Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра мережних технологій Факультету інформатики

КУРСОВА РОБОТА

на тему «Проблеми NLP для української мови»
за спеціальністю 122 «Комп'ютерні науки»

Керівник курсової роботи

д.т.н., проф. Глибовець А. М.

(*прізвище та ініціали*)

_____

(*підпис*)

“____” _____ 2024 р.

Виконав студент Сов’як В.Б

(*прізвище та ініціали*)

_____

“____” _____ 2024 р.

Київ-2024

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра мережних технологій Факультету інформатики

ЗАТВЕРДЖУЮ

Зав.кафедри мережних технологій,

проф., д.ф.-м.н. Г.І. Малашонок

(підпис)

„____”_____2024 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на курсову роботу

студенту Сов’яку Віктору Борисовичу факультету інформатики 3-го курсу

ТЕМА: Проблеми NLP для української мови

Зміст ТЧ до курсової роботи:

1. Індивідуальне завдання
2. Календарний план
3. Зміст
3. Анотація
4. Вступ
4. Огляд і характеристика NLP
5. Аналіз NLP для англійської мови
6. NLP для української мови
7. Висновки
8. Список використаних джерел

Дата видачі “___” _____ 2024 р. Керівник _____

Завдання отримав _____

Тема: Проблеми NLP для української мови

Календарний план виконання роботи:

№ з/п	Назва Етапу	Термін Виконання	Примітка
1	Вибір теми	Жовтень 2023	
2	Затвердження теми з керівником	Листопад 2023	
3	Дослідження і огляд NLP	Листопад 2023 – січень 2024	
4	Прикладний огляд технологій	Січень 2024 - Лютий 2024	
5	Дослідження проблем NLP для української мови	Лютий 2024 – Березень 2024	
6	Висновки	Березень 2024 – Квітень 2024	

Студент Сов’як В.Б.

Керівник Глибовець А.М.

“____” _____ 2024 р.

Зміст

Анотація

Мета курсової роботи полягає у досліджені проблеми NLP для української мови, які стримують розвиток у цій сфері. В роботі було проаналізовано основні проблеми, що розв’язуються з використанням підходів та алгоритмів обробки природної мови. Серед них: розбір залежностей, виявлення іменованих сутностей, аналіз тональності, моделювання тем та питально відповідальні підсистеми. Також, проаналізовано основні інноваційні технології, зокрема BERT та GPT. Аналіз базувався на описі підходів та технологій, що працюють з англійською та українською мовою. В ході роботи були сформовані основні припущення про причини, що стримують розвиток засобів обробки української мови та запропоновані напрями подальшого розвитку.

Скорочення та умовні позначення

NLP – Natural Language Processing

AI – artificial intelligence

POS – part-of-speech

NER – named entity recognition.

DP – dependency parsing

SA – sentiment analysis

TM – topic modeling

Q&A – Question and Answer

LM – language model

LLM – large language model

RNN – recurrent neural network

URL – Uniform Resource Locator

Вступ

У сучасному світі, важливість обробки природньої мови (NLP) не можливо перебільшити, що зумовлено стрімким розвитком інформаційних технологій разом із неспинним збільшенням текстових даних, які потребують аналізу певного рівня. Інструменти, наявні в провідних природніх мовах світу, таких як англійська, вражають своєю точністю виконання завдань, хоча і мають все ж недоліки. Розвиток NLP-систем для таких low-resource-language, як українська, необхідний для забезпечення доступу, до новітніх технологій та підвищення якості обробки текстів українською мовою. Постановка проблем та надання їх потенційних рішень сприятиме подальшому прогресу, забезпечуючи відповідний рівень технологічного розвитку для обробки української мови.

Мета даної роботи є виявлення ключових проблем обробки української мови, які стримують розвиток у цій сфері. Для досягнення цілі, були поставлені основні завдання, такі як: опис сучасного стану NLP сфери, аналіз існуючих алгоритмів NLP для найбільш розвинутої мови – англійської, визначення стану для української мови та опис проблем.

Робота складається з трьох розділів.

Перший розділ присвячено характеристиці сфери NLP, разом із основними техніками (нормалізація, стемінг, лематизація) та завданнями: розбір залежностей (DP), виявлення іменованих сутностей (NER), аналіз тональності (SA), моделювання тем (TM) та питально відповідальні системи (Q&A).

У другому розділі описано алгоритми та фреймворки, що використовуються для роботи з англійською мовою (найбільш розвиненої у даній сфері), а саме BERT, GPT, SQuAD, AI2 Datasets, GLUE Benchmark.

Третій розділ присвячено визначенню стану NLP для української мови, зокрема провідних інструментів як: GPT-2 UA, RoBERTa-WECHEL, youscan-roberta, UberText 2.0, UA SQuAD, CulturaX. Після детального аналізу, висвітлено знайдені проблеми та запропоновані потенційні рішення.

Для розв’язання поставлених задач використовувалися методи аналізу, порівняльного огляду статей, наукової літератури та експертні оцінки.

У результаті дослідження вдалося виявити основні проблеми, які стримують розвиток технологій обробки української мови, що стануть підґрунтям для подальших наукових розвідок у цій області. Отримані результати будуть корисні для фахівців у галузі обробки природної мови.

1. ОГЛЯД І ХАРАКТЕРИСТИКА NLP

1.1 Визначення NLP

Обробка природної мови (NLP) - це галузь комп'ютерних наук, яка поєднує комп'ютерну лінгвістику (моделювання природньої мови на основі правил) із моделями машинного навчання, що дає змогу обробляти мову у формі тексту. Такі системи виконують задачі перекладу текстів, формування відповідей на запитання, виявлення сутностей, тощо. У сучасну епоху, люди уже активно взаємодіють з технологіями обробки природної мови за допомогою систем цифрових асистентів або чат-ботів, що у свою чергу спрощує критично важливі бізнес-процеси, підкреслюючи важливість розвитку NLP і для підприємців. [1]

Через те, що автоматичні методи обробки мови ще не були настільки розвинуті, перші моделі базувалися на правилах, враховуючи основні граматичні конструкції. Через необхідність у інноваційних проривах, почали використовувати ймовірності, для визначення вірогідності приналежності до якогось певного класу, що започаткувало появу статистичних моделей. Сучасні ж системи NLP використовують інструменти глибокого навчання (deep learning) для того, аби навчатися із великою точністю розуміти, генерувати та аналізувати природню мову. Проте, розуміння сильно контекстуальних мовних особливостей, зокрема сарказму – проблема зовсім іншої складності, яка досі актуальна. Наразі, не можна сказати, що такі системи мають здатність повністю розуміти те, що аналізують. Вони здійснюють складний аналіз, для того, щоб знайти найкоректнішу відповідь. [2]

1.2 Історія розвитку

Розвиток NLP починається ще із середини 20 століття і включає у себе декілька етапів [2]:

а) 1950-ті роки. Обробка природної мови бере свій початок у цьому десятилітті, коли Алан Тюрінг розробив тест-перевірку, щоб визначити, чи є комп’ютер справді інтелектуальним [2]. Перевірка була такою: експерт віддалено спілкується із ЕОМ та намагається зрозуміти чи веде діалог із людиною. Якщо відповідь була позитивною, то за тестом Тюрінга таку систему можна було назвати інтелектуальною. Згодом виявилось що такої перевірки не є достатньо.

б) 1950-1990-ті роки. У цей період, NLP значною мірою базувалася на правилах, використовуючи правила ручної роботи, розроблені лінгвістами, щоб визначити, як комп'ютери будуть обробляти мову. Експеримент Джорджтаун-IBM у 1954 році став помітною демонстрацією машинного перекладу, автоматично перекладаючи понад 60 речень з російської на англійську. У 1980-х і 1990-х роках розвивався парсинг на основі правил, морфологія, семантика та інші форми розуміння природної мови. [2] Хоча такий підхід і стикнувся з обмеженням, через слабку адаптивність, це все ж спричинило прогрес в імітації логічного розуміння мовних структур.

в) 1990-ті роки. Оскільки комп’ютери стали швидкими, їх почали використовувати для створення статистичних правил. NLP, яка керується даними, стала основною протягом цього десятиліття, оскільки вона тепер базується на широкому спектрі наукових сфер, не обмежуючись лише лінгвістикою. [2]

г) 2000-2020-ті роки. Обробка природної мови не лише стрімко набрала популярність, а й значно проснулась завдяки розвитку обчислювальної потужності. Також вивчені методи нові з використанням неконтрольованих (unsupervised) та напівконтрольованих (semi-supervised) алгоритмів машинного навчання. NLP також почали інтегрувати з іншими додатками, такими як чат-боти та віртуальні помічники. Сьогодні підходи до NLP передбачають поєднання класичної лінгвістики та статистичних методів. [2]

Обробка природної мови продовжує розвиватися, поєднуючи ефективні методи для створення масштабних моделей. Перехід від систем на основі правил до статистичних методів відкрив нові перспективи для розвитку штучного інтелекту, що дає можливість ефективніше пристосовуватися до розмаїтих завдань і культурних контекстів.

