні методи.

Результатом роботи є реалізація методу розпізнавання іменованих сутностей локацій та місцезнаходження у неструктурованому тексті.

NAMED ENTITIES, RECOGNITION, UNSTRUCTURED TEXT, LOCATION SEARCH

The object of research is the recognition of named entities.

The subject of research is the methods of recognizing named entities in an unstructured text.

The aim of the work is to study and analyze the methods of recognizing named entities in an unstructured text and their further implementation.

Research methods used: analysis and synthesis, comparison methods, mathematical methods.

The result of the work is the implementation of the method of recognizing named entities of locations and locations in unstructured text.

Я, Люліна Катерина Петрівна ,
(прізвище, ім’я, по-батькові)

студент(ка) групи ІПЗм-20-2 здобувач вищої освіти на другому (магістерському) рівні

кафедра програмної інженерії ,
(повна назва кафедри)

заявляю: моя кваліфікаційна робота на тему
Дослідження методів розпізнавання іменованих сутностей в структурованому тексті,

що буде представлена до ЕК для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений(а) з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

CRF – Conditional Random Fields.

DL – Deep Learning.

LSTM – Long Short Term Memory.

NER – Named Entity Recognition, розпізнавання іменованих сутностей.

NLP – Natural Language Processing, обробка природної мови.

ML – Machine Learning.

POS – Part of speech.

RE – Relative extraction.

UD – Universal Dependencies.

ВСТУП

З розвитком інформаційних технологій та штучного інтелекту усе більш актуальним постає питання створення систем для аналізу та обробки природних мов. Обробка текстів природних мов (Natural Language Processing, NLP) є одним із напрямків штучного інтелекту, що використовується для аналізу та синтезу текстів на природних мовах. Однією із задач NLP є розпізнавання іменованих сутностей (Named Entity Recognition, NER) [1]. NER використовується для семантичної розмітки тексту природною мовою задля виявлення іменованих сутностей різних класів [2].

Актуальність дослідження методів NER полягає у тому, що достатня частина соціальної взаємодії переходить у площину цифрових технологій, тому виникає потреба у автоматизації обробки природніх мов в рамках існуючих систем, де передбачена взаємодія з користувачами за допомогою повідомлень у текстовому вигляді. До таких систем належать служби підтримки користувачів, системи доставки товарів та послуг, програмні системи для управління екстреними ситуаціями, аналіз публікацій та новин.

Метою дослідження є синтез методу розпізнавання іменованих сутностей у неструктурованому тексту на прикладі виявлення сутностей локацій та місць.

Задачами дослідження є наступні:

– аналіз існуючих методів розпізнавання іменованих сутностей у неструктурованому тексті;

– розробка методу NER для розпізнавання локацій, позицій та місцезнаходжень у неструктурованому тексті на прикладі української мови;

– порівняння ефективності існуючих методів та розробленого.

Об’єктом дослідження є розпізнавання іменованих сутностей.

Предметом дослідження є методи розпізнавання іменованих сутностей у неструктурованому тексті.

До методів дослідження, використаних у рамках даної роботи, належать аналіз та синтез, методи порівняння, математичні методи. Метод аналізу та синтезу необхідний для аналізу властивостей та особливостей існуючих методів, а також розробки методу розпізнавання іменованих сутностей у неструктурованому тексті. Методи порівняння необхідні для виявлення найбільш ефективних методів NER.

Наукова новизна одержаних результатів полягає в адаптації методів розпізнавання іменованих сутностей локацій, місцеположень та місці у неструктурованому тексті до української мови.

Практично результати досліджень можуть бути застосовані для впровадження у програмні системи обробки повідомлень у сфері доставки товарів та послуг, а також систем реагування на екстрені ситуації задля розпізнавання корисної інформації про локації та місця з заданих повідомлень користувачів.

Наукова стаття за даною тематикою була опублікована у збірнику «Сучасні напрями розвитку інформаційно-коммунікаційних технологій та засобів управління. Тези доповідей дванадцатої міжнародної науково-технічної конференції 27 – 28 квітня 2022 року».

1 ОГЛЯД НАУКОВОЇ ТА ПАТЕНТНОЇ ЛІТЕРАТУРИ

1.1 Загальні відомості

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. На цій конференції вперше було сформульовано визначення іменованої сутності, що полягало у наступному: іменована сутність - це вставлений у текст тег стандартної узагальненої мови розмітки задля маркування слів. Були визначені наступні категорії тегів:

– ENAMEX – група тегів, що використовуються для позначення слів, що відображають імена людей, організацій або локацій;

– TIMEX – група тегів, що використовуються для позначення слів, що відображають дату або час;

– NUMEX – група тегів, що використовуються для позначення слів, що відображають валюту або процент [3].

З тих часів список тегів розширився і наразі може включати назви подій, мов, витворів мистецтва, юридичних термінів та законів тощо.

Розпізнавання іменованих сутностей (NER) використовується для вирішення наступних задач:

– для ідентифікації та класифікації слів у неструктурованому тексті за категоріями;

– для автоматичного визначення того, до якої частини мови слово відноситься та які граматичні ознаки має (POS tagging);

– для виявлення відношень та зв’язків між словами (RE – relationship extraction).

Методи NER можуть базуватися окремо або одночасно на основі лексики (lexicon-based), правил (rule-based) та машинного навчання (machine learning based) [4].

Методи NER, що базуються на лексиці, можуть бути використані в умовах, коли немає у наявності корпусу. Ці моделі поєднують у собі результати морфологічного аналізу, набір лексиконів та методи стеммінгу та лематизації. Суть стемінгу полягає у нормалізації вихідного слова та виділенні з нього суттєвої складової, виключаючи граматичну складову (інформацію про рід, відмінок, число, час тощо). Лематизація дозволяє привести слово до його канонічної форми (лемми) [5]. При приведенні вихідного слова до його інфінітиву можливе використання відомих словників слів завчасно класифікованих за певними категоріями іменованих сутностей [6].

Методи NER, що базуються на правилах, можуть бути використані для тих сутностей, що мають певну структуру, яка може будуватися за завчасно визначеними патернами та правилами. Ці методи використовують регулярні вирази для пошуку іменованих сутностей. З їх допомогою можуть бути розпізнані такі сутності як номери телефонів, адреси електронних поштових скриньок, персональні дані (ідентифікаційні номери, номери банківських карток та рахунків), дати та інші дані, що мають завчасно визначений формат [6].

NER є задачею класифікації, тому для її вирішення можна застосовувати алгоритми статистичного машинного навчання. Існують наступні підходи машинного навчання для NER:

- прихована модель Маркова (Hidden Markov Model);
- умовні випадкові поля (CRF);
- машини опорних векторів (Support Vector Machines);
- моделі максимальної ентропії [7, 8].

Методи NER, що базуються на машинному навчанні, є більш адаптивними, ніж моделі засновані на правилах, де підтримка нових текстів потребує додаткового налаштування правил, а отже і додаткових витрат. Слід зазначити, що методи

машинного навчання для NER потребують створення навчальних даних для прогнозованого маркування тексту на основі завчасно визначених іменованих сутностей та їх категорій.

1.2 Огляд патентної літератури

Тематика розпізнавання іменованих сутностей є досить актуальною, а тому знаходить своє відображення у винахідницькій діяльності. Нижче наведено опис декількох запатентованих методів та підходів щодо обробки тексту.

Для розгляду було обрано наступні патенти:

– ефективний і точний метод і пристрій розпізнавання іменованих сутностей [9];

– метод і система автоматизованого розпізнавання сутностей [10];

– розпізнавання іменованих сутностей з використанням deep learning (DL) [11].

У рамках патенту «Ефективний і точний метод і пристрій розпізнавання іменованих сутностей» було описано алгоритм, що дозволяє вилучати та розпізнавати іменовані сутності в цифрових документах та текстах в умовах обмеженого об’єму пам’яті для обчислень без втрати швидкості отримання результату. Стисла модель NER навчається за допомогою даних, що містить анотований корпус, що оптимізується за рахунок векторної таблиці оптимізації та параметрів оптимізації. Векторна таблиця оптимізації включає в себе кластеризацію речень у навчальному датасеті за векторами слів, а також вибір корпусів з кожного кластеру для включення у набір даних, на основі якого алгоритм буде навчатися. Параметри оптимізації необхідні для зменшення розміру векторного простору слів у наборі навчальних даних, зменшення розмірності символьного векторного простору навчального набору даних, спрощення представлення символів з використанням схеми даних із меншим споживанням пам'яті замість векторів багатовимірних символів.

На рисунку 1.1 зображена схема формування стислої моделі NER.

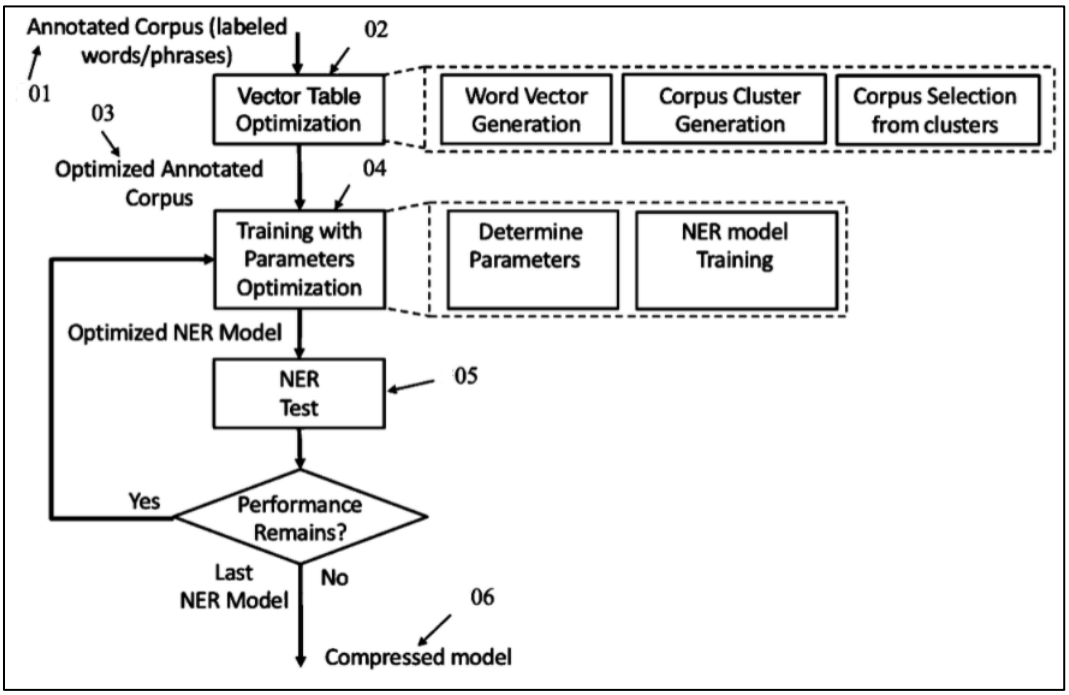

Рисунок 1.1 – Схема формування стислої моделі NER [9]

Ця схема включає наступні етапи. Спочатку анотований корпус 01 оптимізується за допомогою оптимізації векторної таблиці 02 для створення оптимізованого анотованого корпусу 03, який у подальшому має бути оптимізований за допомогою оптимізації параметрів 04, мета якої полягає в тому, щоб ітеративно зменшити якомога більше вимірів і, в свою чергу, споживання пам’яті для навчального набору даних, перш ніж точність розпізнавання почне погіршуватися. На виході 06 - стиснута модель NER.

