Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
(повна назва)

Кафедра програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів розробки лінгвістичних ресурсів перетворення корпусів текстів
(тема)

Виконав:
студент (ка) 2 курсу, групи ІПЗм-22-6

Коваленко Є.В.
(прізвище, ініціали)

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

Тип програми освітньо-наукова

Керівник проф. Шубін І.Ю.
(посада, прізвище, ініціали)

Допускається до захисту
Зав. кафедри _____ З.В.Дудар
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук

Кафедра програмної інженерії

Рівень вищої освіти другий (магістерський)

Спеціальність 121 – Інженерія програмного забезпечення

Тип програми освітньо-наукова програма

Освітня програма Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«___» _____ 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Коваленку Євгенію Владиславовичу
(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження методів розробки лінгвістичних ресурсів перетворення корпусів текстів»

Затверджена наказом по університету від «29» березня 2024 р. №250 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 12.06.2024 р.

3. Вихідні дані до роботи: *інформаційний пошук, корпуси текстів*

4. Перелік питань, що потрібно опрацювати в роботі: *аналіз існуючих методів методів розробки лінгвістичних ресурсів та перетворення корпусів текстів*

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналіз предметної галузі та постановка задачі	30.03 – 14.04.24	*виконано*
2	Аналіз та вибір API для дослідження	15.04 – 24.04.24	*виконано*
3	Аналіз та моделювання предметної області	25.04 – 28.04.24	*виконано*
4	Планування експериментів	29.04 – 08.05.24	*виконано*
5	Програмна реалізація кожного з обраних для дослідження API	08.05 – 18.05.24	*виконано*
6	Експериментальні дослідження	19.05 – 30.05.24	*виконано*
7	Аналіз результатів експериментальних досліджень та розробка рекомендацій	01.06 – 03.06.24	*виконано*
8	Написання та оформлення статті та тез доповіді	03.06 – 05.06.24	*виконано*
9	Підготовка пояснювальної записки	06.06 – 09.06.24	*виконано*
10	Підготовка презентації та доповіді	09.06 – 11.06.24	*виконано*
11	Нормоконтроль	12.06 – 14.06.24	*виконано*
12	Рецензування	15.06 – 17.06.24	*виконано*
13	Занесення диплома в електронний архів	17.06.2024	*виконано*
14	Попередній захист	18.06.2024	*виконано*
15	Допуск до захисту у зав. кафедри	20.06.2024	*виконано*

Дата видачі завдання 30 березня 2024 р.

Студент _____ Коваленко Є.В.
(підпис)

Керівник кваліфікаційної роботи _____ проф. Шубін І.Ю.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до кваліфікаційної роботи, 51 с., 4 рис., 20 джерел, 5 додатків.

АНАЛІЗ ДАНИХ, КОРПУСИ ТЕКСТІВ, ЛІНГВІСТИЧНІ РЕСУРСИ, ОБРОБКА ДАНИХ, ПЕРЕТВОРЕННЯ ТЕКСТУ.

Об'єктом дослідження є методи та процеси розробки лінгвістичних ресурсів, зокрема механізми та інструменти для ефективного перетворення корпусів текстів.

Мета проекту полягає у розробці та дослідженні методів, які б підвищили ефективність обробки текстів і забезпечили більш точну та швидку роботу з лінгвістичними корпусами.

У процесі дослідження використовуються методи обробки природної мови, математичного моделювання, а також аналізу великих обсягів даних.

Результати проекту передбачають розробку рекомендацій та методичних вказівок для створення та оптимізації лінгвістичних ресурсів.

DATA ANALYSIS, TEXT CORPORATIONS, LINGUISTIC RESOURCES, DATA PROCESSING, TEXT TRANSFORMATION.

The object of research is the methods and processes of developing linguistic resources, in particular the mechanisms and tools for the effective transformation of text corpora.

The goal of the project is to develop and research methods that would increase the efficiency of text processing and ensure more accurate and faster work with linguistic corpora.

The results of the project provide for the development of recommendations and methodological guidelines for the creation and optimization of linguistic resources.

Я, Коваленко Євгеній Владиславович, студент гр. ІПЗм-22-6, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя робота на тему «Дослідження методів розробки лінгвістичних ресурсів перетворення корпусів текстів», що буде представлена в екзаменаційну комісію 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(на) 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

ВСТУП

В епоху інформаційних технологій, зі зростанням доступності та обсягів текстових даних, розробка та вдосконалення лінгвістичних ресурсів набуває особливої актуальності. Лінгвістичні ресурси – це інструменти та дані, які використовуються для обробки, аналізу та перетворення текстової інформації. Вони мають широке застосування в таких областях як обробка природної мови, комп'ютерний лінгвістичний аналіз, автоматизований переклад та багато інших. Завдяки цьому, ефективне перетворення корпусів текстів стає ключовою умовою для розвитку багатьох сфер, що базуються на знаннях та інформації.

Проект "Дослідження методів розробки лінгвістичних ресурсів перетворення корпусів текстів" спрямований на аналіз існуючих та розробку нових підходів до створення та оптимізації лінгвістичних ресурсів. Головна увага приділяється методам, які дозволяють підвищити ефективність перетворення текстових даних, забезпечуючи їхню високу точність, швидкість обробки та адаптивність до різноманітних завдань.

У рамках цього дослідження буде здійснений огляд сучасних технологій та методологій, визначені найбільш ефективні стратегії обробки даних, а також розроблені рекомендації та вказівки для подальшого розвитку і вдосконалення лінгвістичних ресурсів. Особлива увага приділяється таким аспектам, як масштабованість, універсальність та можливість інтеграції розроблених ресурсів у різноманітні інформаційні системи.

Цей проект має на меті не лише поглибити розуміння існуючих методів розробки лінгвістичних ресурсів, але й стимулювати розвиток нових технологій та підходів, які сприятимуть прогресу в обробці та аналізі текстових даних. Таким чином, робота спрямована на збагачення інструментарію лінгвістики та інформаційних технологій, а також на підвищення якості та доступності інформаційних послуг.

1 ФУНДАМЕНТАЛЬНИЙ АНАЛІЗ ЛІНГВІСТИЧНИХ РЕСУРСІВ ТА МЕТОДІВ ЇХ РОЗРОБКИ

1.1 Історія та розвиток лінгвістичних ресурсів

Лінгвістичні ресурси, відомі своєю значимістю у мовознавстві та обробці природної мови, мають довгу та багату історію. Їх розвиток розпочався з основних друкованих словників та граматик, які вперше стали доступними завдяки винаходу друкарства в XV столітті [1]. Ці перші ресурси були спрямовані на документування та систематизацію мовних знань, що дозволяло людям навчатися і використовувати мову більш ефективно.

В епоху Ренесансу словники, такі як "Лексикон" Роберта Естена та "Лексикографія" Еліаса Хоймера, почали охоплювати не лише лексичний склад, але й пропонували пояснення і приклади використання слів у різних контекстах. Це був важливий крок у напрямку більш детального та глибокого розуміння мовних явищ [2]. Граматики, на кшталт "Граматики французької мови" Роберта Естена, також стали більш детальними, зосереджуючись на структурі та правилах мови, що сприяло їх кращому розумінню та навчанні.

З розвитком технологій у XIX столітті почалося створення більш складних лінгвістичних ресурсів. Важливим досягненням цього періоду стало видання Оксфордського англійського словника (Oxford English Dictionary), який зібрав і систематизував величезну кількість слів англійської мови разом з їх історичними варіантами та прикладами використання. Це видання стало золотим стандартом у словниковій роботі і мало значний вплив на подальший розвиток лексикографії.

Перехід до електронної ери надав новий імпульс у розвитку лінгвістичних ресурсів. У 1960-х роках з’явився перший електронний корпус текстів – Brown Corpus, який збирався на основі друкованих текстів і використовувався для лінгвістичного аналізу [3]. Це було початком нового етапу в обробці текстів, коли комп'ютери почали активно використовуватися для зберігання, обробки та аналізу великих обсягів мовних даних.

У 1980-х роках виникло кілька важливих проектів, які стали ключовими у розвитку електронних корпусів. Наприклад, Британський національний корпус (British National Corpus) зібрав великий обсяг текстів, що охоплюють різноманітні жанри та стилі сучасної англійської мови. Цей корпус став важливим інструментом для дослідників і викладачів, надаючи їм можливість аналізувати мовні явища в реальному контексті.

З початку XXI століття розвиток лінгвістичних ресурсів отримав новий імпульс завдяки інтернету та великим обсягам цифрових даних. Сучасні лінгвістичні ресурси, такі як Google Ngram Viewer, дозволяють дослідникам аналізувати зміни у використанні слів і фраз у мільйонах книг протягом багатьох століть. Інші проекти, як-от Universal Dependencies, спрямовані на створення уніфікованих синтаксичних описів для багатьох мов світу, що значно сприяє розвитку мультилингвістичних досліджень та обробки природної мови [4].

Сьогодні лінгвістичні ресурси продовжують розвиватися з неймовірною швидкістю. Використання штучного інтелекту та машинного навчання дозволяє автоматизувати процеси збору та аналізу мовних даних, створюючи нові можливості для досліджень. Такі ресурси, як OpenAI GPT, здатні генерувати тексти та проводити складні лінгвістичні аналізи, що відкриває нові горизонти у вивченні та застосуванні мов.