1.3 Морфологічний аналіз

У NLP критично важливо чітко розуміти мову на різних етапах, що відображають послідовність обробки, від форми слів до їх сенсу. Вони дозволяють системам не просто зчитувати текст, але і створювати так зване “розуміння”, відповідаючи на питання, перекладаючи мову, та виконуючи інші завдання, які вимагають поглибленого розуміння людського мовлення.

Виділяють чотири основні етапи аналізу в NLP, а саме: морфологічний, синтаксичний, семантичний та прагматичний.

Розглянемо перший етап – морфологічний – на ньому відбувається аналіз тексту на рівні окремих слів з точки зору їх будови та граматики. Засоби, які найчастіше використовуються для досягнення такої мети є стемінг, лематизація, позначення частини мови (PoS tagging) та нормалізація.

Стемінг - це процес зменшення перегину в словах до їх кореневої форми [3]. Наприклад, слова “connect”, “connected”, “connection”, “connections”, “connects” після методу будуть зведені до “connect” (Рисунок 1.1).

	original_word	stemmed_word
0	connect	connect
1	connected	connect
2	connection	connect
3	connections	connect
4	connects	connect

Рисунок 1.1 – Зменшення перегину за допомогою стемінгу. [4]

“Корінь” в цьому випадку може бути не справжнім кореневим словом, а просто канонічною формою оригінального слова. Наприклад, “trouble”, “troubled”, “troubles”, “troublemsome” скоротяться до “trouble” та “troublemsom” (Рисунок 1.2). Стеммінг корисний для вирішення проблем розрідженості, а також стандартизації словникового запасу. [3]

	original_word	stemmed_word
0	trouble	troubl
1	troubled	troubl
2	troubles	troubl
3	troublemsome	troublemsom

Рисунок 1.2 – Зріз до несправжнього кореня [4]

Лематизація, на перший погляд, дуже схожа на стемінг, де мета полягає в тому, щоб видалити перегини і відобразити слово до його кореневої форми [3]. Єдина відмінність полягає в тому, що лематизація намагається зробити це належним чином, не просто відрубуючи частину, а насправді перетворюючи слова на фактичний корінь. Звідси, слова “touble”, “troubling”, “troubled”, “troubles” будуть зведені до спільної основи “trouble”, оскільки це їх лема (базова форма).

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

Рисунок 1.3 – Зріз слова за допомогою лематизації. [4]

Лематизація не дає значної переваги перед стемінгом для цілей пошуку та класифікації тексту. Насправді, залежно від обраного алгоритму, він може бути набагато повільнішим у порівнянні з використанням дуже базового стемера, і скоріше за все, доведеться знати частину мови відповідного слова, щоб отримати правильну лему [3]. Вимагаючи більше ресурсів для реалізації, вона все ж надає корисні переваги у вигляді точності обробки мови, що значно покращує здатність розуміти зміст текстів та проводити більш детальний аналіз.

Частини мовного тегування (Part-of-Speech tagging) - це лінгвістичний засіб в обробці природної мови, де кожному слову в документі надається певна частина мови (прислівник, прикметник, дієслово тощо) або граматична категорія [5]. Завдяки додаванню шарів синтаксичної та семантичної інформації до слів, ця процедура полегшує розуміння структури та значення речення. Наприклад, англійські символи “NN” означають іменники, “VB” - дієслова, а “JJ” - прикметники тощо. POS tagging використовують для завдань машинного перекладу, розпізнавання іменованих сутностей та вилучення інформації [5]. Даний метод дозволяє провести більш детальний аналіз, а також зрозуміти граматичну структуру тексту.

Нормалізація – процес перетворення тексту в канонічну (стандартну) форму. [3] Такий метод підготовки даних є критично важливим, особливо при роботі із даними, які можуть містити шум (соціальні мережі), наприклад, “2mrw” перетвориться на “tomorrow” (Рисунок 1.4). Нормалізація текстів є надзвичайно ефективною для аналізу неструктурованих клінічних текстів, де лікарі готують нотатки різними нестандартними способами. [3] На жаль, не існує стандартного способу нормалізації текстів, це все залежить від завдання.

Raw	Normalized
2moro 2mrrw 2morrow 2mrw tomrw	tomorrow
b4	before
otw	on the way
:) :-) ;-)	smile

Рисунок 1.4 – Нормалізація слів. [3]

Кожен метод морфологічного аналізу має власні як переваги, так і недоліки. Необхідно враховувати особливості завдання, перш ніж вибрати конкретний метод. Стемінг - простий і швидкий спосіб зменшити слова до їх основної форми, але іноді може відбутися скорочення до неіснуючого кореня. Лематизація зазвичай дає кращі результати, але порівняно зі стемінгом вона більш затратна. Позначення частини мови та нормалізація можуть знадобитись для більш глибокого аналізу мови. Щоб уникнути надмірного використання, важливо розуміти, які методи морфологічного аналізу найкраще відповідають необхідним потребам.

1.4 Синтаксичний аналіз

Синтаксичний аналіз описує логічне значення заданих речень або частин цих речень [6]. Необхідно розглядати правила граматики, щоб

визначити логічне значення, а також правильність речень. Синтаксичний аналіз виконується на рівні речення, на відміну від морфологічного, який проводиться на рівні окремих слів. Наприклад, розглянемо речення: “School go a boy”. Вище зазначене речення нелогічно виражає свій зміст через неправильну граматичну структуру та помилки. Тому, з цим нам допомагає синтаксичний аналіз, який показує, чи передає конкретне речення його логічне значення, а також, чи є його граматична структура правильною. [6]

Для виконання завдання парсингу використовується аналізатор (parser), що визначається як програмний компонент, який призначений для отримання вхідних текстових даних і дає структурне представлення вхідних даних після перевірки правильного синтаксису за допомогою формальної граматики. [6]. Він також генерує структуру даних, як правило, у вигляді дерева розбору або абстрактного синтаксичного дерева. Після дослідження різних можливих структур дерева, аналізатор намагається знайти найбільш оптимальне для даного речення [6].

Shift-reduce – один із доступних аналізаторів [6], будує синтаксичний розбір подібно до того, як це робиться у аналізаторів “знизу-вгору”, тобто дерево розбору будується з листків (знизу) до кореня (вгору). Його узагальненою формою є left-to-right (LR) аналізатор, для роботи якого потрібні буфер введення для зберігання вхідного рядка та стек для зберігання та доступу до правил продукції (production rules). [7]

Операції даного аналізатора включають у себе кілька базових дій:

а) Зсув (Shift) - включає переміщення символів з вхідного буфера на стек. [7] Коли аналізатор зустрічає термінал у вхідному рядку, він переміщує його на вершину стеку.

б) Зменшення (Reduce) виконується за допомогою відповідного правила продукції, якщо на вершині стеку з'являється “handle”. Це означає видалення

правої частини правила (RHS) зі стеку і додавання на його місце лівої частини правила (LHS). [7]

в) Дія аналізатора оголошується прийнятою (accepted), якщо в стеці залишився тільки початковий символ граматики і вхідний буфер порожній, що свідчиться про успішний розбір і припинення роботи аналізатора. [7]

г) Помилка (Error) - це ситуація, коли аналізатор не може виконати зсув, зменшення або прийняти введений рядок. [7] Така ситуація виникає, коли вхідний рядок не відповідає граматиці, або коли відбувається помилка в процесі розбору.

Парсинг залежностей (DP) - це техніка обробки природної мови, яка використовується для аналізу граматичної структури речень. [8] Це тип синтаксичного розбору, який має на меті визначити зв'язки або залежності між словами в реченні наприклад, яке слово є головним (parents), а яке підпорядковується (dependents). Результатом є граф, що відображає структуру речення зі стрілками, що показують залежності між словами.

Парсинг залежностей використовують для вирішення завдань машинного перекладу, де важливо розуміти синтаксичну структуру природної мови для коректного перекладу на цільову мову та визначення настроїв в тексті, що використовує залежності між словами, допомагаючи краще зрозуміти контекст і зміст фраз. [8]

Парсинг залежностей може стикатися з викликами, такими як синтаксична двозначність, коли одне слово може мати кілька ролей залежно від контексту, а також обробка ідіом і стійких виразів, що можуть не слідувати загальним граматичним правилам. [8] Це робить даний метод надзвичайно важливим для розробки ефективних систем NLP.