На рисунку 1.2 (див. с. 15) зображено схему алгоритму для розпізнавання іменованих сутностей. У якості вхідних параметрів алгоритму може служити текст у вигляді цифрового документу або повідомлення. На першому етапі виконується розпізнавання іменованих сутностей з використанням розпізнавача, що був навчений за допомогою стислої моделі NER. На цьому етапі визначається чи існує у тексті хоча б одна іменована сутність і якщо так, то генерується результат першого етапу, що

включає в себе інформацію про розпізнані іменовані сутності, їх класи або типи та ймовірність точності розпізнання. У противному випадку, система NER направляє введений текст для обробки розпізнавачем іменованих об'єктів 06, що базується на основі правил. При цьому, якщо на першому етапі ймовірність розпізнання іменованої сутності вище порогового значення, то результат може перенаправлятися до інтегратору результатів 08, у противному випадку – до етапу 06.

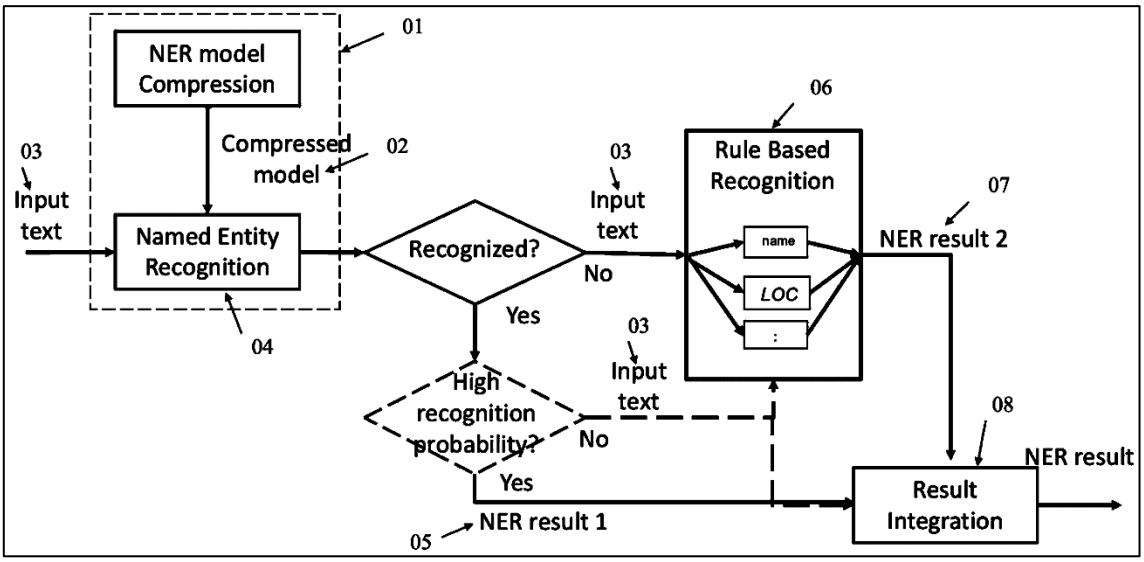

Рисунок 1.2 – Схема роботи алгоритму [9]

Розпізнавач іменованих сутностей на основі правил містить модуль загальних правил і модуль конкретних правил, що призначені для виявлення найбільш часто вживаних слів. Обробка тексту модулями може відбуватися як послідовно, так і паралельно.

Модуль загальних правил являє собою класифікатор NER, навчений за однією або кількома загальними ознаками, включаючи найбільш часто використовувані N слів, а також тегів, що позначують до якої частини мови ці слова відносяться. Загальні ознаки отримують з навчального набору даних шляхом розбиття речень на окремі слова, виділення перших N слів, які найчастіше з'являються в наборі навчальних даних, та призначення POS-тега для кожного окремого слова.

Модуль конкретних правил налаштований на виконання NER для вхідного тексту за допомогою ідентифікації регулярного виразу, специфічного для конкретного класу або типу сутності. Ці регулярні вирази можуть бути визначені у словниках класи та типи іменованих сутностей.

Результати роботи двох модулів NER об'єднуються і утворюють загальний результат другого етапу.

Кінцевий результат алгоритму формується наступним чином:

– якщо результати обох етапів алгоритму не порожні, то обидва результати є вихідними даними;

– якщо результат першого (або другого) етапу порожній, то вихідним значенням є непорожній результат другого (або першого) етапу відповідно [9].

У рамках патенту «Метод і система автоматизованого розпізнавання сутностей» було запропоновано винахід для вилучення понять та іменованих сутностей за допомогою методів, базованих на основі правил, на основі обробки природної мови та на основі знань. На рисунку 1.3 (див. с. 17) зображено блок схему системи, що призначена для ідентифікації концепцій та іменованих сутностей у документі, що містить текст природною мовою.

Система складається з наступних компонентів:

– текстовий процесор 01;

– механізм розпізнавання сутностей 02;

– менеджер лінгвістичних ресурсів 03;

– модуль усунення неоднозначностей 04.

Спочатку вхідний текст обробляється за допомогою текстового процесору, що включає в себе наступні складові: токенізатор 12, синтаксичний аналізатор 14, механізм вилучення акронімів та абревіатур 16 та лексичний процесор для обробки тексту 18.

Токенізатор 12 призначений для розбиття документа на розмічені речення і слова. Його основна роль − створення ряду токенів слів, на основі яких будуть створені відповідні синтаксичні класи або класи частин мови.

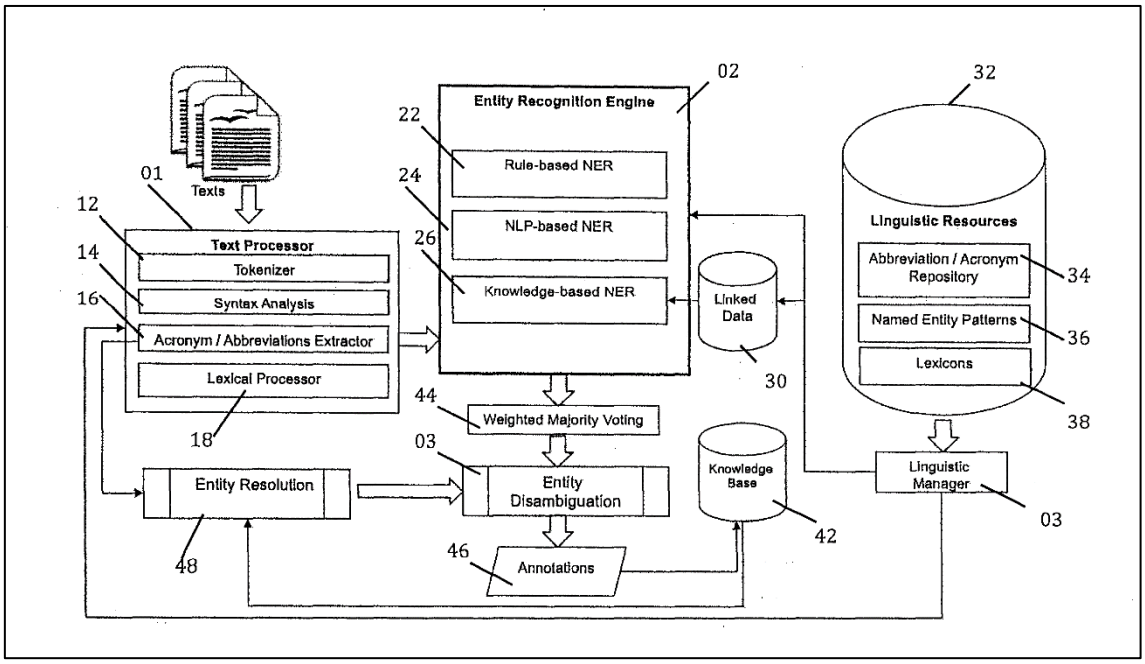

Рисунок 1.3 – Блок схема системи NER [10]

Синтаксичний аналізатор 14 використовується для створення синтаксичної структури і шаблонів сутностей. Він аналізує токенізовані речення та ідентифікує усі можливі словосполучення іменників, які мають бути зіставлені із патернами для розпізнавання сутностей.

Механізм вилучення акронімів та абревіатур 16 знаходить та розширює всі акроніми/абревіатури та передає їх до модулю роздільної здатності сутностей 48, що перетворює акроніми та скорочень у розпізнані сутності з бази знань 42.

Лексичний процесор 18 використовується у якості попередньої обробки тексту та допомагає визначити, чи існує токен як окреме складове слово або іменник, що у подальшому спрощує процес розпізнавання іменованих сутностей.

Менеджер лінгвістичних ресурсів 03 необхідний для управління взаємодії між пристроєм співставлення за патернами та лінгвістичними ресурсами, а також взаємодії між засобом пошуку термінів і пов'язаними даними 30. Також лінгвістичний менеджер 03 надає необхідну інформацію з бази даних лінгвістичних ресурсів 32 до механізму розпізнавання об'єктів 02.

Механізм розпізнавання об'єктів 02 включає в себе логіку розпізнання іменованих сутностей на основі правил (NER 22), на основі обробки природної мови (NER 24) та на основі знань для обробки тексту (NER 26). Даний механізм спочатку зіставляє вхідні дані за патернами з сутностями з лінгвістичних ресурсів, а потім шляхом пошуку термінів і їх співставленням з цільовим рядком.

Оцінки від механізму розпізнавання об'єктів 02 обробляються за допомогою модуля голосування зваженою більшістю 44. Результати NER, які отримали найвищий бал будуть анотовані за допомогою анотатора 46 та збережені у базі знань 42.