1.2 Класифікація та види лінгвістичних ресурсів

Лінгвістичні ресурси є основними інструментами для зберігання, аналізу та використання мовних даних. Вони охоплюють широкий спектр різних видів, кожен з яких має своє призначення, структуру та методи використання. Класифікація лінгвістичних ресурсів дозволяє краще зрозуміти їх функції та сфери застосування, що є важливим для ефективного використання в лінгвістичних дослідженнях та прикладних задачах.

Однією з основних категорій лінгвістичних ресурсів є словники. Вони можуть бути загальними або спеціалізованими, охоплюючи широкий спектр мовних явищ. Загальні словники, такі як "Великий Оксфордський Словник",

зосереджуються на широкому наборі слів і фраз з поясненням їх значень, походження та використання [4]. Спеціалізовані словники, наприклад, технічні або медичні, фокусуються на термінології конкретних галузей, надаючи глибоке розуміння вузькоспеціалізованих лексичних одиниць. Словники тезауруси, такі як "Роджеровий Тезаурус", організовують слова за синонімічними групами, що допомагає користувачам знаходити синоніми та розширювати свої лексичні навички.

Граматики та фонетичні бази є іншими важливими видами лінгвістичних ресурсів. Граматики описують структуру мов, включаючи правила побудови речень, синтаксичні і морфологічні закономірності. Вони є невід'ємною частиною мовного навчання та лінгвістичного аналізу. Фонетичні та фонологічні бази даних, такі як International Phonetic Alphabet (IPA), зберігають інформацію про звуки мов, включаючи їх артикуляцію та акустичні властивості [5]. Ці ресурси є критичними для досліджень у галузі фонетики та фонології, а також для розробки технологій розпізнавання та синтезу мовлення.

Морфологічні бази даних зосереджуються на структурі слів, їх формах та змінах. Вони охоплюють правила утворення слів, таких як відмінювання дієслів чи зміни у словоформах для відображення часу, числа чи роду. Наприклад, база даних WordNet не лише надає інформацію про слова, але й описує відношення між ними, такі як синонімія, антонімія та гіпонімія, що робить її корисною для широкого кола лінгвістичних досліджень.

Синтаксичні та семантичні корпуси є ще однією важливою категорією лінгвістичних ресурсів. Вони включають розмічені тексти, що містять інформацію про граматичні структури та значення слів у контексті. Синтаксичні корпуси, такі як Penn Treebank, надають детальні анотації синтаксичних структур, що дозволяє дослідникам аналізувати граматичні властивості мови. Семантичні корпуси, як-от PropBank, розширюють цей підхід, надаючи анотації, що описують відношення між словами та їх ролі у реченнях.

Важливу роль також відіграють паралельні корпуси, які містять тексти на різних мовах та їх переклади. Ці ресурси, як-от Europarl Corpus, який збирає

документи Європейського Парламенту, є надзвичайно корисними для задач машинного перекладу та багатомовного аналізу. Вони дозволяють дослідникам і розробникам створювати та тестувати моделі перекладу, що здатні працювати з багатьма мовами одночасно.

Корпуси текстів є ключовими лінгвістичними ресурсами, що складаються з великих зібрань текстів, організованих для аналізу мовних даних. Вони можуть бути класифіковані за мовою, жанром, темою або часом створення текстів. Наприклад, корпуси, як-от British National Corpus, охоплюють широкий спектр жанрів та стилів англійської мови, забезпечуючи дослідникам широкий контекст для аналізу. Інші корпуси, такі як COCA (Corpus of Contemporary American English), постійно оновлюються і включають тексти з різних джерел, від газет до наукових журналів, що дозволяє відстежувати сучасні мовні тенденції.

1.3 Сучасні методи та технології в розробці лінгвістичних ресурсів

Розвиток лінгвістичних ресурсів значно прискорився завдяки використанню сучасних методів та технологій. Зростання обчислювальних потужностей, а також удосконалення алгоритмів обробки природної мови (NLP) та машинного навчання (ML) відкрили нові можливості для створення, зберігання та аналізу великих обсягів мовних даних. Цей розділ розгляне ключові методи та технології, що вплинули на сучасну розробку лінгвістичних ресурсів.

Однією з основних технологій, що сприяли революції в обробці мовних даних, є глибинне навчання. Алгоритми глибинного навчання, такі як рекурентні нейронні мережі (RNN) та трансформери, значно покращили якість обробки текстів. Наприклад, модель трансформера GPT (Generative Pre-trained Transformer), розроблена OpenAI, здатна генерувати тексти, що майже не відрізняються від текстів, написаних людиною. Ці моделі використовуються для створення синтаксично та семантично насичених текстів, що значно покращує роботу з корпусами текстів.

Автоматичне анотування текстів стало можливим завдяки прогресу в галузі машинного навчання. Методи автоматичного анотування, такі як використання

CRF (Conditional Random Fields) або нейронних мереж, дозволяють ефективно розмічати текстові дані синтаксичною, семантичною та іншими типами анотацій. Це значно скорочує час і зусилля, необхідні для створення високоякісних лінгвістичних ресурсів. Наприклад, проект Universal Dependencies використовує автоматичні анотатори для створення уніфікованих синтаксичних розміток для різних мов, що дозволяє дослідникам аналізувати синтаксичні структури на багатомовному рівні.

Важливу роль у сучасній обробці мовних даних відіграє статистична обробка природної мови. Статистичні методи, такі як моделі на основі прихованих марковських процесів (HMM) та байєсові мережі, використовуються для побудови систем розпізнавання мовлення, машинного перекладу та інших додатків NLP. Ці методи дозволяють ефективно моделювати ймовірності різних мовних явищ та адаптуватися до різних контекстів використання. Наприклад, системи машинного перекладу, такі як Google Translate, використовують статистичні моделі для покращення точності перекладів між різними мовами, спираючись на величезні обсяги двомовних текстів.

Корпуси текстів та їх автоматичне збагачення також є важливими аспектами сучасних лінгвістичних досліджень. Сучасні технології дозволяють автоматично збирати та оновлювати корпуси текстів з різних джерел, таких як веб-сайти, соціальні мережі та цифрові архіви. Ці корпуси можуть бути збагачені за допомогою методів автоматичного анотування, що включають морфологічний, синтаксичний та семантичний аналіз. Наприклад, корпуси, створені з даних соціальних мереж, дозволяють дослідникам аналізувати сучасні мовні тенденції та культурні зміни у реальному часі.

Інтеграція лінгвістичних ресурсів з іншими системами стала можливою завдяки розвитку технологій API (Application Programming Interface). Використання API дозволяє розробникам інтегрувати лінгвістичні інструменти та дані з іншими додатками та системами, що робить їх доступними для широкого кола користувачів. Наприклад, API Google Cloud Natural Language дозволяє

розробникам використовувати потужні інструменти аналізу текстів у своїх додатках, забезпечуючи автоматичну обробку та аналіз мовних даних.

Нарешті, штучний інтелект та машинне навчання відкривають нові можливості для автоматизації та покращення лінгвістичних досліджень. Вони дозволяють створювати інтелектуальні системи, які можуть навчатися на основі великих обсягів даних та автоматично адаптуватися до нових мовних явищ. Ці технології не тільки покращують точність і швидкість обробки мовних даних, але й сприяють розвитку нових методів і підходів у лінгвістичних дослідженнях.

Таким чином, сучасні методи та технології значно розширили можливості розробки лінгвістичних ресурсів. Використання глибинного навчання, обробки великих даних, статистичних методів, нейронно-мовних моделей, автоматичного перекладу, збагачення корпусів, інтеграції API та штучного інтелекту створює нові перспективи для розвитку лінгвістики та обробки природної мови. Ці інновації продовжують формувати майбутнє мовних досліджень та технологій.

2 СТРАТЕГІЇ ТА ПІДХОДИ В РОЗРОБЦІ ЛІНГВІСТИЧНИХ РЕСУРСІВ

2.1 Планування та розробка лінгвістичних ресурсів

Розробка лінгвістичних ресурсів є багатоступеневим процесом, який вимагає глибокого розуміння мовних потреб та технологічних можливостей. Кожен етап, від початкового визначення цілей проекту до впровадження кінцевого продукту, вимагає ретельного планування і врахування багатьох аспектів.

Перший крок у цьому процесі полягає у чіткому визначенні цілей проекту. Цілі визначають напрямок і масштаб розробки, включаючи типи лінгвістичних ресурсів, які необхідно створити, такі як корпуси текстів, лексичні бази або граматичні моделі. Важливо чітко розуміти, для чого ці ресурси будуть використовуватися. Наприклад, корпус текстів може бути створений для лінгвістичного аналізу, навчання моделей машинного навчання або підтримки систем автоматичного перекладу [6]. Таке визначення цілей дозволяє встановити чіткі критерії успішності та окреслити основні вимоги до розробки.

Після встановлення цілей важливо детально визначити обсяг проекту. Це включає оцінку обсягу даних, які будуть зібрані, структуру ресурсів, які будуть створені, і типи анотацій, що будуть застосовані. Наприклад, при розробці корпусу текстів необхідно вирішити, які жанри і стилі текстів будуть включені, яка їхня географічна та часова репрезентативність, а також які типи лінгвістичної інформації будуть зібрані та розмічені. Визначення обсягу дозволяє краще планувати ресурсні витрати та уникати перевантаження проекту.