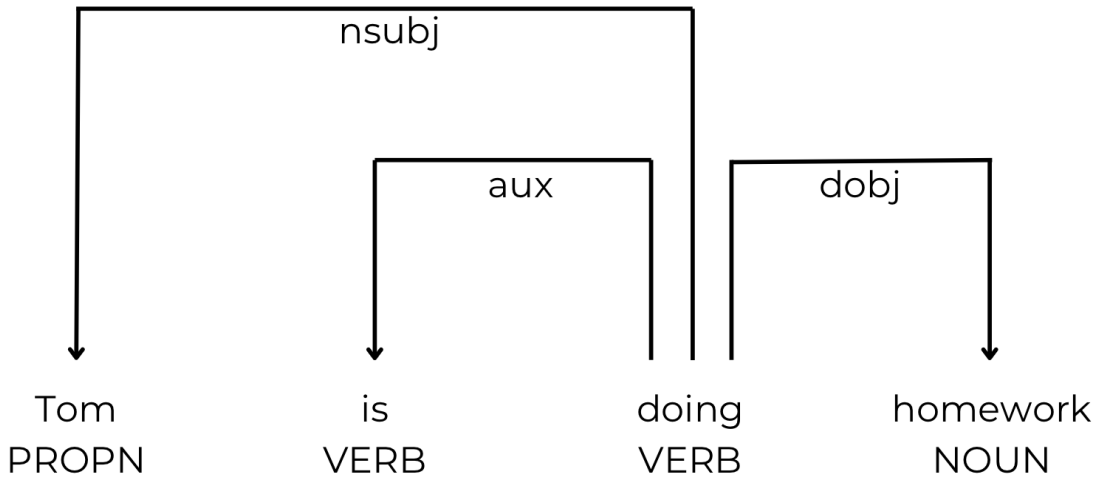

Рисунок 1.5 – Приклад метода DP для речення “Tom is doing homework”.

На Рисунку 1.5 зв’язки представленні у вигляді дерева, де “Tom” - підмет (subject, скорочено “nsubj”). Це основний елемент, навколо якого будується інша частина речення. “is” - допоміжне дієслово (auxiliary verb, “aux”), яке показує час дії та входить в склад складеного присудка. “doing” – дієслово (main verb, “ROOT” у дереві залежностей), яке вказує на дію, що виконується підметом. “homework” – прямий додаток (direct object, “dobj”), який отримує дію і пов'язаний з дієсловом “doing”. У синтаксичному дереві залежностей “Tom” буде з'єднаний з “doing” як підмет, “is” також буде з'єднаний з “doing” як допоміжне дієслово, а “homework” буде з'єднаний з “doing” як прямий додаток.

Синтаксичний аналіз є критично важливим етапом у розумінні структури та значення тексту. Розбір залежностей, як метод синтаксичного аналізу, надає можливість для глибшого розуміння мовних конструкцій, дозволяючи виявити ієрархічні залежності між словами.

1.5 Семантичний аналіз

Розуміючи синтаксичні структури, можна починати заглиблюватись у більш складнішу проблему – визначення “значення” речення. Семантичний аналіз є важливим компонентом обробки природної мови, який зосереджується на розумінні значення, інтерпретації та взаємозв'язків між словами, фразами та реченнями в заданому контексті [9]. Це виходить за рамки синтаксичного аналізу речення і заглиблюється в передбачуване (intended) значення. Такий підхід дозволяє нам витягувати цінну інформацію та втілювати здатність до глибшого розуміння текстів.

Нам легко зрозуміти значення та категорії текстів, що свідчить про інтуїтивне розуміння навколишнього світу, проте для комп’ютерів, через присутність неоднозначності, це перетворюється на складний виклик для розробників. Для таких завдань, ідеально підходить named entity recognition (NER). NER – метод NLP, який відповідає за ідентифікацію та класифікацію конкретних даних із певного текстового вмісту [10]. Даний метод, працює із іменованими сутностями (якими можуть бути як слова, так і їх послідовність) за для визначення їх у заздалегідь визначені категорії.

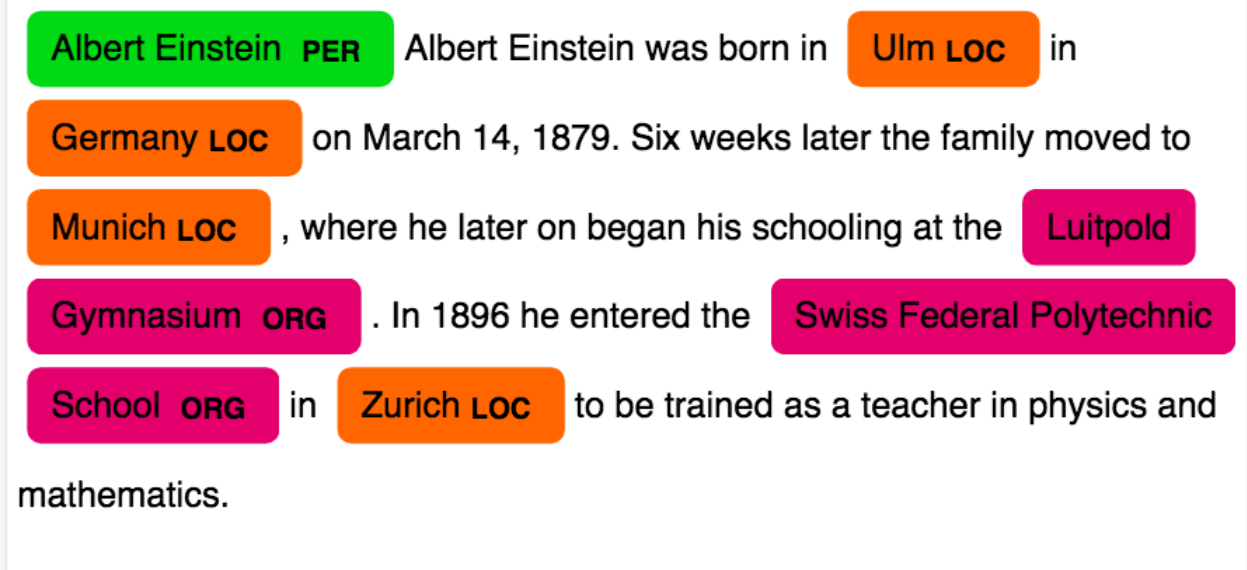

Рисунок 1.6 – Представлення NER у тексті. [10]

На Рисунку 1.6, можемо побачити приклад використання NER у тексті, у якому вдалось класифікувати сутності за трьома класами: PER - person, LOC - location, ORG - organization. Задача NER є ідеальним рішенням для ідентифікації елементів у тексті, які вказують на осіб, об’єкти, місця та часові вказівки. Його також активно використовують у наш час, зокрема у відділах підтримки клієнтів, де компанії отримують велику кількість відгуків щодо команди чи продукту [10]. Він визначає об’єкти у відповідних скаргах і класифікує їх, автоматично направляючи до потрібного відділу. Це дозволяє компаніям створювати автоматизовану систему, яка направляє запити клієнтів до відповідної служби підтримки. Також даний метод знайшов своє прикладне застосування у фільтруванні резюме [10]. Спеціальні навички можна використовувати як сутності для систем NER у процесах найму, таким чином, значно оптимізувавши процес найму.

Загалом, можна виділити два основних кроки у NER: ідентифікація сутностей у тексті та їх категоризація в окремі групи [10]. Виявлення сутності, є початковою фазою процесу, що передбачає ідентифікацію фрагментів

тексту, які потенційно представляють значущі сутності. Класифікація сутностей передбачає віднесення виявлених об'єктів до конкретних, попередньо визначених категорій на підставі їхнього семантичного значення. Категорії можуть охоплювати різноманітні аспекти, від людей та організацій до місць розташування, дат, тощо. Це вимагає глибокого розуміння контексту, в якому ці об'єкти з'являються, бо, наприклад, термін “Apple” у контексті технологій може вказувати на технологічну компанію, а у кулінарному - вказує на фрукт.

Підходи до NER - це різні стратегії, які використовуються для ідентифікації іменованих сутностей у тексті. Ці підходи можна поділити за основними принципами, зокрема: rule-based, learning-based та hybrid. [11]

а) Підходи на основі правил (rule-based approaches) полягають у використанні набору правил, розроблених експертами, що базуються на синтаксично-лексичних шаблонах на лінгвістичних знаннях. [11] NER створені на основі правил, направлені на особливості конкретної галузі для досягнення високої точності. Проте, такі системи мають недоліки: вони вимагають значних витрат та є специфічними для обраного домену (системи не можуть бути адаптовані для використання в інших галузях).

б) Підходи, засновані на навчанні (learning-based approaches), включають в себе використання алгоритмів машинного навчання [11]. Ці методи діляться на три категорії:

- Навчання з вчителем (Supervised Learning): Техніки цієї категорії базуються на тренуванні машин за допомогою маркованих навчальних даних, щоб передбачити результати для вхідних (нових) даних. Моделі здатні розпізнавати патерни та правильно класифікувати набори даних. Приклади таких систем включають Hidden Markov Model (HMM), Support Vector Machine (SVM) і Maximum Entropy Markov Model. [11]

- Частково-навчені підходи (Semi-Supervised Learning): Ці методи поєднують невелику кількість маркованих даних із великою кількістю немаркованих даних [11]. Вони використовують bootstrapping - це техніка повторної вибірки, яка допомагає оцінити невизначеність статистичної моделі [12]. Метод включає вибірку оригінального набору даних із заміною та генерацію декількох нових наборів даних того ж розміру, що і оригінал.