Модуль 04 необхідний для усунення неоднозначностей сутностей з урахуванням контексту речень [10].

Таким чином, даний механізм розпізнавання іменованих сутностей підходить до проблеми ідентифікації комплексно та надає можливості розширення та адаптації системи для обробки текстових даних за рахунок введення нових лінгвістичних словників та тезаурусів у систему у вигляді пов’язаних даних.

У рамках патенту «Розпізнавання іменованих сутностей з використанням DL» було запропоновано метод розпізнавання іменованих сутностей за допомогою глибокого навчання на великомасштабних наборах даних (корпусах тексту) [11]. Корпус даних може включати в себе структуровані та\або неструктуровані тексти або документи. У зв’язку з появою нових термінів та слів було запропоновано даний метод задля підвищення ефективності та надійності систем NER, що могли б генерувати нові та доповнювати існуючі словники сутностей, базуючись на корпусах текстів.

У даному винаході можуть бути використано будь-який алгоритм\метод машинного навчання (ML), що можна навчити на розмічених та\або розмічених

наборах даних. Зокрема, методи контрольованого, частково контрольованого, неконтрольованого, лінійного та нелінійного машинного навчання, а також методи ML, що пов’язані з класифікацією та регресією.

Алгоритми контрольованого ML включають у себе наступні: штучні нейронні мережі, апріорний алгоритм, алгоритм найближчого сусіда, машини опорних векторів, умовні випадкові поля, алгоритм k-ближнього сусіда, байєсовські мережі, прихована модель Маркова та інші.

Алгоритми частково контрольованого ML включають у себе генеруючі моделі, розділення з низькою щільністю, методи на основі графів тощо.

Алгоритми неконтрольованого ML включають у себе наступні: алгоритм максимізації очікування, векторне квантування, генеруюча топографічна карта тощо.

Методи машинного навчання штучної нейронної мережі (ANN – Artificial neural network) включають у себе наступні: штучні нейронні мережі, рекурентні нейронні мережі, згорткові нейронні мережі, LSTM-умовні випадкові поля, двох-направлений LSTM тощо. Методи машинного навчання з DL включають у себе наступні: глибокі машини Больцмана, глибокі нейронні мережі.

На рисунку 1.4 (див. с. 20) зображено діаграму системи розпізнавання іменованих сутностей з використання DL. Ця система 100 включає у себе наступні складові:

– система машинного навчання 104 (NER-ML);

– корпус тексту або документів 102, що є вхідними даними для NER-ML;

– модуль для ідентифікації нових іменованих сутностей 108;

– словники іменованих сутностей 106;

– модуль з результатами розпізнаних іменованих сутностей 110, що використовуються для обновлення та створення нових словників ідентифікованих сутностей.

Система NER-ML 104 може включати в себе декілька моделей NER-ML 104a-104n. Ця модель може бути багато класовою моделлю ML, що може бути навчена

та налаштована для прогнозування та ідентифікації одного або декількох типів сутностей. Після обробки корпусу тексту 102 результат може включати в себе ідентифіковані сутності, їх типи, а також їх положення у початковому тексті.

Результат обробки корпусу тексту 102 системою NER-ML 104 потім потрапляє до модулю ідентифікації сутностей 108, що порівнює отримані сутності з тими, що уже зберігаються у словниках іменованих сутностей 106 та виявляє нові та додаткові сутності, що можуть у свою чергу розширити словники. Кожен словник може включати у себе сутності певного типу. Якщо система ідентифікує відмінні сутності від тих, що уже відомі системі, то ці сутності ідентифікуються як нові результати.

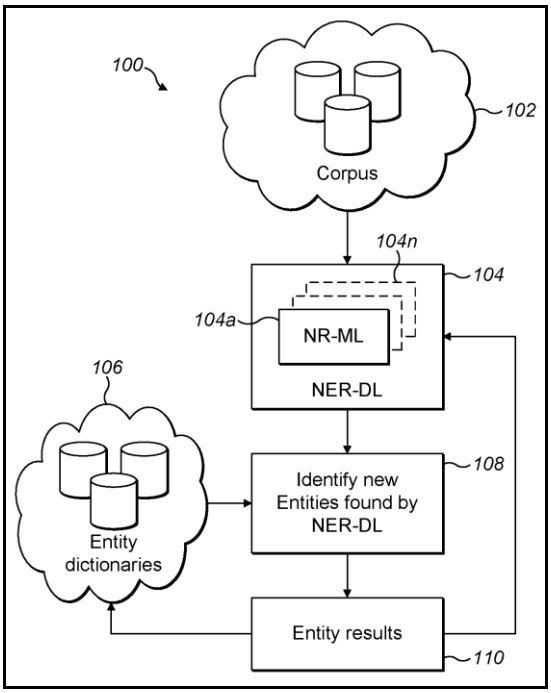

Рисунок 1.4 – Діаграма системи NER з DL [11]

Модуль результатів 110 виконує наступні функції:

– обновлює та доповнює сутності словників 106;

– створює словники іменованих сутностей для нових сутностей або їх типів на основі отриманих результатів;

– створює та обновлює набір навчальних даних, що може у подальшому бути використаний для навчання щонайменше однієї з існуючих моделей системи NER-ML 104 або для додаткових моделей NER-ML.

Ця система розпізнавання іменованих сутностей є досить гнучкою, так як дозволяє працювати як з структурованими, так і з неструктурованими даними, а також включати до системи додаткові алгоритми для ідентифікації іменованих сутностей з використанням уже існуючих оброблених наборів даних.

1.3 Огляд наукової літератури

1.3.1 Огляд закордонних публікацій

У зв’язку з різноманітністю існуючих природних мов виникає потреба у дослідженні особливостей їх обробки.

У роботі «Named Entity Recognition for Sensitive Data Discovery in Portuguese» було розглянуто та проаналізовано методи для розпізнавання іменованих сутностей у вихідних текстах португальською мовою [6]. Було зазначено, що основними методами для розпізнавання іменованих сутностей є наступні:

– методи, що базуються на правилах та граматичних патернах;

– методи, що базуються на словниках;

– методи машинного навчання.

Особливостями першої групи методів, що базуються на правилах є те, що вони дозволяють ефективно розпізнавати іменовані сутності у вихідних даних без використання навчальних даних. Недоліками таких методів наступні:

– потребують досвіду та глибоких знань граматики мови;

– проблематика адаптації методів, що базуються на правилах, до інших мов або предметних областей застосування цих методів;

– вартість та складність підтримки рішень, що базуються на правилах, у подальшому.

Методи, що базуються на словниках, потребують наявності бази знань для розпізнавання сутностей. Для знаходження сутностей у словниках застосовують не лише пошук за повним співпадінням слів, а й пошук за коренями слів, що досягається за рахунок лематизації.

У даній роботі було представлено алгоритм розпізнавання іменованих сутностей, що включає в себе три модулі (див. рис. 1.5).

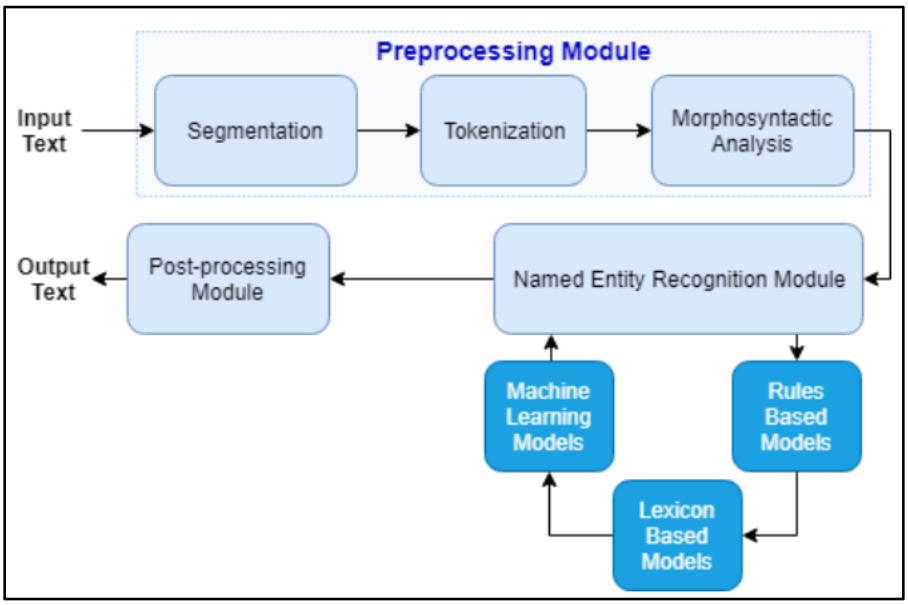

Рисунок 1.5 – Алгоритм розпізнавання іменованих сутностей [6]

Першим модулем є модуль попередньої обробки тексту, який включає наступні етапи:

– сегментація – розбиття вхідного тексту на частини за знаками пунктуації;

– токенізація – розбиття речень на слова або набори слів;

– морфологічний аналіз – визначення, до якої частини мови відноситься слово.

Модуль для розпізнавання іменованих сутностей включає в себе три підмодулі:

– модуль, що базується на правилах та використовується для визначення таких іменованих сутностей, як ідентифікаційні номери, індекси, адреси поштових електронних скриньок, формати дат тощо;

– модуль, що базується на лексиці та використовується для визначення імен, локацій, назв професій, назв валют тощо в умовах, коли немає у наявності достатнього розміченого корпусу;

– модуль, що базується на машинному навчанні та використовується для визначення тих же іменованих сутностей, що і в попередньому підмодулі.

Останній модуль системи використовується для зберігання попередніх результатів обробки текстів, а також компонування нових результатів.

У рамках проведених досліджень з використанням описаного алгоритму розпізнавання іменованих сутностей для португальської мови було отримано наступні результати. Найбільш ефективними підходом для визначення локацій та місць є моделі, що базуються на машинному навчанні. На рисунку 1.6 наведено порівняння ефективності методів для визначення іменованих сутностей різних типів.

Entity Class	Lexicon-Based Model	CRF Model	RF Model	Bi-LSTM Model
TEMPO	**91.7%**	65.9%	75.2%	71.27%
VALOR	**62.8%**	34.6%	18.6%	-
PESSOA	39.5%	69.4%	60.2%	**80.78%**
LOCAL	51.9%	77.5%	58.4%	**80%**
ORGANIZACAO	-	47.1%	34.1%	**80.5%**

Рисунок 1.6 – Точність визначення іменованих сутностей

З рисунку 1.6 очевидно, що найбільш ефективними методами для визначення іменованих сутностей локацій (LOCAL) португальською мовою є методи машинного навчання, що базуються на моделі Bi-LSTM (двонаправлений LSTM) та CRF (Conditional random fields).