Наступним критично важливим кроком є аналіз потреб кінцевих користувачів. Це допомагає зрозуміти, хто буде користуватися розробленими ресурсами і які саме функціональні можливості вони очікують. Наприклад, дослідники можуть потребувати детально анотованих корпусів для проведення лінгвістичних досліджень, тоді як розробники програмного забезпечення можуть зосереджуватися на ресурсах, які легко інтегруються з іншими системами. Аналіз потреб користувачів дозволяє створювати ресурси, які дійсно відповідають їхнім очікуванням та вимогам, і підвищує цінність кінцевого продукту.

Паралельно з аналізом потреб користувачів слід провести огляд існуючих лінгвістичних ресурсів. Це дозволяє виявити, які з доступних ресурсів можна використовувати або адаптувати для потреб проекту. Наприклад, замість створення нового корпусу з нуля, може бути доцільним використати наявні корпуси, якщо вони відповідають вимогам проекту. Аналіз існуючих ресурсів допомагає ефективніше використовувати наявні ресурси та знижує витрати на розробку.

Важливим аспектом планування є вибір методологій та технологій. Сучасні технології, такі як машинне навчання, автоматичне анотування та обробка великих даних, дозволяють значно покращити ефективність створення та обробки лінгвістичних ресурсів. Наприклад, автоматичне анотування текстів з використанням нейронних мереж може значно прискорити процес анотації та підвищити якість анотацій. Вибір технологій також включає визначення платформ та мов програмування, які забезпечать необхідну продуктивність та гнучкість. Наприклад, використання платформ з відкритим кодом може забезпечити більшу гнучкість та знижує витрати на розробку.

2.2 Технологічні аспекти розробки

Успішна розробка лінгвістичних ресурсів значною мірою залежить від правильного вибору технологій та інструментів, які забезпечують ефективне створення, обробку та аналіз мовних даних. У цьому розділі ми розглянемо ключові технологічні аспекти, які відіграють вирішальну роль у розробці сучасних лінгвістичних ресурсів.

Розпочнемо з обговорення платформ та інструментів для зберігання та обробки даних. Однією з головних задач при розробці лінгвістичних ресурсів є ефективне зберігання великих обсягів текстових даних та забезпечення доступу до них. Вибір платформи для зберігання даних значно впливає на продуктивність та масштабованість системи. Сучасні бази даних, такі як MongoDB та Elasticsearch, пропонують потужні можливості для зберігання та пошуку текстових даних, що робить їх популярним вибором для розробки лінгвістичних

ресурсів. MongoDB, зокрема, забезпечує масштабоване зберігання документів у форматі JSON, що є ідеальним для зберігання структурованих текстових даних. Elasticsearch, у свою чергу, забезпечує потужні можливості пошуку та аналізу текстів у реальному часі, що є важливим для роботи з великими корпусами.

Обробка текстових даних потребує використання різних програмних інструментів, таких як платформи для обробки природної мови (NLP). Одним із провідних інструментів у цій області є SpaCy, яка пропонує потужні засоби для токенізації, частеречевої розмітки, іменної розпізнавання сутностей та інших завдань NLP. Набір інструментів NLTK (Natural Language Toolkit) також широко використовується для навчання та досліджень у галузі обробки природної мови, забезпечуючи багатий набір функцій для текстової обробки та аналізу [7]. Ці платформи дозволяють ефективно виконувати складні завдання, пов'язані з обробкою текстів, такі як морфологічний та синтаксичний аналіз, а також забезпечують гнучкість для адаптації до специфічних вимог проекту.

Розвиток технологій глибинного навчання привів до створення потужних нейронних мережевих моделей, таких як BERT, GPT та ELMo, які значно покращили якість обробки текстів. Ці моделі використовуються для завдань, які потребують розуміння контексту та семантики текстів, таких як автоматичний переклад, генерація текстів та розпізнавання емоцій. BERT (Bidirectional Encoder Representations from Transformers), наприклад, використовує двонаправлене навчання для кращого розуміння контексту слів у реченні, що забезпечує високу точність у завданнях класифікації текстів та аналізу настроїв. GPT (Generative Pre-trained Transformer) від OpenAI демонструє вражаючу здатність до генерації зв'язного тексту, що майже не відрізняється від людського, і широко використовується для створення чат-ботів та інших додатків, що потребують генерації природного мовлення.

Важливим аспектом є автоматизація процесу анотації текстових даних. Інструменти, такі як Brat та Prodigy, дозволяють значно спростити процес анотації, забезпечуючи інтерактивні інтерфейси для маркування текстів [8]. Ці інструменти підтримують різні типи анотацій, включаючи синтаксичні,

семантичні та лексичні, і дозволяють користувачам легко додавати або редагувати анотації. Prodigy, наприклад, використовує методи активного навчання для оптимізації процесу анотації, зменшуючи кількість даних, які потрібно вручну анотовані, шляхом автоматичного вибору найбільш інформативних прикладів для маркування.

Ще одним критичним аспектом є інтеграція різних лінгвістичних ресурсів з існуючими інформаційними системами. Використання API (Application Programming Interface) дозволяє легко інтегрувати лінгвістичні інструменти та ресурси з іншими додатками, забезпечуючи доступ до потужних можливостей обробки мовних даних. Наприклад, API Google Cloud Natural Language надає можливість аналізу текстів на основі машинного навчання, включаючи визначення сутностей, аналіз настроїв та розпізнавання синтаксичних структур. Це дозволяє розробникам легко інтегрувати передові технології обробки природної мови у свої додатки.

Технологічний аспект розробки також включає забезпечення масштабованості та продуктивності системи. При роботі з великими обсягами даних важливо забезпечити, щоб система могла ефективно обробляти дані без затримок або збоїв. Використання хмарних обчислювальних платформ, таких як AWS або Google Cloud, дозволяє масштабувати обчислювальні ресурси відповідно до потреб проекту, забезпечуючи високу доступність та надійність системи. Це особливо важливо при роботі з великими корпусами текстів або при реалізації додатків з високими вимогами до обчислювальної потужності.

2.3 Впровадження та використання ресурсів

Впровадження та використання лінгвістичних ресурсів є завершальним, але надзвичайно важливим етапом у їхньому життєвому циклі. Це етап, на якому забезпечується доступність, зручність використання та інтеграція ресурсів у різні інформаційні системи. Успішне впровадження гарантує, що ресурси будуть ефективно використовуватися кінцевими користувачами, включаючи дослідників, розробників програмного забезпечення та інші зацікавлені сторони.

Процес впровадження розпочинається з розгортання ресурсів на відповідних платформах. Це може включати хмарні сервіси, локальні сервери або інтеграцію з існуючими інформаційними системами. Вибір платформи залежить від конкретних потреб користувачів, обсягу даних та вимог до продуктивності. Наприклад, хмарні платформи, такі як Amazon Web Services або Google Cloud, пропонують гнучкість і масштабованість, що робить їх ідеальними для розгортання великих корпусів текстів або інших лінгвістичних ресурсів. Ці платформи дозволяють автоматично масштабувати обчислювальні ресурси відповідно до змін у навантаженні, що гарантує стабільну роботу системи навіть при значних коливаннях у використанні.

Після розгортання ресурси мають бути доступними та зручними для використання. Це означає розробку інтуїтивно зрозумілих інтерфейсів користувача, які дозволяють легко взаємодіяти з ресурсами. Наприклад, веб-інтерфейси для доступу до корпусів текстів або лексичних баз повинні бути простими у використанні і забезпечувати можливості для пошуку, фільтрації та аналізу даних. Інтерактивні візуалізації, такі як графіки або діаграми, можуть значно покращити розуміння користувачами складних мовних даних. Важливо також забезпечити підтримку різних форматів даних та можливість експорту результатів для подальшого використання у інших додатках.

Документація та навчальні матеріали відіграють важливу роль у забезпеченні успішного впровадження. Вони забезпечують користувачів необхідною інформацією для ефективного використання ресурсів. Документація повинна бути детальною, з інструкціями щодо встановлення, конфігурації та використання ресурсів, а також з прикладами коду та типових сценаріїв використання. Наприклад, керівництва користувача для API мають включати опис всіх доступних кінцевих точок, формати запитів та відповідей, а також приклади використання для різних мов програмування [9]. Крім цього, можуть бути створені відеоуроки, вебінари або інтерактивні посібники, що допомагають користувачам швидко освоїти роботу з новими інструментами.

Підтримка користувачів є ще одним важливим аспектом впровадження. Це включає надання технічної підтримки, відповіді на питання користувачів та вирішення проблем, що виникають у процесі використання ресурсів. Ефективна підтримка може значно підвищити задоволеність користувачів та сприяти більш широкому використанню ресурсів. Наприклад, підтримка може надаватися через спеціалізовані служби підтримки, форуми або системи обробки заявок, що дозволяє користувачам отримувати допомогу у реальному часі або звертатися до спільноти за порадою.

Зворотний зв'язок з користувачами є критичним для виявлення можливих проблем та потреб в удосконаленні ресурсів. Це може включати регулярні опитування, форми зворотного зв'язку або спеціальні зустрічі з користувачами, які дозволяють зібрати інформацію про їхній досвід використання ресурсів та виявити можливі напрямки для покращення. Наприклад, аналіз зворотного зв'язку може виявити необхідність у додаткових функціях або вдосконаленнях інтерфейсу, що дозволяє краще відповідати потребам користувачів та підвищити ефективність роботи з ресурсами.удосконаленні. Також важливою є оцінка впливу ресурсів на розвиток галузі, наукові дослідження та практичні застосування.