- Навчання без вчителя (Unsupervised Learning): Алгоритми навчання без вчителя використовують немарковані дані. Два основні підходи - кластеризація, що базується на статистичних даних для виявлення іменованих сутностей на основі схожості контексту та асоціативні правила [11].

в) Ще один підхід до NLP – гібридний. Він є комбінацією найкращих правил машинного навчання (наприклад, learning-based) та правил, заснованих на людському досвіді (human expertise) [11]. Існує чимало гібридних моделей NER, оскільки вони є більш гнучкими порівняно з іншими.

У приклад NER-системи можна привести дослідження Н. Синтаеху (2020) [13] , який вивчав методологію, яка порівнює два частково-навчені підходи LP (label propagation) та EM (expectation maximization) для Ethiopian News Агентство (ENA). Запропонована методологія була спрямована на вирішення завдання класифікації об'єктів за допомогою NER для амхарської мови, що є складною проблемою для методів навчання з вчителем коли набір даних великий. Зазначено, що запити обробляються на даних Ethiopian News Agency (ENA). Цей текстовий корпус складається з 4700 речень, приблизно по 83 слова у кожному реченні. У результаті використання методу було визначено п'ять ключових тем: “Людина”, “Гроші”, “Організація”, “Дата” і “Локація”. Результати оцінки підходу EM показали ефективність на рівні 64%, тоді як підхід LP демонстрував більш високу точність - 79% при використанні 100% маркованого набору даних [13]. Це свідчить про потенційну

ефективність обраних методів у розв'язанні завдань NER для амхарської мови на великому обсязі даних.

Ще одним методом семантичного аналізу є сентимент аналіз (SA) - який визначає емоційне забарвлення у текстах. Він визначає точку зору або емоцію, що стоїть за ситуацією. У основному це означає проаналізувати та знайти емоцію або намір, що стоїть за фрагментом тексту або мови або будь-яким способом спілкування. [14]

Розглянемо приклад мережі закладів, які вони продають різноманітні продукти. Вони створили веб-сайт для продажу своєї їжі, і тепер клієнти можуть замовити будь-який продукт харчування і надавати відгуки, наприклад, чи сподобалася їм їжа, чи ні. З відгуків компанія може зробити висновки та працювати над своїми недоліками. Проте, кількість таких відгуків дуже велика і це вимагає автоматизації, з чим якраз і допомагає SA.

Ось кілька найчастіше використовуваних технік, які використовують для аналізу настроїв:

а) Лексичний аналіз - полягає у використанні попередньо маркованих словників для оцінки тексту. [15] У даному методі текст розбивається на токени, і кожен токен порівнюється зі словником. Якщо слово відповідає позитивному тегу у словнику, то до загального рейтингу тексту додається позитивний бал, а якщо негативний - бал знижується. Незважаючи на простоту методу, його модифікації показують значний успіх у класифікації емоційних відтінків тексту. Використання лексичного аналізу дозволяє швидко обробляти текст на основі статичного набору правил. [15]

б) Використання машинного навчання для SA полягає у навчанні на позначених даних і класифікацію нових даних [15]. У цьому методі використовуються алгоритми, такі як SVM (Support Vector Machines) та наївний байєсівський класифікатор, для створення моделей, які можуть

автоматично аналізувати настрої у тексті [15]. Перевага цього підходу полягає в його здатності до адаптації та здобуванні високої точності, але він вимагає великої кількості анотованих даних для обробки.

в) Гібридний підхід поєднує лексичний аналіз та машинне навчання для оптимізації точності та швидкості. У одному з варіантів використовується двослівний лексикон, який ділиться на позитивні та негативні класи [15]. Створюються псевдодокументи, що включають всі слова з вибраного лексикону, а потім обчислюється косинусна подібність між псевдодокументами та немаркованими текстами. Залежно від ступеня подібності текстам присвоюється позитивний чи негативний настрій, які згодом використовуються для навчання наївного байєсівського класифікатора [15]. Такий підхід дозволяє використовувати переваги обох методів, зберігаючи високу точність машинного навчання та швидкість лексичного аналізу.

Семантичний аналіз - ключовий компонент для сучасних технологій NLP, що відкриває безліч можливостей для розвитку інтелектуальних систем. NER виявляє ключові елементи в тексті, спрощуючи подальший аналіз та категоризацію. SA, у свою чергу, виявляє емоційні компоненти тексту, що дозволяє визначити відношення до певних аспектів. Завдяки такому аналізу, виникає можливість здійснення глибокого розуміння природної мови, що покращує різні сфери обслуговування та дає бізнесу аналітику про їхніх користувачів

1.6 Прагматичний аналіз

Прагматичний аналіз – це фаза розуміння природної мови, що пов’язана з інтерпретацією мови в контексті. Це виходить за рамки буквального значення слів, щоб зрозуміти намір оратора, інтерпретацію одержувача та

контекст розмови [16]. Це дозволяє відтворювати і реагувати на нюанси в комунікації, які залежать від культурних, соціальних та ситуаційних чинників.

Людини щодня генерують надзвичайно велику кількість цифрових даних. Важливо аби ці дані були переглянуті та віднесені відповідно до певної теми. Проблема у тому, що таке завдання є нездійсненим для людини. У таких випадках, на допомогу приходить моделювання тем. Topic Modeling – це метод NLP, який використовується для автоматичного виявлення тем в колекціях текстових документів [17]. Головною метою є ідентифікація груп слів або тем, які часто спільно зустрічаються в текстах, щоб визначити приховану структуру в даних. Більшість алгоритмів, що виконують таке завдання, розкладають матрицю термінів документа на дві або більше. Залежно від алгоритму, записи в матриці термінів документа можуть бути розраховані за допомогою підходу мішків слів (bags-of-words), зворотної частоти документів (TF-IDF) або TF-IDF на основі класів. Розподіл тем у документах та термінів темі можуть бути імовірнісними або детермінованими. Найбільш класичними алгоритмами є латентне розміщення Діріхле (LDA) та невід'ємна матрична факторизація (NMF) [17].

Латентне розміщення Діріхле (LDA) - інструмент для аналізу змісту великих текстових корпусів, який базується на ймовірнісних припущеннях [17]. Алгоритм базується на гіпотезі, що кожен документ є набором з різних тем, які визначаються розподілом слів. Такий підхід дозволяє виявляти приховані (латентні) тематичні структури в текстових даних. Модель починає з припущення, що існує набір "тем", з яких кожен документ "вибирає" свою унікальну комбінацію. Цікаво, що скільки результати не є детермінованими, тому вони можуть бути різними кожен раз, коли запускається модель, навіть на одному наборі даних [17].

Невід'ємна матрична факторизація (NMF) розкладає невід'ємну матрицю на дві невід'ємні матриці, де кожен рядок є темою, а кожен стовпець - документом [17]. NMF припускає, що кожен документ є лінійною комбінацією тем, а кожна тема є лінійною комбінацією термінів. Метою NMF є зменшення розмірності та вилучення ознак. Оригінальна матриця розкладається на матрицю ознак і матрицю коефіцієнтів. Метод найкраще підходить для менших наборів даних та коротких текстів. [17]

Системи автоматизованих відповідей на запитання (Q&A) відкривають нові можливості для взаємодії між людиною та машиною. Вони генерують відповіді на питання, задані людьми у вільній формі, шляхом аналізу великих обсягів текстової інформації. Процес роботи таких систем включає декілька основних етапів: попередню обробку тексту, розуміння поставленого питання, пошук інформації та генерацію відповіді, а також ранжування відповідей за релевантністю [18].

Попередня обробка тексту передбачає стандартизацію формату запитання, видалення нерелевантної інформації, маркування, лематизацію та видалення стоп-слів. На наступному етапі аналізується оброблений текст, щоб визначити ключові поняття та типи запитань [18]. Використання таких методів NLP, як NER, DP і PoS tagging є важливим для ефективного вилучення необхідної інформації із запитань. Пошук інформації здійснюється шляхом перетворення запитань на запити для пошуку в базах даних і корпусів текстів для вибору найбільш релевантної інформації. Метод ґрунтується на пошуку за ключовими словами чи семантиці, залежно від специфікацій системи запитань і відповідей [18]. Після створення відповідей ця вибрана інформація аналізується для отримання конкретних результатів (відповідей на запитання). Останній крок, ранжування, визначає релевантність отриманих відповідей, щоб надати користувачеві найбільш точну інформацію [18]. Ефективність

такої системи залежить від багатьох факторів, включаючи якість вхідних даних, ефективність попередньої обробки та точність моделі.