У рамках роботи «Location Named-Entity Recognition using Rule-Based Approach for Balinese Texts» було досліджено ефективність використання методів, що базуються на правилах, для визначення іменованих сутностей локацій у вихідних текстах балійською мовою [12].

Загальна схема запропонованого алгоритму зображена на рисунку 1.7.

Алгоритм, описаний у роботі, включає модуль попередньої обробки тексту, який має наступні етапи:

– видалення зайвих символів пунктуації, таких як \"#$%&()*+:;<=>@[\]^_`{|}~\n;

– нормалізація тексту, що полягає у зміні специфічних символів балійської мови до стандартних. Наприклад, заміна символу «é» на символ «e»;

– токенізація – розбиття речень на токени у вигляді словосполучень.

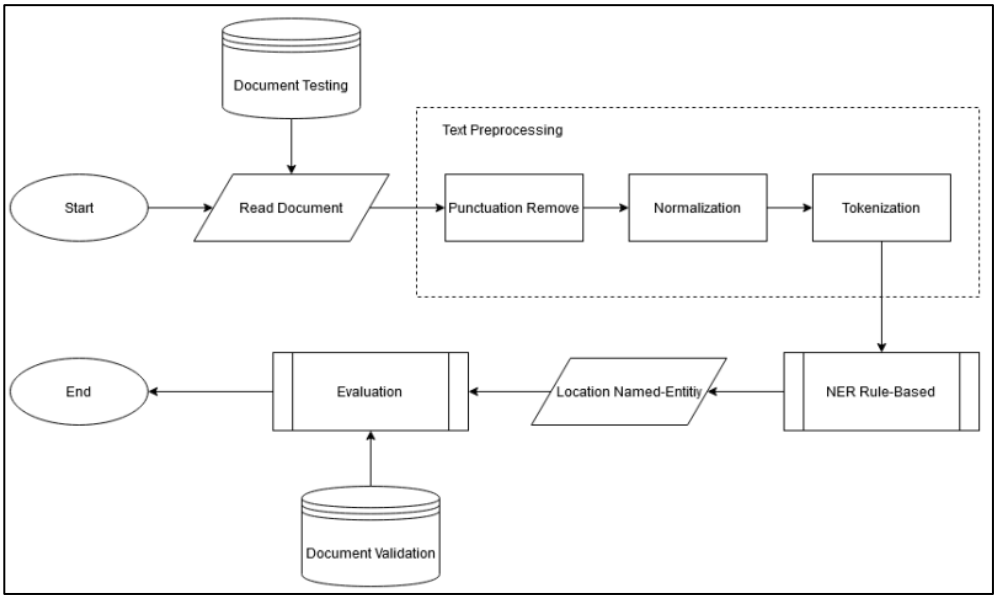

Рисунок 1.7 – Алгоритм розпізнавання іменованих сутностей [12]

Наступним модулем – є модуль розпізнавання іменованих сутностей, що базується на правилах. Цей модуль включає в себе визначення іменованих сутностей локацій, логіка якого побудована на основі знань про граматичні форми балійської мови. Ці правила охоплюють лише специфічні форми та звороти балійської мови, що можуть позначати місця та локації. Для цього були визначені прийменники місця, префікси та суфікси, що можуть вказувати на місця, а також формати, притаманні для балійської мови. Наприклад, у якості префікса можуть використовуватися наступні

слова «Gunung» – «гора», «Desa» – «село» тощо. Суфіксами у балійській мові можуть бути наступні слова: «Utara» – «північ», «Barat» – «захід» тощо. Слова, що вказують на локації: «wewidangan» – «область», «wawengkon» – «знаходиться в».

У залежності від правил балійської мови, було визначено логіку знаходження ключових слів у тексті та виявлення наступних слів, що вказують на локації у залежності від правил граматики балійської мови.

На рисунку 1.8 наведено результати проведення досліджень з використанням наведеного алгоритму.

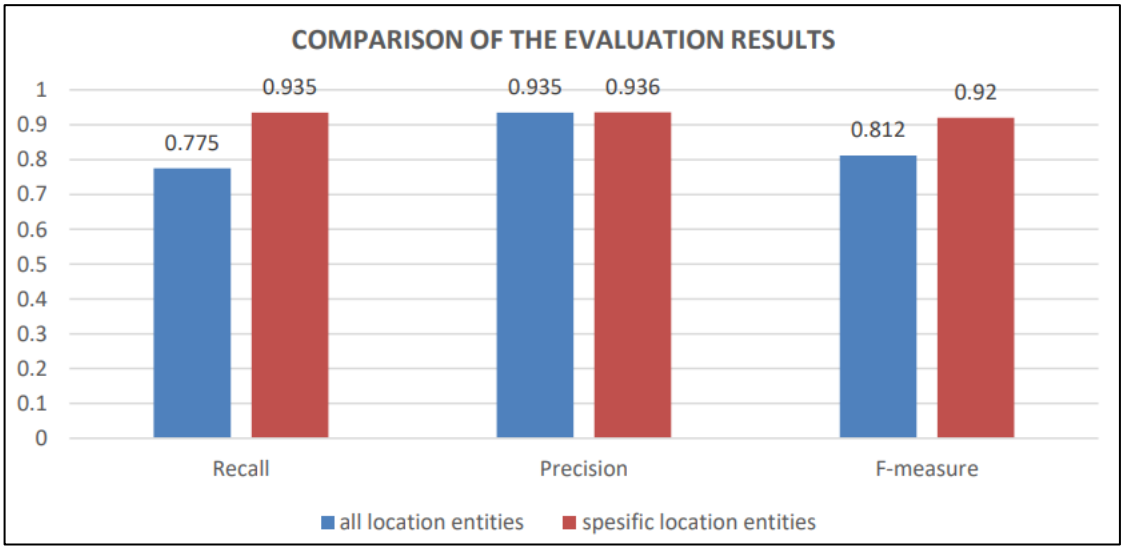

Рисунок 1.8 – Результати ефективності алгоритму [12]

Описаний алгоритм є досить ефективним для визначення специфічних локацій за заданими правилами та патернами. Проте автор у своїй роботі зазначає, що для визначення іменованих сутностей, що позначають локації, але не можуть бути задані правилами, слід використовувати алгоритми машинного навчання.

1.3.2 Огляд вітчизняних публікацій

У вітчизняних публікаціях знайшли своє відображення дослідження у сфері обробки природної мови. Зокрема, у роботі «Ukrainian News Corpus As Text

Classification Benchmark» було розглянуто проблематику формування датасетів українською мовою [13]. Проблема полягає у тому, що сучасні алгоритми машинного навчання потребують датасетів значних розмірів задля проведення досліджень. На просторах інтернету досить складно знайти готові датасети українською мовою, проте можна знайти достатню кількість статей та публікацій, що не є систематизованими та потребують обробки. Автори [13] дослідили та виявили, що для вирішення задачі автоматизованого формування датасетів досить ефективними є моделі ukr-RoBERTa та ukr-ELECTRA для текстів середнього розміру, у той час як XLM-R модель є ефективною при обробці текстів великого розміру.

Одним із найбільш поширених корпусом розмічених даних українською мовою є корпус lang-uk [14]. Цей датасет може бути використаний при вирішенні задачі розпізнавання іменованих сутностей у вихідних текстах українською мовою.

Також актуальним напрямком досліджень у сфері обробки природної мови серед вітчизняних науковців є генерація природної мови (NLG – Natural Language Generation) [15, 16]. У рамках двох робіт [15, 16] було досліджено вирішення проблеми створення знаходження відповідей на питання в автоматичному режимі. Експерименти проводилися на прикладі вихідних даних українською та англійською мовами з використанням натренованих моделей сімейства *BERT. Було виявлено, що модель mBERT є більш ефективною у випадку знаходження відповіді на питання, подібних тим, що використовувалися для тренування моделі. При цьому, для знаходження відповіді на відмінні питання більш ефективною виявилась модель RoBERTa. При цьому, у випадку тренування моделі mBERT двома мовами (англійською та українською), точність отримання результатів була вищою у порівнянні з тренуванням моделі лише однією із вищевказаних мов [15].

Проблематику вирішення задачі NLG також було висвітлено у роботі «Neural Natural Language Generation: A Survey onMultilinguality, Multimodality, Controllability and Learning» [17]. Дане дослідження пропонує підхід до вирішення стандартних задач NLP не декларативним шляхом, коли фактично задається формальна чи

математична сукупність правил, яким має відповідати модель, а генеративно-змагальним шляхом, коли корекція роботи моделі відбувається за рахунок порівняння з реальними даними, а до процесу генерації вносяться відповідні зміни. У роботі [17] розглянуто декілька основних задач NLG (машинний переклад, генерація опису, автоматичне розпізнавання тексту, абстракційне реферування, спрощення тексту, генерація питань та відповідей генерація діалогів) та описано напрямки їх подальшого вивчення. Було розглянуто нейронні методи вирішення задач NLG з точки зору наступних аспектів: багатомовності, багатомодальності (наявності різних форматів даних, що потребують обробки), а також багатозадочності.

Таким чином, дослідження вітчизняних науковців вказують на актуальність проблематики обробки природніх мов, у тому числі української мови, а також відкривають нові можливості для проведення досліджень у вирішенні сумісних задач NLP. Зокрема, дослідження у сфері автоматичного генерування датасетів українською мовою може значно полегшити процес формування та пошуку даних для навчання моделей та проведення подальших експериментів при вивченні та розв’язанні таких задач, як задачі автоматичного розпізнавання іменованих сутностей у структурованих та неструктурованих текстових даних. Дослідження задач NLG в аспекті багатомовності може відкрити нові можливості для обробки природних мов, у тому числі української.

2 НАУКОВО-ТЕХНІЧНА ЗАДАЧА

У кваліфікаційній роботі магістра необхідно вирішити ряд задач. У першу чергу провести аналіз теоретичних відомостей та публікацій, які характеризують сучасний розвиток методів і технологій розпізнавання іменованих сутностей. Далі необхідно дослідити методи розпізнавання іменованих сутностей на прикладі різних мов, з різною кількістю ресурсів. Дослідити ефективність існуючих методів в для різних мов враховуючи специфіку граматик та лексики різних природних мов.