2.4 Проблеми та виклики у розробці ресурсів

Однією з головних технічних проблем є якість даних. Для створення корисних і точних лінгвістичних ресурсів необхідно забезпечити високу якість вхідних даних. Це включає точність анотацій, відповідність даних реальним мовним явищам та узгодженість між різними частинами ресурсу. Наприклад, при створенні корпусу текстів важливо забезпечити, щоб всі тексти були коректно анотовані та структуровані. Проте забезпечення такої якості може бути складним завданням, особливо при роботі з великими обсягами даних або з різними мовами. Автоматизовані інструменти для анотації можуть допомогти скоротити час і зусилля, але вони не завжди можуть забезпечити необхідну точність, що вимагає додаткової ручної перевірки та корекції.

Ще однією серйозною технічною проблемою є масштабованість систем. При роботі з великими лінгвістичними ресурсами необхідно забезпечити, щоб система могла обробляти та зберігати великі обсяги даних без зниження продуктивності. Це особливо важливо для хмарних сервісів або веб-додатків, які повинні обробляти запити користувачів у реальному часі. Використання хмарних обчислень може допомогти вирішити цю проблему, надаючи можливість масштабувати обчислювальні ресурси відповідно до потреб проекту. Проте це також потребує ретельного планування і управління ресурсами, щоб уникнути перевантаження системи і забезпечити стабільну роботу.

Організаційні виклики є ще одним важливим аспектом у розробці лінгвістичних ресурсів. Координація роботи різних команд і управління проектом можуть бути складними завданнями, особливо якщо проект включає в себе міжнародні команди або багато різних завдань. Наприклад, проекти зі створення багатомовних корпусів часто потребують співпраці дослідників і анотувальників з різних країн, що може призвести до труднощів у комунікації та узгодженні робочих процесів. Впровадження гнучких методологій розробки, таких як Scrum або Kanban, може допомогти покращити координацію і підвищити ефективність роботи команд, забезпечуючи регулярні зустрічі, чітке планування та постійний зворотний зв'язок.

Етичні питання є невід'ємною частиною розробки лінгвістичних ресурсів, особливо якщо мова йде про роботу з особистими або конфіденційними даними. Забезпечення конфіденційності та захисту персональних даних є критичним аспектом, що вимагає ретельного дотримання нормативних вимог і етичних стандартів [10]. Наприклад, при створенні корпусів текстів, що містять особисту інформацію, необхідно забезпечити її анонімізацію або отримання згоди від власників даних. Також важливо впроваджувати протоколи безпеки, такі як шифрування даних і контроль доступу, щоб захистити дані від несанкціонованого доступу та зловживань.

Фінансові аспекти також можуть створювати значні виклики у розробці лінгвістичних ресурсів. Розробка високоякісних ресурсів часто потребує значних

фінансових інвестицій, включаючи витрати на обладнання, програмне забезпечення, оплату праці анотувальників та дослідників. Залучення фінансування може бути складним завданням, особливо для академічних або некомерційних проектів. Важливо ретельно планувати бюджет проекту і ефективно використовувати наявні ресурси. Співпраця з іншими організаціями або пошук зовнішнього фінансування, наприклад, через гранти або інвестиції, може допомогти знизити фінансовий тиск і забезпечити успішне завершення проекту.

Важливо також враховувати виклики, пов'язані з підтримкою та оновленням ресурсів після їхнього впровадження. Лінгвістичні ресурси мають бути актуальними і постійно оновлюватися для відображення сучасних мовних тенденцій і потреб користувачів. Це вимагає постійного моніторингу, збору зворотного зв'язку від користувачів і регулярного оновлення даних. Наприклад, корпуси текстів мають постійно оновлюватися новими текстами, щоб відображати поточні зміни в мові. Це потребує значних ресурсів і координації, але є критично важливим для забезпечення довгострокової корисності та релевантності ресурсів.

Сучасні технології, такі як штучний інтелект і машинне навчання, відкривають нові можливості для автоматизації процесів розробки та підтримки лінгвістичних ресурсів. Проте їх впровадження також може створити додаткові виклики, пов'язані з необхідністю забезпечення високої точності та надійності автоматизованих систем.

Вирішення цих викликів вимагає ретельного планування, гнучкого підходу до управління проектами та ефективного використання сучасних технологій. Забезпечення високої якості даних, масштабованості систем, координації роботи команд, дотримання етичних стандартів і ефективного управління фінансовими ресурсами є ключовими аспектами успішної розробки лінгвістичних ресурсів. Прийняття стратегій для подолання цих викликів дозволяє створювати високоякісні, корисні та актуальні ресурси, що сприяють розвитку мовознавства та обробки природної мови.

3 РОЗРИШЕНИЙ АНАЛІЗ ТА ЗАСТОСУВАННЯ КОРПУСІВ ТЕКСТІВ

3.1 Значення та види корпусів текстів

Корпуси текстів є основними інструментами у лінгвістичних дослідженнях і прикладних завданнях обробки природної мови. Вони дозволяють дослідникам аналізувати велику кількість текстових даних, виявляти мовні закономірності та тенденції, а також навчати та тестувати алгоритми машинного навчання. Важливість корпусів текстів полягає в їхній здатності відображати реальні мовні практики та забезпечувати основу для різноманітних мовних досліджень. У цьому розділі ми розглянемо різні види корпусів текстів та їх значення у сучасних лінгвістичних дослідженнях.

Корпуси текстів можна класифікувати за кількома основними ознаками. Однією з найпоширеніших класифікацій є поділ за мовами, що включає моноязичні, білінгвальні та мультиязичні корпуси. Моноязичні корпуси містять тексти однієї мови і використовуються для аналізу лексичних, граматичних та стилістичних особливостей цієї мови. Наприклад, British National Corpus (BNC) представляє сучасну англійську мову у всіх її формах і жанрах, надаючи дослідникам широкий спектр текстів для аналізу [11]. Білінгвальні корпуси, які включають тексти на двох мовах і їхні переклади, є надзвичайно корисними для задач машинного перекладу. Вони дозволяють дослідникам і розробникам порівнювати структури різних мов і покращувати якість перекладу. Мультиязичні корпуси, які містять тексти на багатьох мовах, є важливими для багатомовного аналізу і розробки систем, що працюють з багатьма мовами одночасно. Прикладом є Europarl Corpus, що збирає документи Європейського Парламенту на різних європейських мовах.

Іншим важливим критерієм класифікації є поділ за часом, що включає діахронічні та синхронічні корпуси. Діахронічні корпуси збирають тексти, написані у різні історичні періоди, що дозволяє дослідникам вивчати зміни в мові з плином часу. Вони є незамінними для історичної лінгвістики, оскільки дозволяють аналізувати еволюцію лексики, граматики та стилістики. Наприклад, Corpus of Historical American English (COHA) охоплює понад 400 мільйонів слів

американської англійської мови з 1810 до сучасності і є цінним ресурсом для вивчення мовних змін у американській англійській. Синхронічні корпуси, навпаки, збирають тексти, що відображають сучасний стан мови в певний момент часу. Вони дозволяють дослідникам аналізувати актуальні мовні явища і тенденції, що існують у сучасному суспільстві.

Жанрова класифікація також відіграє важливу роль. Корпуси можуть включати тексти різних жанрів, таких як літературні твори, наукові статті, журналістські тексти, повсякденні діалоги та інші. Загальні корпуси намагаються представити мову в цілому, охоплюючи широкий спектр жанрів і стилів. Наприклад, корпус COCA (Corpus of Contemporary American English) містить тексти з різних джерел, включаючи художню літературу, газети, наукові статті та інші, забезпечуючи репрезентативний огляд сучасної американської англійської [12]. Спеціалізовані корпуси, навпаки, зосереджені на певних жанрах або доменах використання. Наприклад, The Michigan Corpus of Academic Spoken English (MICASE) зосереджений на академічному усному англійському, забезпечуючи унікальний ресурс для вивчення академічної лексики та стилістики.

Крім того, корпуси текстів можуть відрізнятися за ступенем анотації. Вони можуть бути нерозміченими, частково або повністю анотованими. Нерозмічені корпуси містять тексти без будь-якої додаткової інформації про структуру або зміст. Вони використовуються для загального аналізу текстів або як сировина для подальшої анотації. Частково анотовані корпуси містять деяку структурну або лінгвістичну інформацію, таку як маркування речень або токенізація. Повністю анотовані корпуси, як-от Penn Treebank, включають детальні анотації, такі як синтаксичні дерева або семантичні ролі, що робить їх надзвичайно корисними для лінгвістичних досліджень і навчання алгоритмів машинного навчання.

Значення корпусів текстів важко переоцінити. Вони є основним джерелом даних для лінгвістичних досліджень і мають широкий спектр застосувань. У лінгвістиці вони слугують основою для вивчення лексичних, граматичних і стилістичних особливостей мови, дозволяючи дослідникам проводити кількісні та якісні аналізи. Наприклад, аналіз колокацій (сполучень слів) в корпусах дозволяє

виявляти типові моделі використання слів і фраз, що є важливим для розуміння семантики та синтаксису мови. Крім того, корпуси текстів є критичними для задач обробки природної мови, таких як розпізнавання мовлення, машинний переклад і генерація тексту. Наприклад, системи машинного перекладу, такі як Google Translate, навчаються на багатомовних паралельних корпусах, що дозволяє їм забезпечувати високоякісний переклад між різними мовами.