Для навчання моделей Q&A необхідно зібрати великий набір даних із питаннями та відповідями, що дозволяє системі ефективно вирішувати поставлені задачі. Основні типи систем автоматизованих відповідей на запитання складають:

а) Системи Q&A, що базуються на інформаційному пошуку (IR). Вони здійснюють пошук за ключовими словами або семантичний пошук, для ідентифікації найбільш релевантних документів. Цей підхід є відносно простим у реалізації та широко використовується. Однак, актуальність індексованого тексту, разом із ефективністю методів пошуку інформації значно впливає на продуктивність таких систем [18].

б) Системи автоматичної відповіді на питання можуть використовувати також базу знань, наприклад онтологію, для пошуку відповідної інформації. Такі системи Q&A більш точні та надійні, завдяки використанню добре організованих знань, проте їх продуктивність обмежується охопленням даних [18].

в) Генеративні системи Q&A автоматично генерують відповідь на задане питання використовуючи нейронні мережі. Метод базується на ідеї, що машину можна навчити розуміти та відтворювати текст на природній мові, щоб надавати коректну відповідь. Такі системи є потужними, оскільки здатні відповідати на широкий спектр питань і генерувати відповіді, які схожі на людські [18].

г) Гібридні системи Q&A здатні поєднувати усі вище перераховані методи, за для покращення загальної продуктивності. Даний підхід керується припущенням, що різні методи мають свої сильні та слабкі сторони, і їх комбінація може покращити ефективність системи. Проте, їх проектування та

реалізація вимагає більше ресурсів, ніж створення на основі одного підходу [18].

Незважаючи на складність і вимоги до обчислень, роль прагматичного аналізу в NLP - незамінна для розробки більш інтуїтивно зрозумілих і точних систем обробки мови. У епоху, коли соціальні мережі охоплюють різні людські сфери, здатність систем розуміти та відтворювати тексти цілком є критично важливим, ніж будь-коли.

1.7 Оцінка якості

Після дослідження різних методів аналізу у галузі NLP, виникає питання про оцінку якості кожної моделі. Для цього використовуються спеціальні метрики, які дозволяють оцінити їх ефективність. Важливо розуміти основні поняття, такі як істинно позитивні (TP), хибно негативні (FN), хибно позитивні (FP) та істинно негативні (TN) результати:

а) TP (True Positive) - кількість правильно ідентифікованих позитивних об’єктів [19].

б) TN (True Negative) - кількість правильно ідентифікованих негативних об’єктів [19].

в) FP (False Positive) - кількість неправильно ідентифікованих позитивних об’єктів [19].

г) FN (False Negative) - кількість неправильно ідентифікованих негативних об’єктів [19].

Однією з найпростіших і широко використовуваних метрик є правильність (accuracy) - це частка TP від загальної кількості екземплярів [19]. Обчислюється за допомогою формули:

$$(TP + TN) / (TP + TN + FP + FN).$$

У завданнях NLP правильність використовується для вимірювання загальної продуктивності моделі. Якщо з 100 електронних листів, 95 були правильно класифіковані моделлю, то правильність буде 95 / 100 = 0.95.

Точність (precision) - це частка TP до загальної кількості випадків, визначених як позитивні [19]. Обчислюється за допомогою формули:

$$TP / (TP + FP).$$

У завданнях NLP точність використовується для вимірювання того, скільки випадків, визначених як позитивні, насправді є позитивними. Якщо з 100 електронних листів модель визначила 98 як позитивні, а насправді лише 95 з них були дійсно позитивними, то точність буде 95 / 98 ≈ 0.9694.

Повнота (recall) - це TP до загальної кількості випадків, які насправді є позитивними. Формула для обчислення така:

$$TP / (TP + FN).$$

Дана метрика використовується для вимірювання того, скільки позитивних випадків було правильно ідентифіковано моделлю. Якщо з 100 електронних листів модель правильно визначила 90 як позитивні з усіх 100 дійсно позитивних, то повнота буде 90/100 = 0.9.

F1-міра - гармонійне середнє між значеннями точності та повноти, і використовується для їх збалансування та обчислюється за допомогою формули:

$$2 * (precision * recall) / (precision + recall) [19].$$

Метрика обчислюється в межах від 0 до 1, де 1 - найкращий можливий бал. У завданнях NLP F1-score використовується для оцінки загальної продуктивності моделі. Використовуючи значення точності та повноти з попередніх прикладів, можемо обчислити F1-міру: 2 * ((0.9694 * 0.9) / 0.9694 + 0.9)) ≈ 0.933.

Усі ці метрики широко використовуються, зокрема для завдань SA, NER та класифікація тексту [19].

Для завдань, які вимагають машинний переклад та узагальнення тексту, використовуються метрики BLEU (Bilingual Evaluation Understudy) та ROUGE (Recall-Oriented Understudy for Gisting Evaluation) відповідно [20].

Двомовне оцінювання (BLUE) - вимірює схожість між машинним перекладом тексту та довідковими перекладами з використанням n-грамів, які є суміжними послідовностями з n слів [20]. Найпоширенішими n-грамами, що використовуються - це уніграми (єдині слова), біграми (послідовності з двох слів), триграми (послідовності з трьох слів). Метрика BLEU обраховується за допомогою формули:

$$BLEU = BP * \exp(\sum pn).$$

BP (Brevity Penalty) - це так званий penalty term, або штраф, який застосовується до оцінки якості машинного перекладу, щоб компенсувати короткі переклади порівняно з референсними перекладами [20]. Він розраховується як:

min(1, (reference_length / translated_length)), де:

reference_length - це загальна кількість слів у довідкових перекладах, а translated_length - загальна кількість слів у машинному перекладі.

pn - це точність n-грамів, яка розраховується як кількість n-грамів, які з'являються у машинно-згенерованому перекладі і в довідкових перекладах, поділені на загальну кількість n-грамів у машинно-згенерованому перекладі [20].

Оцінка BLEU коливається від 0 до 1, де 1 - найкраща якість перекладу. BLEU широко використовується в завданнях машинного перекладу, оскільки забезпечує ефективний спосіб оцінки їх якості.

Метрика ROUGE - розроблена для оцінки якості висновків, які згенеровані машинами, порівнюючи їх з довідковими, наданими людьми [20]. Вона вимірює схожість з використанням overlapping n-grams. Найпоширенішими n-грамами - уніграми, біграми та триграми. Формула оцінки така:

$$ROUGE = \sum (\text{overlapping n-grams}), \text{ де:}$$

Overlapping n-grams - це кількість n-грамів, які з'являються як у машинно-зведеному, так і в довідкових висновках, поділених на загальну кількість n-грамів [20].

Оцінка ROUGE варіюється від 0 до 1, де 1 – найкраща зведена якість. Вона часто застосовується в завданнях узагальнення тексту, забезпечуючи спосіб об'єктивної оцінки якості машинно створених узагальнень порівняно з довідковими.

Ще одним способом оцінки моделей NLP є перевірка їх здатності прогнозувати наступний токен, що особливо важливо для завдань моделювання мови, генерації тексту та розпізнавання мовлення. Метрики, такі

як перплексія та логарифмічна ймовірність, кількісно визначають невизначеність моделі, коли вона стикається з новими даними. Перплексія вимірює обернену ймовірність даних, показуючи скільки виборів в середньому модель повинна здійснити, щоб передбачити наступний токен [21]. Логарифмічна ймовірність вимірює ймовірність генерації спостережуваних даних моделлю. Нижча перплексія та вища логарифмічна ймовірність вказують на кращу продуктивність моделі [21].

Наостанок, інколи найкращим способом вимірювання якості в NLP є звернення до людей за їхніми думками [21]. Даний спосіб оцінювання доповнює усі вищезгадані метриками, оскільки люди можуть надати більш суб'єктивні оцінки природної мови, такі як читабельність, зв'язність, стиль, тон та емоційне забарвлення. Людську оцінку можна провести різними способами, такими як шкали оцінювання, ранжування, попарне порівняння або відкриті питання [21].

Метрики відіграють ключову роль у вимірюванні продуктивності моделей, кожна з яких, відображає різні аспекти якості. Жодна метрика не може повністю описати цілком якість моделі, тому важливо використовувати їхню комбінацію для коректного визначення ефективності.

1.8 Проблеми та майбутнє

NLP - сфера, яка швидко розвивається, а також відкриває безліч можливостей для взаємодії між людьми та машинами. Розглянемо очікуване майбутнє NLP у 2024 році, а також ключові можливості:

а) Завдяки популярності голосових помічників, таких як Alexa, Siri та Google Assistant, NLP відкрила нові можливості для технології голосового інтерфейсу. У подальших роках можемо очікувати подальшого прогресу в

голосових інтерфейсах, що зробить їх більш інтуїтивно зрозумілими та сприйнятливими до команд природної мови. [22]

б) Зі зростанням світових ринків зростає потреба в системах NLP, які можуть обробляти кілька мов, тому очікується розвиток багатомовних систем NLP, які можуть розуміти та обробляти мови з усього світу. [22]

в) Чат-боти стають все більш популярними в таких галузях, як обслуговування клієнтів та електронна комерція, тому очікуємо побачити чат-ботів, які є більш складними та здатними обробляти складні запити завдяки NLP. [22]

Тепер розглянемо основні виклики та проблеми:

а) Моделі NLP можуть бути упередженими (bias) щодо певних груп або демографічних характеристик, що призводить до несправедливих результатів [22]. Наприклад, Gemini створив історично неточні та потенційно образливі зображення, припускаючи диверсифікацію особистостей у Вермахті (німецькій армії часів Другої світової війни), включаючи представників різних рас, що нереалістично для того історичного контексту [23]. Це могла бути спроба боротьби з існуючими упередженнями у навчальних даних, але такий підхід виявився неефективним і викликав чимало критики.