Основною метою дослідження є синтез методу розпізнавання іменованих сутностей у неструктурованому тексту на прикладі виявлення сутностей локацій та місць. У рамках дослідження було прийнято рішення розглядати лише іменовані сутності локацій та місць з наступних причин:

– зменшення області досліджень іменованих сутностей;

– дослідження специфіки обробки іменованих сутностей локацій та місць;

– широка сфера застосування алгоритмів та методів для розпізнавання локацій та місць в рамках існуючих програмних системах.

Задачами дослідження є наступні:

– вивчення стану наукових надбань у питаннях обробки природних мов, зокрема розпізнавання іменованих сутностей, для української мови;

– дослідження існуючих методів та наукових робіт щодо питання розпізнавання іменованих сутностей для локацій в українській та іноземних мовах;

– розгляд та опис наявних технологій, розробка програмного забезпечення та опис його архітектури;

– визначення метрик для вимірювання ефективності розробленого методу для вирішення питання розпізнавання іменованих сутностей локацій та місць;

– розробка та опис плану проведення експериментів, опис отриманих результатів проведених експериментів, підбиття підсумків дослідження.

3 ОПИС ТЕОРЕТИЧНИХ ТА ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

3.1 Опис теоретичних досліджень

У ході огляду наукової та патентної літератури було виявлено, що досить поширеними методами для розпізнавання іменованих сутностей є ті, що базуються на правилах, та ті, що базуються на машинному навчанні. Комбінація цих методів у рамках однієї системи для розпізнавання іменованих сутностей забезпечує ефективність та гнучкість цих систем.

У рамках даного дослідження пропонується наступна система для задачі розпізнавання локацій та місцезнаходжень у неструктурованому тексті, що включає у себе наступні модулі:

– модуль для попередньої обробки вхідного тексту;

– модуль для розпізнавання іменованих сутностей локацій за допомогою машинного навчання;

– модуль для розпізнавання прийменників місця за допомогою машинного навчання;

– модуль для розширення розпізнаних іменованих сутностей за допомогою алгоритму, що базується на правилах;

– модуль для виводу результату обробки тексту.

На рисунку 3.1 (див. с. 30) зображена загальна структура системи.

Першим етапом системи є попередня обробка тексту, суть якої полягає у розбитті вхідного тексту на токени, а також проведення морфологічного аналізу речень. При проведенні морфологічного аналізу проводиться POS-тегування слів речень, виявлення зв’язків між словами, а також виявлення властивостей слів відповідно до універсальної анотації лексичних та граматичних властивостей слів [18].

Другим етапом є розпізнавання іменованих сутностей. Модуль розпізнавання іменованих сутностей включає в себе два компоненти:

– розпізнавання іменованих сутностей локацій за допомогою машинного навчання;

– розпізнавання іменованих сутностей за допомогою правил.

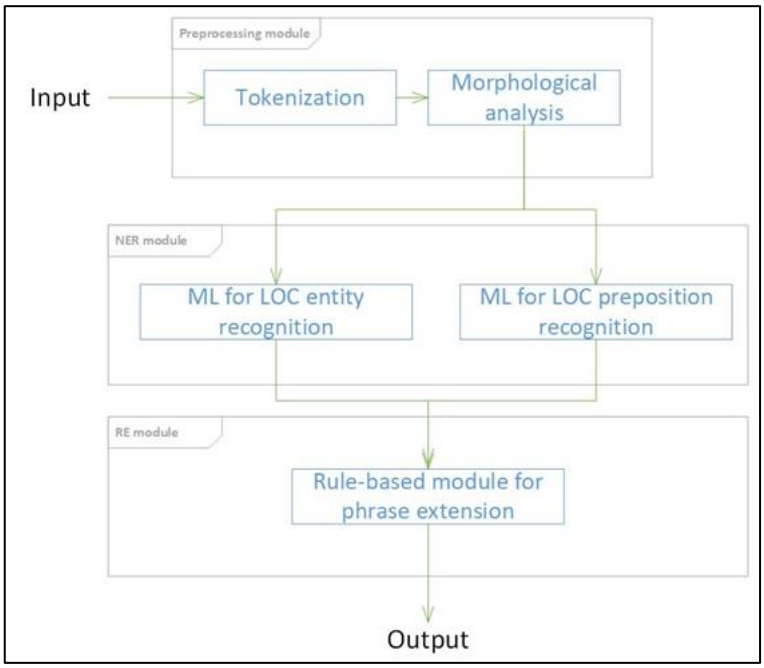

Рисунок 3.1 – Схема системи для розпізнавання іменованих сутностей

Розпізнавання іменованих сутностей локацій за допомогою машинного навчання може бути виконане за допомогою тренування моделей на корпусі українською мовою та використанні алгоритмів машинного навчання.

Розпізнавання іменованих сутностей локацій у неструктурованому тексті українською мовою за допомогою правил може бути реалізований на базі аналізу зв’язків між словами у реченнях. Механізм RE дозволяє виявити типи залежностей між словами у реченні і виявити слова, що вказують на місця та локації.

Фреймворк Universal Dependencies (UD) описує та систематизує анотацію граматик різних природніх мов світу. Ця анотація призначається для опису частин мов, морфологічних ознак слів та синтаксичних зв’язків між ними [19].

При аналізі особливостей граматики української мови можна виявити, що у більшості випадків для задання місця та локацій у реченнях використовуються просторові прийменники [20]. Приклади просторових прийменників та приклади їх використання наведені у таблиці 3.1. Слід зазначити, що в таблиці наведена лише частина прийменників, що можуть вказувати на місця.

Таблиця 3.1 – Просторові прийменники

Прийменник	Приклади використання
у	У місті, у будинку, у саду, у магазині
в	В Ужгороді, в озері, в театрі
на	На дорозі, на столі, на дворі, на площі
до	До магазину, до зупинки, до школи
від	Від зупинки, від берега, від дороги
під	Під столом, під стріхою, під яблунею
над	Над головою, над річкою
за	За хатою, за будинком, за рогом вулиці
через	Через річку, через двір, через степ
при	При дорозі, при долині
перед	Перед машиною, перед двором, перед під’їздом
поза	Поза гаєм, поза річкою, поза садом, поза двором
біля	Біля кафе, біля пам’ятнику, біля школи
обабіч	Обабіч дороги
коло	Коло будинку, коло запинки, коло школи
по	По дорозі, по воді, по вулиці

При утворенні словосполучень з вищевказаними прийменниками формуються залежність між словами, де прийменник місця є залежним словом, а

іменник/займенник, що утворює словосполучення, виступає у ролі основного слова. Таку залежність можливо виявити за допомогою RE. Зокрема, фреймворк UD анотує такі залежності як «Case». «Case» залежність використовується для позначення зв'язків між іменники та іншими частинами мови (прикметниками, числівниками, прийменниками). При цьому існують підвиди «Case» залежностей, одна з яких є «Loc». «Loc» ідентифікатор використовується для позначення зв'язків між словами, що вказують на місце або час. Наприклад, у реченні «У червні я був у Києві» є дві «Case» залежності типу «Loc»: перша – «У червні», що позначає час; друга – «у Києві», що позначає місце. При знаходженні таких залежностей можна виявити основні слова, що вказують або на місце, або на час.

Аналіз «Case» залежностей типу «Loc» може бути застосовано при розробці алгоритму, що базується на правилах, для виявлення локацій у неструктурованому тексті.

Цей підхід дозволяє виявити додаткові іменовані сутності локацій, що не були виявлені при спробі розпізнати іменовані сутності за допомогою алгоритмів, що базуються на машинному навчанні. Також результати розпізнавання іменованих сутностей за допомогою алгоритмів, що базуються на правилах, у подальшому можуть бути використані для удосконалення моделей, що використовуються у алгоритмах машинного навчання.

Третім етапом системи є розширення розпізнаних іменованих сутностей локації. Цей етап необхідний для того, щоб виявляти не лише ключові слова, що вказують на місце, а й залежні слова, що можуть надавати додаткову інформацію при заданні локацій. Наприклад, у реченні «Він перебуває у Харківській обласній лікарні» на місце вказують такі слова як «у лікарні», проте також наявна додаткова інформація, що позначається словами «Харківській» та «обласній».

При цьому результати розпізнаних іменованих сутностей на другому етапі за допомогою обох алгоритмів з цього етапу окремо можуть розширені з використанням третього етапу системи.

Додаткову інформацію про місцезнаходження можливо виявити за допомогою аналізу зв’язків між словами у реченні. Нижче наведено опис типів зв’язків з наявних у фреймворці Universal Dependencies, які можна використовувати для вирішення цієї задачі [21].

Зв’язок «nmod» використовується для позначення неузгодженого означення відносно іменника. Наприклад, у словосполученні «берег річки» слово «берег» є основним, а слово «річки» – залежним та описує основне слово.

Зв’язок «nummod» використовується для позначення кількісних числівників. Наприклад, словосполучення «третій поверх» містить основне слово «поверх» та залежне слово «третій».

Зв’язок «amod» використовується для позначення прикметників. Наприклад, словосполучення «старий садок» містить основне слово «садок» та залежне слово «старий», що у свою чергу описує основне слово.

Зв’язок «det» використовується для позначення визначень. Наприклад, словосполучення «той будинок» містить основне слово «будинок» та залежне слово «той». У якості залежних слів можуть бути такі слова як «той», «цей».

Зв’язок «flat» використовується для з’єднання слів. Цей вид зв’язку може бути використаний для позначання залежності слів між ім’ям та по батькові, в датах – для залежності слів між днем та місяцем, для адрес – для позначення залежності між назвою вулиць та номерами будинків. Наприклад, словосполучення «Гагаріна 35» містить основне слово «Гагаріна», що позначає назву вулиці, та залежне слово «35», що позначає номер будинку.

Зв’язок «compound» використовується для позначення чисел. Наприклад, словосполучення «мінус другий поверх» містить зв’язок «compound» відносно слів «мінус другий», при цьому основним словом є «другий», а залежним – «мінус».

Зв’язок «reparandum» використовується для позначення виправлених слів. Наприклад, у реченні «Йдіть прав-… ліворуч» існує зв’язок цього типу між словами «прав-» та «ліворуч», де перше слово є залежним, а друге – основним.

При цьому при аналізі вищенаведених зв’язків та знаходженні залежних слів від уже відомого основного слова, додатковий аналіз зв’язків щодо залежних слів проводити немає необхідності. Проте, якщо відомо залежне слово та за допомогою вищенаведених слів було знайдено основне слово, то додатковий аналіз зв’язків відносних цих знайдених слів проводити необхідно.