Корпуси текстів також є важливими для аналізу соціальних і культурних контекстів мови. Вони дозволяють дослідникам аналізувати, як мова змінюється під впливом соціальних, політичних і культурних факторів. Наприклад, аналіз корпусів з соціальних медіа може виявити, як сучасні технології та комунікаційні платформи впливають на мовні практики та вираження емоцій. Крім того, корпуси, які включають тексти з різних культурних контекстів, дозволяють вивчати мовні варіації та культурні відмінності у використанні мови.

3.2 Розробка та структуризація корпусів

Розробка та структуризація корпусів текстів є основним етапом у створенні лінгвістичних ресурсів, що дозволяють здійснювати комплексний аналіз великих обсягів текстових даних. Цей процес включає кілька важливих аспектів, які необхідно враховувати для забезпечення якості та репрезентативності кінцевого продукту.

Визначення мети та обсягу корпусу є першочерговим завданням. Важливо зрозуміти, які конкретні дослідження або прикладні завдання будуть виконуватися з використанням цього корпусу. Це допомагає сформулювати основні критерії відбору текстів, а також визначити необхідний обсяг даних, що має бути включений до корпусу. Наприклад, корпус, розроблений для аналізу сучасної лексики в новинних медіа, повинен включати тексти з різних видань і тематик, щоб забезпечити широку репрезентативність.

При розробці корпусу важливо враховувати його репрезентативність. Це означає, що вибір текстів повинен відображати різноманітність мови або специфічну тематику, для якої створюється корпус. Наприклад, якщо корпус

призначений для дослідження наукового стилю, він повинен включати тексти з різних наукових дисциплін і жанрів, що забезпечить всебічне покриття предметної області. Технології обробки тексту, такі як машинне навчання та методи обробки природної мови, можуть бути використані для автоматизації процесів відбору текстів та забезпечення їхньої відповідності вимогам.

Анотація текстів є критично важливим етапом у процесі створення корпусу. Вона включає додавання лінгвістичної інформації до текстів, що полегшує їх подальший аналіз. Наприклад, морфологічна розмітка дозволяє ідентифікувати частини мови, а синтаксична розмітка описує структуру речень. Сучасні технології, такі як автоматичні анотатори на основі глибинного навчання, значно спрощують цей процес, забезпечуючи високу точність і швидкість анотації. Такі інструменти, як SpaCy або Stanford NLP, часто використовуються для цієї мети, оскільки вони надають потужні можливості для розпізнавання мовних структур і елементів [13].

Структурування корпусу включає організацію текстів у відповідні бази даних або інші структури зберігання, що дозволяє ефективно управляти великими обсягами даних. Це може включати використання реляційних баз даних або документно-орієнтованих систем, таких як MongoDB, які забезпечують гнучкість і швидкий доступ до текстів. Структуровані дані полегшують виконання пошукових запитів і забезпечують можливість швидкого отримання необхідної інформації. Наприклад, в системах, де зберігаються великі корпуси текстів, використання Elasticsearch може значно прискорити пошук і аналіз даних завдяки його можливостям реального часу.

Забезпечення якості даних є важливим аспектом при розробці корпусів. Якість текстів та їх анотацій має вирішальне значення для точності подальшого аналізу. Тому необхідно проводити ретельну перевірку та корекцію текстів, а також використовувати методи контролю якості, такі як перевірка на відповідність стандартам анотації та використання спеціалізованих алгоритмів для виявлення і виправлення помилок. Це особливо важливо для корпусів, які

будуть використовуватися в наукових дослідженнях або для навчання моделей машинного навчання.

Автоматизація процесів створення і структуризації корпусів за допомогою сучасних технологій дозволяє значно скоротити час і зусилля, необхідні для їх розробки. Використання технологій машинного навчання, таких як автоматичне анотування і класифікація текстів, допомагає зменшити ручну роботу і підвищити точність і швидкість обробки даних. Наприклад, автоматичні інструменти можуть ефективно розмічати великі обсяги текстів, забезпечуючи їхню готовність для подальшого аналізу.

3.3 Використання корпусів у дослідженні та прикладних завданнях

Корпуси текстів є невід'ємною частиною сучасних лінгвістичних досліджень та різноманітних прикладних задач у галузі обробки природної мови (NLP). Вони забезпечують основу для проведення комплексного аналізу мовних явищ і слугують джерелом даних для навчання та тестування моделей машинного навчання [14].

У лінгвістичних дослідженнях корпуси текстів дозволяють глибше вивчати мовні закономірності та структури. Наприклад, дослідження колокацій, тобто частих сполучень слів у текстах, може виявити типові патерни використання слів і фраз. Це сприяє кращому розумінню семантики та синтаксису мови. Аналіз корпусів також допомагає досліджувати діалектні та соціолінгвістичні відмінності. Наприклад, розгляд регіональних варіантів англійської мови у різних текстах дозволяє виявити, як використання лексики та граматичних структур змінюється залежно від географічного або соціального контексту.

Освітні технології широко використовують корпуси текстів для створення навчальних матеріалів та тестів, а також для оцінки мовних навичок учнів. Сучасні системи вивчення мов, такі як Duolingo або Babbel, застосовують корпуси для формування завдань, що відображають реальні мовні ситуації [15]. Це дозволяє користувачам практикувати мову в контекстах, які максимально наближені до повсякденного життя. Крім того, системи автоматичної перевірки

граматики та стилю, такі як Grammarly, використовують корпуси текстів для розробки алгоритмів, що аналізують тексти учнів і надають рекомендації з покращення мовлення.

Машинний переклад значною мірою залежить від корпусів текстів. Багатомовні паралельні корпуси, що містять тексти на різних мовах разом з їх перекладами, служать основним джерелом даних для навчання моделей перекладу. Це дозволяє системам, таким як Google Translate, забезпечувати якісний переклад між різними мовами [16]. Сучасні моделі нейронного машинного перекладу, такі як трансформери, використовують ці корпуси для навчання, що дозволяє їм враховувати контекст та семантичні зв’язки між словами, забезпечуючи більш природні та точні переклади.

Аналіз соціальних мереж на основі корпусів текстів надає унікальні можливості для дослідження сучасних мовних тенденцій і поведінкових патернів. Наприклад, дослідження даних з Twitter або Facebook може показати, як нові слова і фрази входять в ужиток, а також як люди виражають свої думки та емоції в онлайн-спілкуванні. Це має значення для маркетингових досліджень і політичного аналізу, оскільки компанії і політичні аналітики можуть використовувати ці дані для відстеження реакцій на продукти, послуги або події.

У технологіях розпізнавання та синтезу мови корпуси текстів грають вирішальну роль. Вони забезпечують навчальні дані для моделей, які розпізнають і синтезують мовлення. Наприклад, голосові помічники, такі як Siri або Google Assistant, використовують корпуси текстів для навчання своїх моделей, щоб вони могли розуміти та відповідати на запити користувачів природною мовою. Системи синтезу мови, як-от WaveNet від DeepMind, створюють голосові відповіді на основі текстів, використовуючи корпуси для навчання моделей, що генерують високоінтелектуальні і природні звуки [17].

Робота з корпусами текстів супроводжується кількома викликами. Забезпечення якості даних і підтримка їхньої актуальності є ключовими завданнями. Якість текстів та їх анотацій впливає на точність подальшого аналізу, тому важливо проводити ретельну перевірку та корекцію даних. Також важливо

регулярно оновлювати корпуси новими даними, щоб вони відображали сучасні мовні тенденції. Крім того, необхідно враховувати етичні аспекти роботи з текстовими даними, особливо з огляду на конфіденційність інформації та авторські права.

Майбутнє роботи з корпусами текстів виглядає перспективно завдяки постійному розвитку технологій штучного інтелекту та машинного навчання. Ці технології відкривають нові можливості для автоматизації створення і анотації корпусів, що дозволяє ефективніше обробляти великі обсяги текстових даних. Зростаючий інтерес до багатомовних та мультимодальних корпусів сприяє розширенню можливостей у сфері багатомовної обробки текстів.

3.4 Проблеми та перспективи в роботі з корпусами

Перетворення корпусів текстів є важливим етапом у підготовці даних для лінгвістичних досліджень і різноманітних застосувань у галузі обробки природної мови (NLP). Цей процес передбачає застосування ряду інструментів та технологій, які забезпечують ефективну обробку та аналіз текстових даних.

Одним із ключових аспектів перетворення корпусів є нормалізація текстів, яка включає приведення тексту до стандартної форми. Це може включати видалення пунктуації, перетворення всіх літер у нижній регістр або обробку текстів різними способами для забезпечення їхньої однорідності [18]. Нормалізація дозволяє зменшити різноманітність текстових даних і спрощує подальший аналіз. Наприклад, нормалізація тексту є важливою для забезпечення ефективної роботи алгоритмів машинного навчання, які використовуються для обробки текстових даних.

Інша важлива задача – це токенізація, яка передбачає розбиття тексту на окремі одиниці, такі як слова або речення. Токенізація є основою для багатьох лінгвістичних аналізів, оскільки вона дозволяє виділяти окремі лексичні одиниці з тексту. Існують різні підходи до токенізації, зокрема, використання регулярних виразів або спеціалізованих інструментів, таких як NLTK або SpaCy, які можуть

адаптуватися до особливостей різних мов [19]. Важливо зазначити, що правильна токенізація може значно вплинути на точність подальшого аналізу.