б) Збір великих обсягів даних у сфері NLP породжує питання щодо конфіденційності [22]. Її порушення може призвести до серйозних проблем, включаючи неправомірний доступ до особистих даних або витік інформації.

в) Що стосується моделей NLP, які стають все складнішими, може виникати проблема у зрозумінні, як вони приймають певні рішення, що ускладнює визначення причин та наслідків деяких дій, зроблених моделями NLP. [22]

г) Розвиток складних моделей вимагає значних обчислювальних ресурсів, які не завжди доступні, що призводить до створення бар'єрів для дослідників, які не мають достатньої обчислювальної потужності. [22]

д) Недостатня якість даних призводить до неправильних результатів обробки природної мови.

Даний розділ, дає повну картину процесу NLP, разом із визначенням ключових понять, оглядом історії, детальним розглядом усіх аспектів аналізу із їхніми основними завданнями та оцінюванням якості даних систем. Прогнозоване майбутнє підкреслює важливість вирішення поставлених проблем, за для розвитку сфери NLP. Цей розділ створює необхідну базу знань для подальшого розгляду прикладних інструментів NLP для англійської та української мов.

2. АНАЛІЗ NLP ДЛЯ АНГЛІЙСЬКОЇ МОВИ

2.1 Огляд State-of-the-art Моделей

Застосування трансферного навчання (transfer learning) та попередньо навчених мовних (pre-trained language) моделей відкриває нові можливості для сфери обробки природної мови (NLP) [24]. Трансферне навчання полягає в тому, що модель, навчена для виконання однієї задачі, може застосуватися також для вирішення інших, причому задачі можуть бути як схожими так і різними [25]. За допомогою цього методу, можна ефективно використовувати набуті знання (з попередніх моделей) для розв'язання нових проблем без необхідності перенавчання з нуля. Попередньо навчені мовні моделі - ключовий компонент трансферного навчання. Наприклад, моделі BERT, GPT, RoBERTa навчаються на великих корпусах текстів, що дозволяє їм вивчити різноманітні лінгвістичні особливості та властивості мови.

Трансформери - представляють собою архітектуру нейронних мереж, основною перевагою яких є здатність ефективно працювати з послідовностями будь-якої довжини та засвоювати залежності в тексті без необхідності використання рекурентних або зворотних зв'язків, як це було раніше у звичайних рекурентних нейронних мережах (RNN) [26]. Вони стали основною тенденцією в нових дослідженнях, адже дозволяють будувати дуже гнучкі мовні моделі. У той же час, виникла дискусія щодо значення досягнень великих попередньо навчених мовних моделей, які займають провідні позиції в рейтингах. Хоча багато експертів погоджуються з статтею професора копенгагенський університету - Анни Роджерс[27], про те, що досягнення найкращих результатів завдяки лише більшій кількості даних та обчислювальній потужності не є справжнім науковим проривом, інші фахівці з сфери NLP вказують на деякі позитивні аспекти в поточному напрямку,

наприклад, можливість виявлення основних обмежень поточної парадигми. [28]

У будь-якому випадку, останні вдосконалення, зумовлені не тільки значним збільшенням обчислювальної потужності, але й виявленням геніальних способів полегшення моделей зі збереженням високої продуктивності.

Мовна модель BERT (Bidirectional Encoder Representations from Transformers) - це фреймворк машинного навчання з відкритим кодом для обробки природної мови (NLP), яка натренована на текстовому корпусі з Вікіпедії і здатна до підлаштування (fine-tuned) за допомогою наборів даних для запитань і відповідей (Q&A datasets) [29]. Модель базується на трансформаторах глибокого навчання, в якій кожен вихідний елемент підключений до кожного вхідного елемента, а ваги між ними динамічно розраховуються на основі їх з'єднання. Раніше мовні моделі зчитували вхідний текст лише послідовно - зліва направо або справа на ліво, а одночасно не могли. Унікальність BERT полягає саме у можливості читання у обох напрямках одночасно (bidirectional encoder). Використання двонаправленості дає можливість попередньо навчитись на двох різних завданнях NLP – masked language modeling (MLM) та next sentence prediction (NSP) [29]. Ідея маскованого моделювання (MLM) полягає у прихованні або зміненні слова у реченні, для перевірки, чи зможе модель виправити помилку на основі заданого контексту. Для передбачення наступного речення (NSP), мета полягає в тому, щоб програма передбачила, чи певні два речення мають логічний, послідовний зв'язок, чи їх взаємозв'язок просто випадковий.

Як уже вище згадувалось, ідея будь-якої техніки NLP полягає у розумінні людської мови. Для BERT – це передбачення слова в певному пропущеному місці. Для тренування такі моделі використовують велике

сховище позначених (labeled) навчальних даних, розроблених лінгвістами, які роблять складне маркування вручну [29].

BERT, попередньо навчений на немаркованих колекціях Вікіпедії та Brown Corpus. Навчання продовжується навіть під час використання у продакшені, зокрема в Google search, за допомогою неконтрольованого навчання (unsupervised learning) і вдосконалюється [29]. BERT може адаптуватися до постійно зростаючого обсягу пошукових запитів та точно налаштувати на специфікації користувача за допомогою трансферного навчання.

Bidirectional Encoder Representations from Transformers (BERT) є open-source проектом, тобто будь-хто може ним користуватися. Сама компанія “Google” стверджує, що користувачі можуть навчити найсучаснішу систему Q&A всього за 30 хвилин на блоці обробки хмарних тензорів і за кілька годин за допомогою графічного блоку обробки [29]. Багато інших організацій, дослідницьких груп та окремих фракцій Google тонко налаштовують архітектуру моделі за допомогою контрольованого навчання, щоб або оптимізувати її для ефективності, або спеціалізувати для конкретних завдань, попередньо навчаючи BERT з певними контекстними уявленнями. [29] Розглянемо приклади з Таблиці 1:

Модель	**Призначення**
DocBERT	Класифікація документів
VideoBERT	Навчання на немаркованих даних YouTube
SciBERT	Науковий текст
G-BERT	Медичні рекомендації
TinyBERT	Оптимізація ефективності
DistilBERT	Спрощена версія BERT
ALBERT	Зменшення споживання пам'яті

Модель	Призначення
RoBERTa	Покращення ефективності навчання
ELECTRA	Високоякісне представлення тексту

Таблиця 2.1 – Моделі BERT та їхнє призначення. [29]

BERT – інноваційний фреймворк, який надає потужний та зручний інструмент розробникам. Його розробка, фактично започаткувала нову еру в обробці природної мови, значно покращивши попередні методи.

GPT (Generative Pre-trained Transformers) - це моделі, які здатні аналізувати, витягувати, узагальнювати та іншим чином використовувати інформацію для створення контенту [30]. Одним з найбільш відомих випадків використання моделі GPT є проект Chat-GPT, що здатний імітувати природній діалог та відповідати на поставлені, навіть вузькоспеціалізовані запитання. З 2018 року, компанія OpenAI випустила чотири версії GPT:

а) GPT-1: Перша версія моделі GPT, випущена в 2017 році, використовувала архітектуру Transformer для генерації тексту [30]. GPT-1 став першим успішним прикладом глибокого навчання для природної мови та зумів відтворювати реальний текст, що викликало значну зацікавленість у спільноти.

б) GPT-2: Друга модель, випущена у 2019 році, мала значно більшу потужність та швидкодію, порівняно з GPT-1. Маючи теж архітектуру Transformer, вона тепер вирізнялася своїм великим розміром (1,5 мільярда параметрів) на момент випуску [31]. Ця модель була попередньо навчена на наборі даних WebText, який складається з текстів, вилучених з 45 мільйонів веб-сторінок. GPT-2 в основному дотримується архітектури попередньої моделі GPT, але має деякі модифікації:

1) Нормалізація шару тепер застосовується до вхідних даних кожного підблоку і додаткова нормалізація за рівнем була додана після остаточного self-attention блоку. [31]

2) Застосовується модифікована ініціалізація, яка враховує накопичення на залишковому шляху залежно від глибини моделі. Ваги залишкових шарів масштабуються при ініціалізації за кількість залишкових шарів. [31]

3) Розмір словника розширено до 50,257. Розмір контексту тепер становить 1024 токени, що вдвічі більше, ніж у попередній версії, і використовується більший розмір партії (batch) - 512 [31].