Також необхідно проаналізувати наступні типи зв’язків для знаходження додаткової інформації про сутності локацій.

Зв’язок «appos» використовується для виявлення прикладки. Наприклад, словосполучення «Шевченко, гарний сад» містить основне слово «Шевченко» та залежне слово «сад», що дає характеристику та уточнює значення основного слова.

Зв’язок «acl» використовується для позначення дієприкметникових зворотів або залежної частини означального речення відносно головного слова. Наприклад, у словосполученні «сад, у якому ми були» є зв’язок цього типу між словами «сад», що виступає у якості основного слова, та «були», що є залежним словом. При аналізі даного типу зв’язку при наявності основного слова «сад», слід аналізувати усі залежні слова в інших зв’язках від залежного слова «були» у даному типу зв’язку, незалежно від того виступає знайдене слово «були» в інших зв’язках основним або залежним. Так, як усі інші зв’язки, що будуть включати знайдене залежне слово «були» будуть мати посилання на залежну частину означального речення, що є уточнювальною, або на слова дієприкметникового звороту.

Зв’язок «conj» використовується для позначення однорідних членів речення. Однорідними членами речення є такі слова, що виконують одну синтаксичну роль та сполучаються між за допомогою сурядного зв’язку. Наприклад, у реченні «Звідти він наносить удари по селищам і селам» цей тип зв’язку існує між словами «селищам» та «селам». Ці слова є рівноправними та не залежать один від одного. Проте, якщо у сутності з локацією відоме одне з цих слів, то цей тип зв’язку дозволяє виявити інформацію про інші локації, наявні у реченні. Нове слово, знайдене за допомогою цього типу зв’язку має бути проаналізовано подібно уже відомому слову.

Зв’язок «case» також має бути проаналізовано, проте для тих зв’язків, у яких основне слова має в значенні feats (в ознаках цього слова) позначення «LOC», що вказує на належність цього слова до сутності типу локацій або часу. Знайдене залежне слово є прийменником та дозволяє розпізнати локацію більш точно. Так як прийменники можуть вказувати на відносність основного слова та уточнювати контекст у якому основне слово використовується. Наприклад, для словосполучень «перед будинком» та «за будинком» прийменник значно змінює вказання локацій.

Таким чином, обидва компоненти системи, а саме компонент для ропізнавання іменованих сутностей локацій та місцезнаходжень та компонент для розширення розпізнаних іменованих сутностей разом мають надати можливість знаходити вичерпну інформацію щодо наявних у неструктурованому тексті сутностей, що вказують місце.

3.2 План проведення експериментів

У рамках експерименту має бути досліджена ефективність роботи компонентів системи у наступних конфігураціях:

– розпізнавання іменованих сутностей локацій за допомогою машинного навчання без застосування RE (компонент ML);

– розпізнавання іменованих сутностей локацій за допомогою машинного навчання за допомогою RE (компонент ML with RE);

– розпізнавання іменованих сутностей за допомогою алгоритму, що базується на правилах, з застосування RE (компонент Rule-Based with RE);

– розпізнавання іменованих сутностей з використанням обох алгоритмів (того, що базується на машинному навчанні, та того, що базується на правилах) з використанням RE (ML + Rule-Based + RE).

Для проведення експериментів мають бути виконані наступні кроки:

– реалізація програмної системи, опис якої наведено у розділі 5;

– визначання метрик вимірювання ефективності розпізнавання іменованих сутностей;

– збір даних для проведення експериментів;

– проведення вимірювань ефективності роботи системи при різних її конфігураціях.

3.3 Метрики ефективності розпізнавання іменованих сутностей

Для визначення ефективності розпізнавання іменованих сутностей у неструктурованому тексті в рамках дослідження має бути використана F-міра. F-міра обчислюється на основі значень Precision та Recall. У формулі 1 наведено розрахунок Precision, що дозволяє обчислити відсоток вірно виявлених сутностей серед усіх розпізнаних сутностей.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

де TP – кількість істинних позитивних рішень;

FP – кількість помилково позитивних рішень.

У формулі 2 наведено розрахунок значення Recall, що відображає відсоток вірно розпізнаних іменованих сутностей серед тих, що мали бути розпізнані.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

де TP – кількість істинних позитивних рішень;

FN – кількість помилково негативних рішень.

У формулі 3 наведено обчислення значення F-міри.

$$F = 2\frac{Precision * Recall}{Precision + Recall} \quad (3)$$

Слід зазначити, що при проведенні дослідження, у якості одиниці, що відображає рішення було взято не шукану сутність як набір слів, а окремі слова у реченні. Наприклад, якщо сутність локації включає у себе декілька слів, то кожне окреме слово цієї сутності вважається окремим рішенням при обчисленнях. Це пов'язане з тим, що у рамках досліджень об'єктом для розпізнавання є не лише ключові слова, що мають відношення до задання локацій та місць у реченнях, а й допоміжні слова, що уточнюють ключові.

3.4 Збір даних для проведення експериментів

У якості корпусів українською мовою для навчання моделей для розпізнавання іменованих сутностей був використаний український корпус lang-uk, що включає у себе 262 тексти [14], первинним джерелом яких є відкритий корпус українських текстів – Браунський корпус української мови [22].

У якості додаткових даних для проведення експериментів щодо ефективності розпізнавання іменованих сутностей локацій було використано художню українську літературу, а також набір даних з існуючих новинних чатів та каналів, що наразі містять достатньо інформації з вказанням локацій.

Джерелом текстів для аналізу сутностей локацій з художньої літератури було використано соціально-побутову повість «Кайдашева сім'я» Івана Семеновича Нечуй-Левицького, автобіографічну повість «Зачарована Десна» Олександра Довженка, роман «Хіба ревуть воли, як ясла повні?» Панаса Мирного.

У якості новинних каналів були використані сучасні місцеві telegram-канали, що включають інформацію про події в селах, селищах, районах та областях України, наприклад, новинний канал «ДЕРГАЧИ ПОДСЛУШАНОUA|Харьков, Украина», «информатор мд огт», телеграм-чат «Старомайорское, Макстрой, Макаровка» та «Вн информатор» та інші канали. Ці канали були обрані, так як містять достатньо текстів українською мовою, що містять інформацію про місця подій або запитів про стан рідних з вказанням додаткової інформації про їх місце знаходження.

При формуванні вибірки тестових даних 20% даних було взято з художньої літератури та 80% з новинних каналів. Такий розподіл у першу чергу пов’язаний з тим, що канали містять тексти у тому вигляді, у якому користувачі систем могли б вказувати та описувати локації у реченнях у сторонніх сервісах. Наприклад, телеграм чати включають у себе наступні повідомлення «Якщо хтось зможе, подивіться будь ласка, чи живі люди на вул. Пушкіна 81 та на вул. Леніна (напроти магазину ‘Теремок’).» або «Може хтось знає, де зараз Петро та сім'я, які проживали на Нижній і були у підвалі біля міліцейського будинку.». Ці речення включають у себе як і частини інформації про адресу, так і додаткову інформацію з деталями, що можуть допомогти зрозуміти більш точне місце знаходження.

3.5 Результати проведених експериментальних досліджень

Було сформовано набір тестових даних українською мовою, що містить 200 записів. При цьому 40 з них є записами, що були взяті з художньої літератури, та 160 – з новинних телеграм-каналів.

Експерименти були проведені для конфігурацій системи, що були вказані у плані проведення експерименту. У таблиці 3.2 наведено результати проведених досліджень для усього тестового набору даних.

Таблиця 3.2 – Результати проведених експериментальних досліджень

Конфігурація	F-міра	Precision	Recall
Розпізнавання іменованих сутностей локацій за допомогою машинного навчання без застосування RE	44.31%	97.41%	28.68%
Розпізнавання іменованих сутностей локацій за допомогою машинного навчання з застосування RE	65.22%	81.46%	54.38%
Розпізнавання іменованих сутностей за допомогою алгоритму, що базується на правилах, з застосування RE	71.19%	83.1%	62.27%

Кінець таблиці 3.2

Конфігурація	F-міра	Precision	Recall
Розпізнавання іменованих сутностей з використанням обох алгоритмів (того, що базується на машинному навчанні, та того, що базується на правилах) з використанням RE	80.79%	82.88%	78.81%

Також було проведено заміри ефективності роботи системи для тестових наборів даних, взятих з художньої літератури та новинних телеграм-каналів окремо.

Результати ефективності роботи системи у різних конфігураціях для набору даних, взятих з художньої літератури, наведено у таблиці 3.3.

Таблиця 3.3 – Результати проведених експериментальних досліджень для художньої літератури

Конфігурація	F-міра	Precision	Recall
Розпізнавання іменованих сутностей локацій за допомогою машинного навчання без застосування RE	6.89%	100%	3.57%
Розпізнавання іменованих сутностей локацій за допомогою машинного навчання з застосування RE	11.23%	100%	5.95%
Розпізнавання іменованих сутностей за допомогою алгоритму, що базується на правилах, з застосування RE	56%	85.36%	41.66%
Розпізнавання іменованих сутностей з використанням обох алгоритмів (того, що базується на машинному навчанні, та того, що базується на правилах) з використанням RE	59.37%	86.36%	45.23%

Результати ефективності роботи системи у різних конфігураціях для набору даних, взятих новинних телеграм-каналів, наведено у таблиці 3.4.

Таблиця 3.4 – Результати проведених експериментальних досліджень для новин

Конфігурація	F-міра	Precision	Recall
Розпізнавання іменованих сутностей локацій за допомогою машинного навчання без застосування RE	52.01%	97.34%	35.48%
Розпізнавання іменованих сутностей локацій за допомогою машинного навчання з застосування RE	73.83%	81.1%	67.76%
Розпізнавання іменованих сутностей за допомогою алгоритму, що базується на правилах, з застосування RE	74.63%	82.73%	67.98%
Розпізнавання іменованих сутностей з використанням обох алгоритмів (того, що базується на машинному навчанні, та того, що базується на правилах) з використанням RE	85.16%	82.41%	88.11%

Таким чином, було отримано усі необхідні дані для аналізу ефективності застосування запропонованої системи для розпізнавання іменованих сутностей локацій у неструктурованому тексті.

4 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ

4.1 Аналіз результатів проведення експериментальних досліджень

Відповідно до отриманих результатів ефективності роботи запропонованої системи при різних конфігураціях можна зробити наступні висновки.