Анотація текстів є ще одним важливим етапом перетворення корпусів. Вона включає додавання лінгвістичних міток, таких як частини мови, синтаксичні структури або семантичні ролі, до текстових даних. Це дозволяє проводити більш детальний аналіз мовних особливостей. Наприклад, морфологічна анотація дозволяє ідентифікувати частини мови кожного слова, що є корисним для завдань синтаксичного аналізу або розпізнавання сутностей. Автоматичні анотатори, такі як Stanford CoreNLP або OpenNLP, використовують методи машинного навчання для забезпечення точної і швидкої анотації великих обсягів текстів.

Лематизація і стемінг є процесами, які зменшують слова до їхньої базової форми або кореня. Це дозволяє зменшити різноманітність словоформ і полегшити аналіз текстів. Лематизація використовує словники або алгоритми для перетворення слова до його базової форми (леми), тоді як стемінг просто відрізає кінцеві частини слів для виявлення їхнього кореня. Вибір між лематизацією і стемінгом залежить від конкретних потреб аналізу [20]. Наприклад, лематизація є більш точною, але стемінг може бути швидшим і менш ресурсоємним.

Автоматичне вилучення термінів та іменованих сутностей є ще одним ключовим аспектом перетворення корпусів текстів. Це включає ідентифікацію специфічних термінів або імен (таких як назви людей, місць або організацій) у текстах. Цей процес важливий для завдань, які потребують розпізнавання ключових слів або інформації з текстів. Інструменти, такі як SpaCy або Named Entity Recognition (NER) модулі, використовуються для автоматичного розпізнавання іменованих сутностей у текстах.

Інструменти для роботи з корпусами, такі як Sketch Engine або Corpus Workbench, пропонують широкі можливості для зберігання, обробки та аналізу текстових даних. Вони дозволяють виконувати складні пошукові запити, генерувати статистичні дані та проводити аналіз лінгвістичних патернів. Наприклад, Sketch Engine дозволяє користувачам створювати власні корпуси, додавати анотації і аналізувати тексти за допомогою потужних алгоритмів. Corpus

Workbench забезпечує ефективне управління великими обсягами текстів і дозволяє виконувати пошук та аналіз текстових даних з високою швидкістю і точністю.

Автоматизація перетворення корпусів текстів є важливим аспектом сучасних технологій обробки даних. Використання сценаріїв або програмних засобів для автоматизації рутинних завдань, таких як очищення даних або анотація, дозволяє значно скоротити час і зусилля, необхідні для підготовки текстових даних. Наприклад, Python скрипти можуть бути використані для автоматизації процесів нормалізації, токенізації і анотації текстів, що забезпечує ефективну обробку великих обсягів даних.

4 РОЗРОБКА ЗАСТОСУНКУ ДЛЯ ОБРОБКИ ЛІНГВІСТИЧНИХ РЕСУРСІВ

4.1 Визначення цілей та вимог до застосунку

Для розробки програмного застосунку, призначеного для роботи з корпусами текстів, важливо чітко визначити основні цілі проекту та вимоги, що забезпечать його ефективне функціонування та зручність використання.

Головною метою проекту є створення зручного і функціонального інструменту, який дозволить користувачам працювати з текстовими корпусами. Застосунок має забезпечувати можливість завантаження текстів у популярних форматах, таких як TXT і CSV, що надасть доступ до текстових даних з різних джерел. Завантажені тексти повинні зберігатися у структурованій базі даних, що дозволить легко здійснювати пошук і доступ до текстів. Це забезпечить користувачам можливість ефективно працювати з великими обсягами текстової інформації.

Інтерактивний аналіз текстів є ключовою функцією застосунку. Користувачі повинні мати змогу здійснювати пошук за ключовими словами або фразами у завантажених текстах, що дозволить швидко знаходити необхідну інформацію. Додатково, застосунок повинен надавати статистику текстів, таку як кількість слів і частота використання окремих слів або фраз. Це допоможе користувачам отримати загальне уявлення про зміст та структуру корпусу, забезпечуючи основні аналітичні можливості без складних налаштувань.

Інтерфейс користувача повинен бути простим і інтуїтивно зрозумілим, дозволяючи легко взаємодіяти з текстами. Користувачі повинні мати можливість завантажувати тексти, виконувати пошук і переглядати результати аналізу без необхідності мати глибокі технічні знання. Це забезпечить зручність використання застосунку для широкого кола користувачів, включаючи тих, хто не є експертами у галузі лінгвістики або обробки текстів.

Безпека даних є важливим аспектом, який слід враховувати під час розробки застосунку. Всі завантажені тексти повинні зберігатися у безпечному середовищі, що забезпечує захист від несанкціонованого доступу. Користувачі мають

контролювати свої дані, включаючи можливість видалення або редагування текстів, що забезпечить конфіденційність та безпеку інформації.

4.2 Розробка архітектури та вибір технологій

Основним принципом побудови архітектури є модульність, що дозволяє розділити застосунок на незалежні частини, кожна з яких виконує свою специфічну роль (див.рис.4.1).

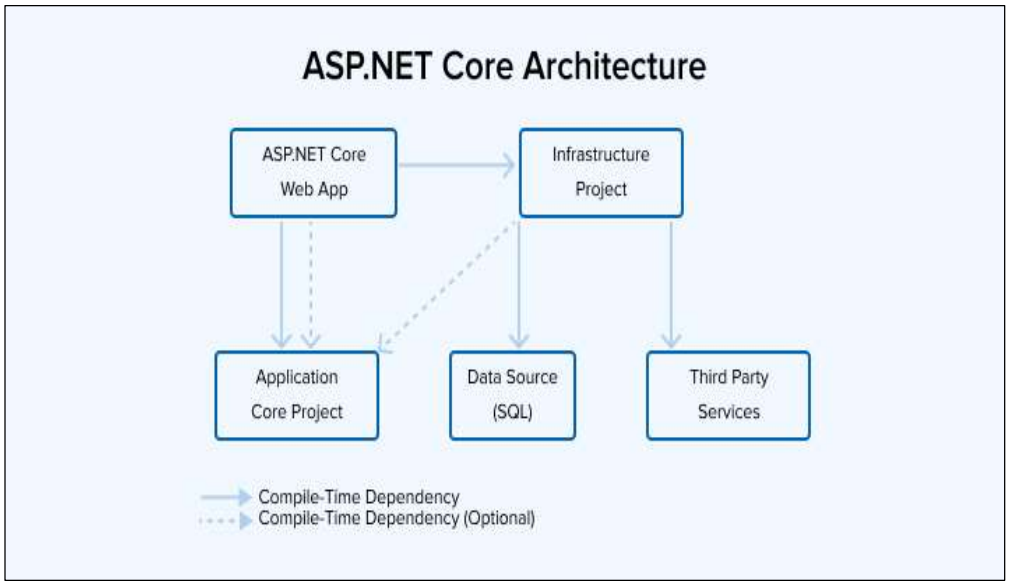

Рисунок 4.1 – Приклад модульної архітектури

Такий підхід сприяє більш гнучкому управлінню проектом та спрощує процес оновлення і розширення функціональності. У контексті застосунку для роботи з корпусами текстів, ми можемо визначити кілька основних компонентів, що будуть взаємодіяти між собою, забезпечуючи виконання всіх необхідних функцій.

Інтерфейс користувача (UI) виступає як основний засіб взаємодії між користувачем та системою. Він повинен бути розроблений таким чином, щоб забезпечувати зручний і інтуїтивно зрозумілий доступ до основних функцій застосунку, таких як завантаження текстів, пошук і аналіз. Сучасні технології, на основі Javascript, дозволяють створити привабливий і функціональний інтерфейс, що забезпечить позитивний користувацький досвід.

Серверна частина (Backend) відповідає за обробку запитів користувачів, управління даними та виконання лінгвістичного аналізу. Використання технологій, таких як ASP.NET Core, допомагає створити надійну і продуктивну серверну інфраструктуру. Серверна частина повинна забезпечувати доступ до бази даних, а також до зовнішніх ресурсів, необхідних для аналізу текстів. Важливо, щоб ця частина системи була розроблена з урахуванням можливості масштабування для обробки великих обсягів даних.

База даних є центральним компонентом для зберігання текстових корпусів та пов’язаних метаданих. Використання сучасних баз даних, таких як SQL Server, дозволяє забезпечити ефективне зберігання та швидкий доступ до текстів. Система повинна бути оптимізована для роботи з великими обсягами текстової інформації, забезпечуючи швидкий пошук та маніпулювання даними.

Лінгвістичні сервіси, що здійснюють аналіз текстів, відіграють ключову роль у функціональності застосунку. Ці сервіси можуть виконувати такі завдання, як розпізнавання іменних сутностей, морфологічна розмітка та синтаксичний аналіз. Використання готових інструментів і бібліотек, таких як SpaCy або Stanford NLP, може значно спростити реалізацію цих функцій, забезпечуючи високу точність і ефективність аналізу.