в) GPT-3: Вона випущена в 2020 році і представляє собою значний прорив у сфері машинного навчання. GPT-3 навчений на набагато більшому корпусі даних (45 терабайтів), ніж його попередники, та має вражаючу кількість параметрів для тренування – 175 млрд [32].

Dataset	**Quantity (tokens)**	**Weight in training mix**	**Epochs elapsed when training for 300B tokens**
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Рисунок 2.1 – Корпуси даних GPT-3 [32]

На Рисунку 2.1 зображені основні корпуси даних, на яких модель навчалась. Корпус Common Crawl містить петабайти даних, зібраних протягом 8 років сканування веб-сторінок. У корпусі є сира (raw) інформація про веб-сторінки, екстракти метаданих та тексти з легким фільтруванням. WebText2 - це тексти веб-сторінок з усіх вихідних посилань Reddit з повідомлень із трьома

або більше голосами (upvotes). Books1 та Books2 - це два корпуси книг з Інтернету. Сторінки Вікіпедії англійською мовою також є частиною навчального корпусу. Третій стовпчик у таблиці “Weight in training mix” вказує на частку прикладів під час навчання, які взято з певного набору даних.
GPT-3 здатна генерувати текст, який складно відрізнити від людського та демонструє вражаючі результати у багатьох завданнях природньої мови.

г) GPT-4: Четверта версія моделі, зуміла досягти ще більших висот у сфері штучного інтелекту. Вражає кількість параметрів, що становить 1.76 трлн, та можливість опрацювання одразу до 25,000 слів, що у 8 разів більше ніж GPT-3 [33]. Архітектура даної моделі приховується. Спочатку вважали, що OpenAI це зробило через стрімкий розвиток і потенційні ризики для людства, проте звіт від SemiAnalysis розкриває деталі про модель, вказуючи на те, що компанія тримає архітектуру GPT-4 закритою через можливість плагіату [33]. Витрати лише на навчання моделі становлять понад 60 млн доларів. Архітектура виведення використовує кластер з 128 GPU з паралелізмом тензорів та каналів. Крім того, GPT-4 містить vision encoder для читання веб-сторінок та транскрибування зображень і відео, що додає більше параметрів для додаткового налаштування. [33] GPT-4 є мультимодальною моделлю, тобто дозволяє працювати не лише з текстовими даними, як попередники, а ще й з іншими типами, зокрема зображеннями. Ця модель вражає своєю здатністю розуміти та генерувати складні тексти, що робить її дуже корисною у різних сферах, від машинного перекладу до генерації контенту.

BERT та GPT відкрили нову еру в розвитку обробки природної мови, відзначившись значними досягненнями в цілих наборах завданнях, показавши надзвичайну здатність розуміти та генерувати текст, що дозволило виконувати завдання з вражаючою точністю та ефективністю. Однак, існують виклики, такі як подолання обмежень, пов'язаних з обчислювальною складністю та

потребою у великій кількості даних для навчання. Тому, для подальшого розвитку NLP потрібно продовжувати вдосконалювати дані моделі для вирішення основних проблем.

2.2 Корпуси даних

Набори даних – це колекції текстової інформації, що використовуються для тренування, оцінювання та тестування моделей. Як правило, вони формуються людськими анотаторами, що підкреслює складність та часозатратність їх створення. Між собою вони різняться за розміром, складністю та спеціалізованістю, задовольняючи конкретні потреби. Наявність таких корпусів має значний вплив на розвиток та вдосконалення моделей для природніх мов. Великий обсяг та різноманітність даних дозволяють ефективно навчати моделі, що покращує їхню точність та роботу в продукційному середовищі. Крім того, корпуси даних дозволяють дослідникам експериментувати з новими методами NLP, розвивати нові алгоритми та архітектури моделей.

Для англійської мови ситуація виглядає яскравіше та динамічніше, ніж для більшості мов, адже майже усі найновіші дослідження, які з’явились в останніх десятиліттях сконцентровані на обробці саме англійської. Розглянемо зокрема SQuAD (Stanford Question Answering Dataset) - це набір даних для моделей відповідей на питання (Q&A), який містить реальні запитання, поставлені людьми в наборі статей Вікіпедії, де відповіддю є певний текст у відповідній статті [34]. Набір даних розділений на тренувальний (train) та тестовий (test). Моделі оцінюються на тестовому наборі і їх продуктивність вимірюється за допомогою таких показників, як точний збіг (EM) та F1-міра (Рисунок 2.2).

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948

Рисунок 2.2 – Оцінка моделей на SQuAD [34]

SQuAD відіграє вирішальну роль у просуванні сфери Q&A, створюючи загальний стандарт для порівняння та оцінки продуктивності різних систем [35]. Його реалізація була одним із найскладніших, а результати BERT на ньому привернули увагу спільноти, що сприяло подальшому розвитку NLP. Важливим є необхідність створення схожих корпусів для інших мов, оскільки це надає важливі ресурси для розвитку сфери та розширює можливості застосування технологій NLP в різних культурних та лінгвістичних контекстах.

Текстові корпуси також публікують різні наукові організації, наприклад, Allen Institute for AI (AI2) [36], зібрала понад 80 сторінок різних корпусів, які наразі в загальному доступі (open-source), що надає розробникам значні

переваги. По-перше, така велика колекція різноманітних даних, дозволяє розширити можливості навчання та тестування, надаючи інструмент, який дає можливість для покращення продуктивності моделей. Окрім цього, наявність даних у відкритому доступі приводить до зростання наукової спільноти, адже розробники з усього світу можуть використовувати ці безкоштовні дані для власних досліджень.

Також важливо відзначити платформу Kaggle [37], яка створена як для змагань, так і для обміну даними. Там можна знайти широкий спектр датасетів не лише з NLP, а й різних галузей. Вона відкриває безліч можливостей для дослідників, які у свою чергу можуть скористатися доступними даними для навчання та тестування своїх моделей. Також, організація змагань сприяє розвитку нових інноваційних рішень у цій галузі.

Hugging Face [38] відіграє ключову роль, надаючи доступ до великого арсеналу моделей та датасетів. Розроблені тут інструменти дозволяють дослідникам легко експериментувати, вдосконалюючи свої рішення. Ця платформа сприяє обміну знаннями, що допомагає спільноті швидко реагувати на нові виклики та досягати значних проривів разом.

Корпуси даних – критично важливі для розвитку NLP, оскільки являються основним джерелом даних для навчання та тестування моделей. Наявність великої кількості різних корпусів збільшує кількість завдань, які можуть бути розв’язані та покращує якість моделей.

2.3 Огляд GLUE benchmark

У контексті AI орієнтири (benchmarks) передбачають систематичну оцінку моделей за конкретними параметрами [39]. Це дозволяє зрозуміти сильні та слабкі сторони різних методів, дозволяючи розробникам приймати обґрунтовані рішення та досягнення. Орієнтири використовують

стандартизовані тести для здійснення порівнянь, що стимулює розвиток передових технологій [39]. Вони містять як поодинокі завдання (наприклад, класифікація тексту), так і їх сукупність (Рисунок 2.3).