По-перше, алгоритми, що базуються на машинному навчанні, мають високу точність при розпізнанні іменованих сутностей локацій, проте є не досить ефективними при спробі розпізнати сутності локацій (F-міра становить 44.31%), що мають описовий характер. Неефективність цього алгоритму для вирішення задачу розпізнавання локацій, що не обмежуються назвами міст, вулиці, номера будинків та іншими ключовими загальновідомими словами, може бути пов’язана з недостатньою натренованістю моделі та обмеженості анотування наявного корпусу даних.

По-друге, алгоритми, що базуються на правилах, у запропонованій системі є досить обмеженими, так як орієнтовані на виявлення лише одного залежного слова від слів-маркерів, що можуть вказувати на місце, зокрема маються на увазі просторові прийменники. Саме тому цей алгоритм має використовуватися разом з механізмом RE, що допомагає знаходити слова, пов’язані з словом, що вказує на місце. На основі результатів аналізу алгоритму, що базується на правилах, з застосуванням RE, можна зробити висновок, що ефективність цього підходу є значно вищою (F-міра становить 71.19%). Слід зазначити, що зменшення точності розпізнавання іменованих сутностей пов’язана у першу чергу з морфологічним аналізом речень.

По-третє, застосування RE для результатів розпізнавання іменованих сутностей окремо для алгоритмів, що базуються на машинному навчанні, та тих, що базуються на правилах, відповідно до отриманих результатів доводить свою ефективність. Зокрема, при застосуванні RE для результатів розпізнавання іменованих сутностей за допомогою машинного навчання, ефективність розпізнавання сутностей зросла з 44.31% до 65.22%. При цьому точність зменшилась приблизно на 16%, у той час як повнота виявлення іменованих сутностей зросла на 25.4%. Тобто, Механізм RE у

цілому дозволяє ефективно виявити додаткові слова, що можуть вказувати на локації, про що свідчить ріст повноти виявлення іменованих сутностей. Зменшення точності виявлення залежних слів може свідчити про необхідність додаткового теоретичного аналізу можливих зв'язків слів у реченні в українській мові з метою зменшення хибно позитивно розпізнаних слів.

При поєднанні алгоритмів, що базується на правилах та машинному навчанні, та механізму RE експериментально було отримано F-міру рівну 80.79%, що свідчить про ефективність використання запропонованої системи для розпізнавання іменованих сутностей локацій.

4.2 Аналіз отриманих результатів для текстів з різних типів джерел

Було виявлено, що при проведенні експериментів на наборі даних, що містить записи з телеграм-каналів з новинами та телеграм-чатів з повідомленнями користувачів ефективність системи при всіх конфігураціях є вищою порівняно з результатами експериментів, проведеними на усіх наборах даних, на відповідних конфігураціях.

Це може бути пов'язано з тим, що корпус, що використовувався для навчання моделі, був побудований на засадах корпусу англійської мови, у якому більшою частиною джерел для корпусу є саме статті з новинами. Це є важливим фактором, так як модель, що тренувалася з використанням корпусу, використовується не тільки для виявлення іменованих сутностей за допомогою машинного навчання, але й для виявлення зв'язків між словами у реченні та для проведення морфологічного аналізу слів.

У свою чергу менші показники ефективності системи для розпізнавання іменованих сутностей при різних конфігураціях також пояснюється структурою початкового корпусу для навчання моделі. Проте слід зазначити, що алгоритм, що базується на правилах з використанням RE вносить значний вклад в ефективність

системи для розпізнавання іменованих сутностей локацій при наявності недоліків у структурі початкового корпусу.

Таким чином, можна зробити висновок, що запропонована система може використовуватися для систем, що потребують розпізнавання іменованих сутностей локацій у неструктурованому тексті, де наявні повідомлення користувачів з вказаними локаціями та місцями, або для аналізу статей та новин. Для того, щоб система була ефективною при роботі з текстами художньої літератури, необхідно удосконалити набір даних для тренування моделі, у тому числі розглянути можливість повторного тренування моделі на основі тих даних, що можуть бути отримані у результаті роботи системи.

5 ОПИС ПРОГРАМНОЇ СИСТЕМИ

5.1 Опис технологій

У якості програмної мови для реалізації системи для розпізнавання іменованих сутностей було обрано Python. Цей вибір пов’язаний з наявністю достатньої кількості готових рішень для обробки природніх мов.

Для морфологічного аналізу слів та розпізнавання іменованих сутностей були розглянуті наступні бібліотеки:

– spaCy [23];

– NLTK [24];

– Stanza [25].

Для вибору бібліотеки, на базі якої має бути побудована програмна реалізація системи, було визначено наступні критерії для оцінки програмних рішень:

– підтримка української мови;

– наявність механізму для виявлення залежностей між словами;

– наявність механізму для виявлення морфологічних ознак слів;

– наявність механізму для розпізнавання іменованих сутностей;

– підтримка фреймворку Universal Dependencies.

У таблиці 5.1 наведено порівняння запропонованих бібліотек за визначеними критеріями.

Таблиця 5.1 – Таблиця порівняння бібліотек

Критерій	spaCy	NLTK	Stanza
Підтримка української мови	–	–	+
Наявність механізму для виявлення залежностей між словами	+	+	+
Наявність механізму для виявлення морфологічних ознак слів	+	+	+

Кінець таблиці 5.1

Критерій	spaCy	NLTK	Stanza
Наявність механізму для виявлення морфологічних ознак слів	+	+	+
Наявність механізму для розпізнавання іменованих сутностей	+	+	+
Підтримка фреймворку UD	+–	+	+

При порівнянні трьох бібліотек за Парето найкращою альтернативою є бібліотека Stanza. Вибір пояснюється тим, що у першу чергу Stanza підтримує українську мову та містить готові натреновані моделі з використанням корпусу uk-lang. При цьому бібліотека spaCy не підтримує фреймворк Universal Dependencies для усіх наявних моделей, а лише для частини підтримуваних мов.

У якості середовища розробки було обрано PyCharm Community Edition 2022.1.

5.2 Етапи розробки програмної системи

Етапами розробки програмного забезпечення були наступні:

– збір та підготовка тестових даних;
– проектування програмного забезпечення;
– реалізація програмного забезпечення;
– тестування реалізованої системи.

На етапі збору тестових даних було обрано 200 наборів текстів з різних джерел даних. 40 записів було взято з художньої літератури та 160 – з телеграм-чатів та каналів з новинами.

На етапі проектування було визначено формат вхідних та вихідних даних. Також на етапі проведення проектування системи мають бути визначені та описані її компоненти та їх взаємодію. На етапі реалізації має бути реалізована спроектована система. Результатом цього етапу має бути готова програмна реалізація системи. На

етапі тестування має бути проведений аналіз коректності роботи системи, а також проведено експерименти щодо ефективності роботи системи.

5.3 Опис розробленої програмної системи

Готова програмна реалізація містить наступні складові:

– компонент для зчитування вхідних даних;

– компонент для морфологічного аналізу речень;

– компонент для розпізнавання іменованих сутностей локацій за допомогою машинного навчання;

– компонент для розпізнавання ключових слів локацій за допомогою алгоритму, що базується на правилах;

– компонент для запису результатів обробки даних у файл.

У якості вхідних даних використовується файл з реченнями. При цьому вхідні дані зчитуються рядок за рядком. Приклад вхідних даних наведено на рисунку 5.1.

```
Він живе на третьому поверсі.
Він живе у другому під'їзді.
У вівторок на Плеханівській, 73 відновлює роботу відділ державної реєстрації актів цивільного стану.
У різних населених пунктах Харківської області набори відрізняються, іноді якимись нюансами, інколи ж зовсім інший асортимент.
В албанській Тирані автобус після ДТП застряг над річкою.
О 18:30 вони обіцяють розпочати на площі Франції безстроковий страйк.
У результаті пожежі на об'єкті у Білгородській області різні пошкодження отримали сім будинків.
Губернатор Білгородської області повідомив, що на кордоні трьох муніципалітетів – Борисівського та Білгородського районів та Як
У мережі з'явилися фото мосту.
```

Рисунок 5.1 – Приклад формату вхідних даних

Морфологічний аналіз речень було реалізовано з використанням бібліотеки Stanza. При цьому морфологічний аналіз речень включає у себе наступні аспекти:

– розбиття речень на токени;

– розмітка слів за частинами мови;

– аналіз морфологічних особливостей слів;

– виявлення іменованих сутностей у реченні;

– виявлення залежностей між словами у реченні.

Нижче наведено реалізацію морфологічного аналізу речень:

```
LANGUAGE_CODE = 'uk'

nlp = stanza.Pipeline(lang='uk', processors='tokenize,pos,ner,depparse')
sentence: Sentence = nlp(text)
```

Для виявлення розпізнаних сутностей локацій було використано наступну логіку:

```
def get_location_entities(sentence: Sentence):
    locations = []

    for recognised_entity in sentence.entities:
        if 'LOC' in recognised_entity.type:
            locations.append(recognised_entity)

    return locations
```

Для розпізнавання іменованих сутностей за допомогою правил було використано наступну логіку:

```
def __is_location(word: Word):
    return word.feats is not None and ('Loc' in word.feats)

def get_locations_by_prepositions(sentence: Sentence):
    locations = []
    for main_word, dependency, subordinate_word in sentence.dependencies:
        if not is_location(subordinate_word) or not 'case' in dependency:
            continue

        locations.append(main_word)
    return list(dict.fromkeys(locations))
```

Для розширення іменованих сутностей за допомогою RE було реалізовано два алгоритми. Один для обробки номінальних, другий – для обробки усих інших залжностей. До номінальних типів зв’язків було віднесено настпуні: 'nmod', 'nummod', 'amod', 'det', 'flat', 'compound', 'reparandum'.