4.3 Функціональність та можливості застосунку

Застосунок дозволяє користувачам завантажувати тексти у різних форматах, таких як TXT, CSV та JSON. Це забезпечує гнучкість у роботі з текстовими даними з різних джерел. Завантаження може здійснюватися як з локальних файлів, так і з зовнішніх ресурсів через URL. Після завантаження тексти зберігаються у структурованій базі даних, що забезпечує зручний доступ і ефективне управління великими обсягами текстових даних. Застосунок використовує MSSQL базу для зберігання, що дозволяє обробляти великі корпуси текстів швидко і ефективно (див.рис.4.2).

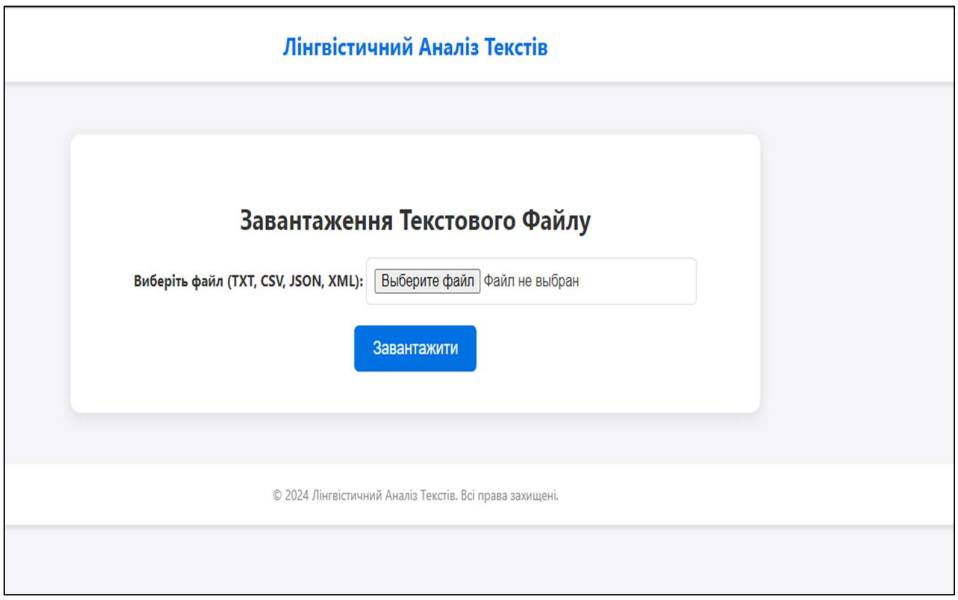

Рисунок 4.2 – Завантаження текстового файлу

Важливою функцією застосунку є анотація текстів. Користувачі можуть додавати лінгвістичні маркування до текстів, такі як морфологічна розмітка та розпізнавання іменних сутностей (див.рис.4.3).

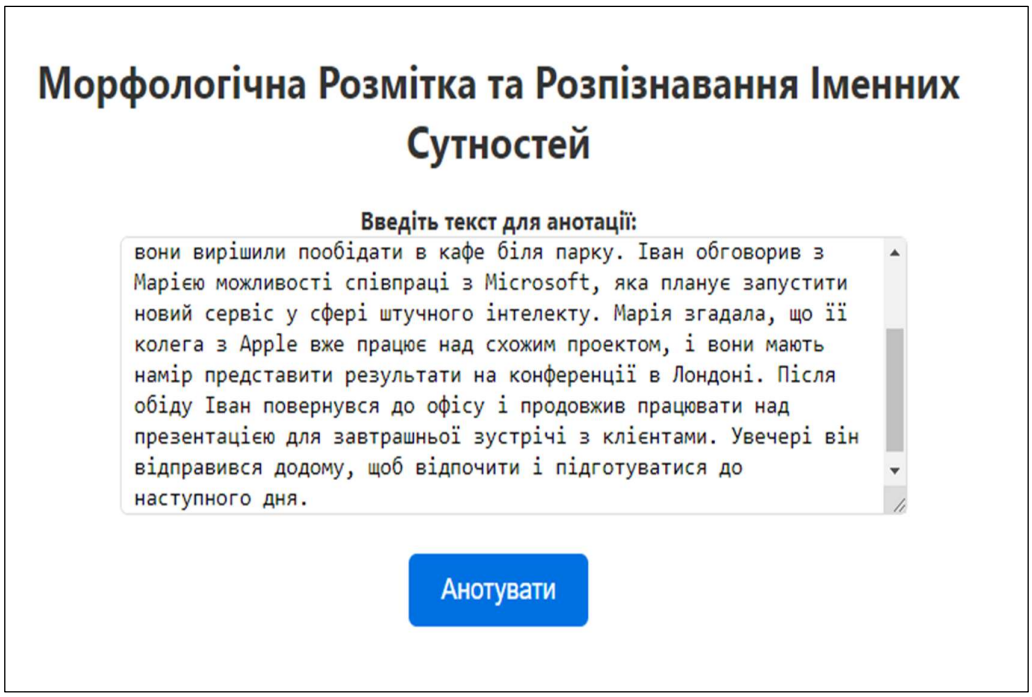

Рисунок 4.3 – Анотація тексту

Застосунок використовує бібліотеку Stanford NLP, що дозволяє

автоматизувати процес анотації, зменшуючи потребу у ручному втручанні. Це особливо корисно для дослідників, які потребують додаткової інформації про структуру текстів для своїх аналізів.

Пошукові можливості застосунку дозволяють користувачам швидко знаходити необхідну інформацію у завантажених текстах. Система підтримує повнотекстовий пошук, що дозволяє здійснювати пошук за ключовими словами або фразами у всьому корпусі текстів. Додатково, користувачі можуть застосовувати фільтри для звуження результатів пошуку за певними критеріями, такими як дата створення тексту або автор. Це значно спрощує процес навігації і допомагає зосередитися на найбільш релевантних фрагментах текстів.

Застосунок надає інструменти для аналізу текстових даних. Користувачі можуть отримувати базову статистику про тексти, включаючи кількість слів і частоту використання окремих слів або фраз. Для більш глибокого аналізу застосунок підтримує візуалізацію даних через графіки і діаграми (див.рис.4.4).

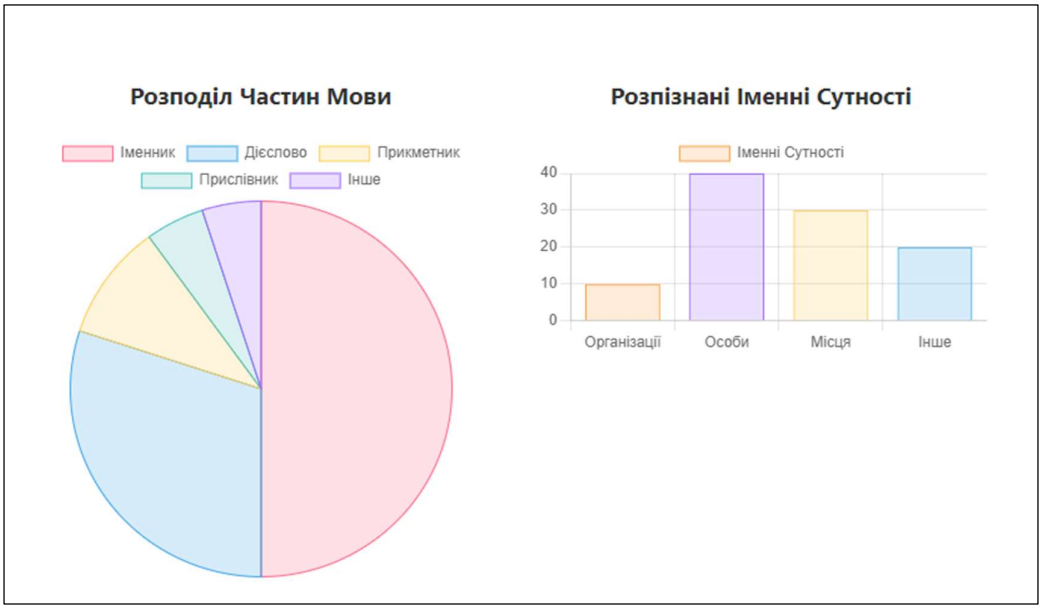

Рисунок 4.4 – Результат морфологічного аналізу

Це дозволяє користувачам краще розуміти мовні патерни і тенденції, надаючи зрозуміле і наочне представлення інформації. Візуалізація допомагає ідентифікувати важливі аспекти текстових даних, що може бути корисним для різних аналітичних задач.

Безпека і конфіденційність даних залишаються важливими аспектами функціональності застосунку. Всі тексти зберігаються у захищеному середовищі, і доступ до них мають лише авторизовані користувачі. Система використовує методи шифрування для захисту даних під час передачі і зберігання, що забезпечує високий рівень безпеки і конфіденційності інформації. Застосунок забезпечує надійний контроль доступу, що дозволяє тільки авторизованим користувачам взаємодіяти з конфіденційними даними, гарантуючи захист від несанкціонованого доступу.

4.4 Впровадження, тестування та розгортання застосунку

Впровадження, тестування та розгортання є критичними етапами у життєвому циклі розробки програмного забезпечення. Ці процеси забезпечують високу якість, надійність та готовність застосунку до використання. Далі ми розглянемо кожен з цих аспектів у контексті розробки застосунку для роботи з корпусами текстів на платформі .NET з використанням C#.