Алгоритм для обробки номінальних типів зв’язків наведено нижче:

```
def __get_subordinate_words(sentence: Sentence, word: Word):
```

```
    related_words = []
    for main_word, dependency, subordinate_word in sentence.dependencies:
        if not is_nominal_dependency(dependency):
            continue

        if word == main_word:
            related _words.append(subordinate_word)
        elif word == subordinate_word:
            related_words.append(main_word)
            related_words.extend(get_subordinate_words(sentence, main_word))
            related_words.extend(get_subordinate_phrase(sentence,main_word))

        return subordinate_words
```

Алгоритм для обробки усих інших зв’язків, що були описані у попередніх розділах, наведено нижче:

```
def __get_subordinate_phrase(sentence: Sentence, word: Word):
    related_phrase_words = []

    for main_word, dependency, subordinate_word in sentence.dependencies:
        if main_word == word:
            if 'appos' in dependency:
                related_phrase_words.append(subordinate_word)
                related_phrase_words.extend(
                    get_subordinate_words(sentence, subordinate_word))
            elif 'conj' in dependency:
                related_phrase_words.append(subordinate_word)
                related_phrase_words.extend(
                    get_subordinate_words(sentence, subordinate_word))
            elif 'acl' in dependency:
                related_phrase_words.append(subordinate_word)
                subordinate_phrase_words.extend(
                    get_all_dependent_words(sentence, subordinate_word))
            elif 'case' in dependency and is_location(subordinate_word):
                related_phrase_words.append(subordinate_word)
            continue
```

```
        if 'conj' in dependency:
            related_phrase_words.extend(
                get_subordinate_words(sentence, main_word))

    return related_phrase_words
```

Вихідними даними алгоритму є файл, у якому вказано початкове речення та набори слів з виявленими сутностями та розширеною інформацією про сутності, а також набір даних, виявлених у результаті аналізу речення за допомогою алгоритму, що базується на правилах з використанням RE. Приклад вихідних даних наведено на рисунку 5.2.

Може хтось знає, де зараз Петро та сім'я, які проживали **на Нижній** і були **у підвалі.**

Entity
Нижній

Entity extended
на

Locations
на
Нижній
у
підвалі
будинку

Рисунок 5.2 – Приклад вихідних даних

Таким чином, було програмно реалізовано запропоновану систему для розпізнавання іменованих сутностей локацій.

6 МОЖЛИВОСТІ ВПРОВАДЖЕННЯ У НАУКОВІЙ І ПРАКТИЧНІЙ ДІЯЛЬНОСТІ

Запропонована система для розпізнавання іменованих сутностей знайшла своє рішення у статті «Дослідження методів розпізнавання іменованих сутностей у неструктурованому тексті» [26], де було розглянуто проблематику розпізнавання іменованих сутностей, а також піднято питання ефективності поєднання методів машинного навчання та методів, що базуються на правилах, для вирішення задачі розпізнавання іменованих сутностей локацій описового характеру.

Практичне продовження запропонована система може мати при її впровадженні у системи з наступних предметних областей:

– служби доставки продуктів та сервісів;

– служби для екстренної допомоги, зокрема в системи пожежних служб, медичної допомоги, поліцейських служб тощо;

– для аналізу вказання місця подій у новинах та чатах;

– у системах служб таксі.

Дане дослідження має потенціал для розвитку у майбутньому. Одним із відомих аспектів для вивчення, що може у подальшому покращити ефективність роботи системи, є питання розпізнавання просторових прийменників та прийменників часу.

Також, систему можна розширити шляхом використання додаткових аналізаторів тексту, що можуть виявляти сутності місця за ключовими словами.

ВИСНОВКИ

У ході проведення дослідження щодо розпізнавання іменованих сутностей у неструктурованому тексті на прикладі виявлення сутностей локацій та місць було вивчено існуючі засоби та системи розпізнавання іменованих сутностей, їх особливості та запропоновано метод для пошуку іменованих сутностей локацій у неструктурованому тексті українською мовою.

Було розглянуто патенти та наукові роботи, присвячені розпізнаванню іменованих сутностей, у тому числі і локацій, на прикладі іноземних мов, зокрема англійської та португальської. Було виявлено, що у більшості випадків дослідники рекомендують поєднувати у системах розпізнавання іменованих сутностей алгоритми, що базуються на правилах, та ті, що базуються на машинному навчанні.

У рамках даної роботи було досліджено ефективність використання алгоритмів машинного навчання, алгоритмів, що базуються на правилах, а також механізму RE для розпізнавання іменованих сутностей локацій у неструктурованому тексті українською мовою. Було виявлено, що завдяки застосуванню аналізу залежностей між словами у реченні, ефективність розпізнавання іменованих сутностей зростає. Зокрема, було встановлено, що при розпізнаванні іменованих сутностей локацій за допомогою машинного навчання F-міра становить 44.31%, при застосуванні машинного навчання разом з механізмом RE – 65.22%. При цьому при використанні алгоритму, що базується на правилах разом з мезанізмом RE F-міра становить 71.19%. При комбінації обох алгоритмів та механізму RE F-міра досягає 80.79%.

Слід зазначити, що запропонована система є більш ефективною для даних, що були взяті з телеграм-чатів та каналів з новинами, аніж для тих, что були взяті з художньої літератури. Це пов’язано з початковим набором даних, що був використаний для навчання моделей для розпізнавання іменованих сутностей та анотування вхідних даних.

ПЕРЕЛІК ПОСИЛАНЬ

1. What is natural language processing? URL: https://www.ibm.com/cloud/learn/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can (дата звернення 23.02.2022).
2. Software > Stanford Named Entity Recognizer (NER). URL: https://nlp.stanford.edu/software/CRF-NER.html (дата звернення 23.02.2022).
3. Overview of results of the muc-6 evaluation. URL: https://aclanthology.org/X96-1048.pdf (дата звернення 26.02.2022).
4. Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications. 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 (дата звернення 26.02.2022).
5. Нормализация и лемматизация текста с использованием тезауруса. URL: http://www.solarix.ru/for_developers/docs/text-normalization.shtml (дата звернення 26.02.2022).
6. Mariana Dias та ін. Named Entity Recognition for Sensitive Data Discovery in Portuguese. *Applied Sciences*. 2020. pages 1-15.
7. Named Entity Recognition System. URL: https://www.sciencedirect.com/topics/computer-science/named-entity-recognition-system (дата звернення 26.02.2022).
8. A Hidden Markov Model Based Named Entity Recognition System: Bengali and Hindi as Case Studies : веб-сайт. URL: https://link.springer.com/content/pdf/10.1007/978-3-540-77046-6_67.pdf (дата звернення 26.02.2022).
9. EFFICIENT AND ACCURATE NAMED ENTITY RECOGNITION METHOD AND APPARATUS: патент WO2020118741 КНР : МПК G06F 17/27 2006.1. №121846, заявл. 18.12.2018, опубл. 18.06.2020. URL: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2020118741&_cid=P11-L0XNNY-11539-1 (дата звернення 04.03.2022).

10. A METHOD AND SYSTEM FOR AUTOMATED ENTITY RECOGNITION: патент WO2015080558 Малайзія : МПК G06F 17/27 2006.1. №000153, заявл. 18.12.2018, опубл. 18.06.2020. URL: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2015080558&_cid=P11-L0XNNY-11539-1 (дата звернення 05.03.2022).

11. NAME ENTITY RECOGNITION WITH DEEP LEARNING: патент WO2020193966 Великобританія : МПК G06F 40/295 2020.1. №050779, заявл. 23.03.2020, опубл. 01.10.2020. URL: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2020193966&_cid=P11-L0XNNY-11539-1 (дата звернення 11.03.2022).

12. Location Named-Entity Recognition using Rule-Based Approach for Balinese Texts. URL: https://www.researchgate.net/publication/349518820_Location_Named-Entity_Recognition_using_Rule-Based_Approach_for_Balinese_Texts (дата звернення 17.03.2022).

13. Panchenko D., Maksymenko D., Turuta O., Luzan M., Tytarenko S., Turuta O. Ukrainian News Corpus As Text Classification Benchmark // Proceedings of the 17th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops, P. 717-726.

14. NER-анотація українського корпусу. URL: https://github.com/lang-uk/ner-uk (дата звернення 17.03.2022)

15. Дашенков Д. С. Дослідження методів автоматичної відповіді на запитання. URL: https://openarchive.nure.ua/bitstream/document/18861/1/2021_M_PI_Dashenkov_DS.pdf (дата звернення 22.03.2022)

16. Дашенков Д. С., Турута О. П. Improving question answering system for Ukrainian language // Інформаційні технології і безпека. Матеріали XX Міжнародної науково-практичної конференції ІТБ-2020. – Київ, 2020. No 20. сс. 125-130.

17. Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning. URL: https://doi.org/10.1613/jair.1.12918 (дата звернення 15.04.2022)

18. Universal features. URL: https://universaldependencies.org/u/feat/index.html (дата звернення 02.04.2022)

19. Universal Dependencies. URL: https://universaldependencies.org/ (дата звернення 02.04.2022)

20. Українська мова та література. Прийменник. URL: https://zno.if.ua/?p=2716 (дата звернення 02.04.2022)

21. Dependencies. URL: https://universaldependencies.org/ru/dep/ (дата звернення 03.04.2022)

22. Браунський корпус української мови. URL: https://github.com/brown-uk/corpus (дата звернення 03.04.2022)

23. spaCy. URL: https://spacy.io/ (дата звернення 08.04.2022)

24. NLTK. Documentation. URL: https://www.nltk.org/ (дата звернення 09.04.2022)

25. Stanza. Named Entity Recognition. URL: https://stanfordnlp.github.io/stanza/ner.html (дата звернення 09.04.2022)

26. Люліна К. П., Турута О. П. Дослідження методів розпізнавання іменованих сутностей у неструктурованому тексті // Сучасні напрями розвитку інформаційно-коммунікаційних технологій та засобів управління. Тези доповідей дванадцатої міжнародної науково-технічної конференції 27 – 28 квітня 2022 року. – Баку – Харків – Жиліна, 2022, т. 2. с. 143.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

13. Panchenko D., Maksymenko D., Turuta O., Luzan M., Tytarenko S., Turuta O. Ukrainian News Corpus As Text Classification Benchmark // Proceedings of the 17th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops, P. 717-726.

15. Дашенков Д. С. Дослідження методів автоматичної відповіді на запитання. URL: https://openarchive.nure.ua/bitstream/document/18861/1/2021_M_PI_Dashenkov_DS.pdf (дата звернення 22.03.2022)

16. Дашенков Д. С., Турута О. П. Improving question answering system for Ukrainian language // Інформаційні технології і безпека. Матеріали XX Міжнародної науково-практичної конференції ІТБ-2020. – Київ, 2020. No 20. сс. 125-130.

17. Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning. URL: https://doi.org/10.1613/jair.1.12918 (дата звернення 15.04.2022)

26. Люліна К. П., Турута О. П. Дослідження методів розпізнавання іменованих сутностей у неструктурованому тексті // Сучасні напрями розвитку інформаційно-коммунікаційних технологій та засобів управління. Тези доповідей дванадцатої міжнародної науково-технічної конференції 27 – 28 квітня 2022 року. – Баку – Харків – Жиліна, 2022, т. 2. с. 143.