Процес впровадження розпочинається після завершення основних етапів розробки. У нашому випадку, це включає інтеграцію всіх компонентів застосунку, таких як інтерфейс користувача, серверна частина, база даних та лінгвістичні сервіси. Першим кроком є створення тестового середовища, яке відображає реальні умови експлуатації. Це середовище дозволяє провести інтеграційне тестування, забезпечуючи, що всі частини системи працюють разом без збоїв.

Після успішного тестування, застосунок готується до впровадження у реальне робоче середовище. На цьому етапі важливо забезпечити належну документацію для користувачів та адміністраторів, що включає інструкції щодо встановлення, конфігурації та використання системи. Документація допомагає користувачам швидко освоїти нові функції і ефективно використовувати застосунок у своїй роботі. Важливим аспектом впровадження є також підготовка персоналу, який буде відповідати за підтримку і експлуатацію системи. Це може включати навчання або проведення інформаційних сесій.

Тестування є невід'ємною частиною процесу забезпечення якості

програмного забезпечення. У нашому застосунку використовуються різні підходи до тестування, щоб забезпечити його надійність і відповідність вимогам користувачів.

Модульне тестування фокусується на перевірці окремих компонентів системи. Кожен модуль тестується окремо, щоб переконатися, що він працює правильно у відриві від інших частин системи. Це допомагає виявити і виправити помилки на ранніх етапах розробки. Наприклад, функції завантаження текстів або пошуку можуть бути перевірені на коректність виконання з використанням тестових даних.

Інтеграційне тестування перевіряє, як окремі компоненти системи взаємодіють один з одним. Це важливо для забезпечення того, що всі частини застосунку працюють разом у гармонії. На цьому етапі ми перевіряємо, чи правильно передаються дані між інтерфейсом користувача і серверною частиною, чи коректно взаємодіє сервер з базою даних, і чи лінгвістичні сервіси надають правильні результати.

Користувацьке тестування проводиться, щоб переконатися, що застосунок відповідає очікуванням і потребам кінцевих користувачів. Це включає тестування функцій з точки зору користувача, що допомагає виявити будь-які проблеми з юзабіліті або функціональністю. На цьому етапі залучаються реальні користувачі або тестові групи, які використовують систему у звичайному режимі і надають зворотний зв'язок щодо її роботи. Це дозволяє зробити необхідні коригування перед остаточним розгортанням.

Автоматизоване тестування використовує скрипти та інструменти для автоматизації процесу тестування. Це особливо корисно для перевірки рутинних або повторюваних завдань, таких як запуск тестових сценаріїв для різних входів і перевірка вихідних результатів. Автоматизація значно прискорює процес тестування і знижує ймовірність людських помилок.

Розгортання застосунку є завершальним етапом у його життєвому циклі. Після успішного тестування система готова до перенесення у виробниче середовище. Цей процес включає встановлення програмного забезпечення на

цільові сервери або пристрої, налаштування необхідної інфраструктури і забезпечення доступу користувачів до нової системи.

На першому етапі розгортання здійснюється встановлення застосунку на сервери. Це може включати як фізичні сервери, так і віртуальні машини у хмарному середовищі. Платформа .NET забезпечує гнучкість у виборі середовища розгортання, дозволяючи використовувати як локальні, так і хмарні ресурси. Серверна частина, включаючи базу даних і бекенд-сервіси, конфігурується відповідно до вимог продуктивності і безпеки.

Після встановлення проводиться налаштування мережевих параметрів, таких як налаштування доступу до застосунку з різних мереж або забезпечення безпеки передачі даних через SSL/TLS. Це гарантує, що користувачі можуть безпечно і ефективно взаємодіяти з системою. Важливим кроком є також налаштування бази даних, включаючи імпорт початкових даних, налаштування прав доступу і оптимізацію продуктивності.

Коли система готова до використання, проводиться тестове розгортання, яке дозволяє перевірити всі аспекти встановлення та налаштування у виробничому середовищі без впливу на реальних користувачів. Це включає перевірку працездатності всіх функцій, продуктивності та стабільності системи. Після успішного тестового розгортання система стає доступною для всіх користувачів.

ВИСНОВКИ

У цій роботі ми детально розглянули важливість, структуру, та використання лінгвістичних ресурсів, особливо акцентуючи увагу на корпусах текстів. Значення цих ресурсів в сучасному світі лінгвістичних досліджень та обробки природної мови є величезним, оскільки вони відіграють ключову роль у розвитку мовних технологій та когнітивних досліджень. Розробка застосунку на .NET і C# для обробки цих ресурсів продемонструвала можливості сучасних технологій та підкреслила важливість програмного забезпечення в цій області.

Розробка програмного рішення відкрила нові перспективи для ефективного використання лінгвістичних даних. Система, яка була розроблена у рамках цієї роботи, демонструє, як інтеграція різноманітних модулів та використання передових алгоритмів може сприяти кращому розумінню та обробці природної мови. Водночас, проект виявив і ряд викликів, зокрема пов'язаних з обробкою великих обсягів даних, забезпеченням якості анотацій та інтеграцією з іншими системами та базами даних.

Процес розробки підкреслив необхідність глибокого аналізу вимог до програмного продукту та ретельного планування його структури. Важливим аспектом стало забезпечення гнучкості застосунку для адаптації до змінних потреб користувачів та вдосконалення інструментів обробки текстів.

Подальший розвиток у цій галузі передбачає не тільки поліпшення та оновлення існуючих корпусів і методів аналізу, але й розвиток нових підходів до обробки природної мови. Це включає в себе інтеграцію зі штучним інтелектом та машинним навчанням для створення ще більш ефективних та інтуїтивно зрозумілих інструментів. Розвиток технологій відкриває нові можливості для аналізу та інтерпретації мовних даних, розширюючи кордони того, як ми розуміємо та використовуємо природну мову.

Завершуючи, можна сказати, що робота з лінгвістичними ресурсами та розробка інструментів для їх обробки - це складна, але надзвичайно захоплююча область, що має великий потенціал для вдосконалення засобів комунікації, аналізу даних, та розвитку наукових досліджень. Результати цієї роботи

підтверджують важливість подальшого дослідження та розвитку у цій сфері, а також необхідність інвестування в нові технології та методики для збагачення наших знань та покращення інструментів, які ми використовуємо для розуміння та обробки мови.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Меннінґ К. Д., Шютце Г. Основи статистичної обробки природної мови. Кембридж: MIT Press, 1999. 620 с.

2. Джурафський Д., Мартін Дж. Г. Обробка мови та мовлення. 3-тє вид. США: Prentice Hall, 2019. 1024 с.

3. Албахарі Дж., Албахарі Б. C# 8.0 як в оріховій шкаралупі: Визначний довідник. США: O'Reilly Media, 2019. 1098 с.

4. Microsoft. NET документація [Електронний ресурс]. Режим доступу: https://docs.microsoft.com/dotnet. (дата звернення: 30.04.2024).

5. Іде Н., Пустеджовські Дж. (ред.). Довідник з лінгвістичної анотації. Кембридж: Springer, 2017.

6. Чакраборті Г., Паголу М., Гарла С. Текстовий аналіз і аналіз: практичні методи, приклади та кейс-стадії з використанням SAS. Кері, Північна Кароліна: SAS Institute, 2014.

7. МакЕнері Т., Харді А. Корпусна лінгвістика: метод, теорія та практика. Кембридж: Cambridge University Press, 2012.

8. Бауер Т. Лінгвістичні ресурси в цифрову еру. В: Т. М. Джонсон (ред.), Досягнення в цифровій лінгвістиці (с. 142-158). Кембридж: Cambridge University Press, 2019.

9. Чен С. Автоматизовані техніки вилучення термінів. В: Матеріали 15-ї Міжнародної конференції з технології мови (с. 67-75), 2020.

10. Шубін І., Снісар С., Литвин С. Формалізація та застосування алгебраїчних методів у автоматизованих інтелектуальних системах. В: Матеріали 8-ї Міжнародної конференції з проблем інфокомунікацій науки та техніки (PIC S and T 2021), Харків, Україна, 2021 (с. 67-70).

11. Джонсон Т., Сміт П. Комплексний огляд технологій мови. Огляди технологій мови, том 35, № 2, 2024, с. 102-134.

12. Андерсон Л. (ред.). Інновації у вивченні мов. В: Л. Андерсон (ред.), Нові напрями в освітніх технологіях (Т. 2, с. 45-78). Лондон: Educational Tech Press, 2021.

13. Ейзенштейн Дж. Введення до обробки природної мови. Кембридж: MIT Press, 2019.

14. Берд С., Кляйн Е., Лопер Е. Природна мова з Python. Кембридж: O'Reilly Media, 2009. 504 с.

15. Бішоп К. М. Розпізнавання образів і машинне навчання. Нью-Йорк: Springer, 2006. 738 с.

16. Кетц Дж. Введення в обробку природної мови. Бостон: Addison-Wesley, 2021. 420 с.

17. Віттен І. Х., Франк Е., & Голл М. А. Data Mining: Practical Machine Learning Tools and Techniques. Сан-Франциско: Morgan Kaufmann, 2016. 654 с.

18. Журавльов О. Автоматичний переклад: теорія і практика. Київ: Видавництво Наукова Думка, 2020. 320 с.

19. Райзен М. Статистичні методи в обробці природної мови. Лондон: Chapman & Hall, 2017. 598 с.

20. Маріо Ж., Пітерс Й. Лінгвістичні корпорації та цифрові інновації. Відень: Springer, 2019. 450 с.