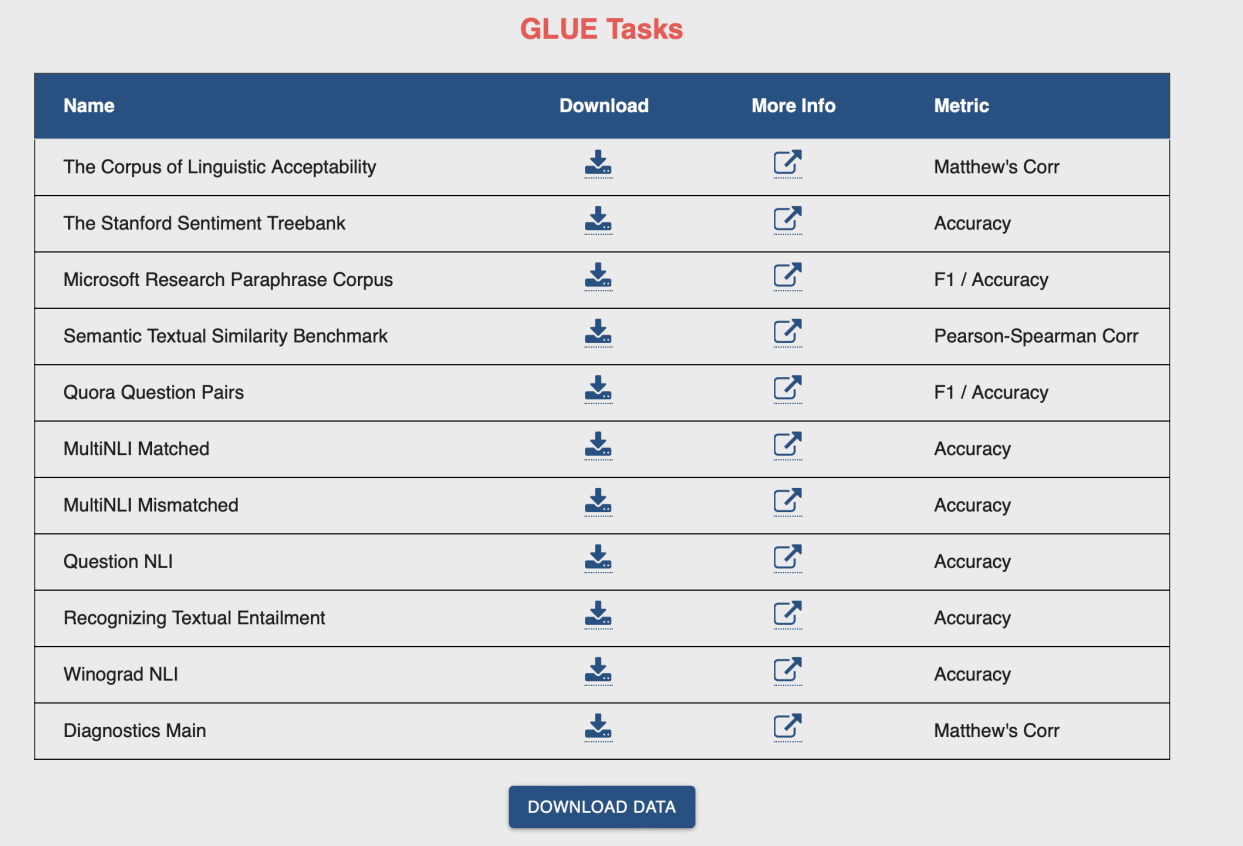

Рисунок 2.3 – Завдання GLUE [40]

GLUE (General Language Understanding Evaluation) – це колекція ресурсів для навчання, оцінки та аналізу систем розуміння природної мови [41]. Основна мета, полягає у заохочені до розвитку моделей, які можуть узагальнювати свої навички на різні завдання та області. Тобто, замість того, щоб оцінювати моделі лише за окремими завданнями, GLUE створює зручну платформу для оцінки їх здатності до розуміння мови загалом. Показуючи хороші результати на різних завданнях в рамках GLUE, з’являється потенціал

для розуміння мови в широкому спектрі контекстів, а не лише в обмеженому діапазоні.

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7

Рисунок 2.4 - Таблиця лідерів для GLUE [42]

На поточній таблиці лідерів (Рисунок 2.4) GLUE benchmark виокремлюються п'ять найкращий моделей у кожної з яких неймовірні результати. Це свідчить про інтенсивну конкуренцію і високий рівень досконалості досягнутих результатів. Розглянемо лідерів:

а) Microsoft Alexander v-team з моделлю Turing ULR v6 демонструє вражаючі результати для усіх завдань, зокрема виділяється у задачі аналізу емоційних оцінок (SST-2) із результатом 97.5%, а також у задачі розуміння природної мови (QNLI) з результатом 96.7%.

б) JDExplore d-team із моделлю Vega v1 показує схожі високі показники, особливо видається у завданні аналізу емоційних оцінок (SST-2) із 97.9% і ледь поступається в завданні відповідей на питання (QNLI) із результатом 96.7%.

в) Ще одна команда від Microsoft Alexander v-team з моделлю Turing NLR v5 також досягла високого загального рейтингу, з особливою у завданні з розуміння природної мови (QNLI) де вони мають 97.9%.

г) DIRL Team зі своєю моделлю DeBERTa + CLEVER показує хороші результати, з найвищим досягненням у завданні аналізу сентименту (SST-2) із результатом 97.6% та в завданні з розуміння природної мови (QNLI) із результатом 96.7%.

д) ERNIE Team - Baidu із моделлю ERNIE вирізняється найвищою оцінкою у завданні з аналізу мови (CoLA) на рівні 75.5% та показує конкурентоспроможні результати в інших завданнях, особливо у завданні аналізу емоційних оцінок (SST-2) з 97.8%.

Так розробники можуть здійснювати аналіз найкращих моделей та приймати участь у так званих змаганнях, що заохочує у розвитку ефективних систем. Результати на таблиці лідерів (leaderboard) свідчать про значний прогрес у сфері обробки природної мови (NLP), демонструючи, як сучасні моделі вже здатні ефективно розуміти людську мову. Цей орієнтир (benchmark) дозволяє визначити ключові напрямки для подальших досліджень, особливо в областях, де моделі ще не досягли рівня людського розуміння. Таким чином, орієнтири розвивають більш універсальні методи розуміння мови, що є критично важливим для створення інтелектуальних систем майбутнього.

2.4 Оцінка стану

Стан обробки природної мови (NLP) для англійської мови наразі можна оцінити як надзвичайно передовий, з великими досягненнями та стрімким розвитком, особливо за останнє десятиліття. Революційні архітектури, такі як BERT (Bidirectional Encoder Representations from Transformers) та GPT (Generative Pre-trained Transformer), суттєво підвищили можливості машинного розуміння та генерації тексту, прокладаючи шлях до більш складних та орієнтованих на нюанси мови систем. Інструменти та платформи, так як SQuAD (Stanford Question Answering Dataset), AI2 (Allen Institute for AI), Kaggle, та GLUE Benchmark, забезпечують важливі ресурси для навчання, тестування та порівняння моделей, стимулюючи інновації та співпрацю в галузі. Їх наявність, неможливо переоцінити для розвитку сфери NLP, бо вони

дозволяють оцінювати свої моделі на великому спектрі завдань, сприяючи розвитку універсальних та адаптивних систем. Проте, існують проблеми, які необхідно вирішити, для подальшого розвитку систем обробки мови, а саме:

а) Упередженість (bias). Важливо уникати впровадження упередження у системах NLP, оскільки це слідує до необ'єктивних результатів. [23]

б) Обрахункова потужність. З огляду на те, що завдання у сфері NLP вимагають великої кількості обчислень, зокрема тренування глибоких нейронних мереж, обчислювальна потужність стає ключовим фактором для хорошої моделі. Однак вирішення цього виклику може бути складним через обмежену доступність ресурсів, складність програмування та оптимізації алгоритмів, а також високі витрати на обладнання та інфраструктуру

в) Трансформери ламають таблиці лідерів NLP [27]. Архітектура Transformer відома своєю високою продуктивністю та здатністю до оптимізації, що приводить до кращих результатів на багатьох завданнях порівняно з іншими моделями. Однак це створює нерівності у змаганнях, оскільки доступ до потужних обчислювальних ресурсів і великих обсягів даних може забезпечити перевагу виключно через розмір моделі, а не через суттєві наукові досягнення. Потенційним вирішенням даної проблеми може бути створення стандартних навчальних корпусів, які б дозволяли об'єктивно порівнювати різні моделі без переваги для тих, хто має доступ до більшого обсягу даних або обчислювальних ресурсів. Крім того, важливо враховувати кількість обчислень та обсяг даних, доступних кожному учаснику, під час оцінки результатів, щоб відображати реальні можливості моделей і уникнути несправедливих переваг.

Підсумовуючи, стан розвитку обробки природної мови для англійської мови вражає своїми досягненнями. Завдяки значним зусиллям розробників, було досягнуто великих успіхів у розвитку революційних технологій, зокрема

BERT та GPT, та створенні масштабних корпусів даних. Однак, незважаючи на це, навіть для найбільш прогресивної мови у сфері NLP, існують певні труднощі, які ще необхідно розв’язати, щоб підняти рівень ефективності та точності систем на ще вищий. Таким чином, незважаючи на стрімкий розвиток, NLP для англійської мови залишається сферою, яка постійно еволюціонує та знаходить нові шляхи для вдосконалення.

3. NLP ДЛЯ УКРАЇНСЬКОЇ МОВИ

3.1 UNLP спільнота

Розвиток технологій обробки різних природніх мов є ключовим аспектом для сучасної епохи, адже вони сприяють підтримці мовного різноманіття у цифровому світі та економічному зростанню. Розквіт інноваційних інструментів для обробки саме української мови припадає на останнє десятиліття, разом із появою корпусів даних і вдосконалених алгоритмів машинного навчання. Це допомогло розвинути ефективніші технології для виконання різних завдань українською мовою, серед яких відзначають семантичний аналіз, автоматичний переклад та генерація тексту. Також значний поштовх до розвитку дали Ukrainian Natural Language Processing Workshops (UNLP) [43], які стартували у 2021 році. Особлива увага на цих заходах приділяється викликам, що стоять перед спільнотою. Ці конференції (workshops) стали місцем для обговорення інновацій, що об'єднують дослідників, які прагнуть досягти однієї мети - підвищення якості NLP ресурсів для української мови. UNLP конференції сприяють обміну досвідом між учасниками, що в свою чергу допомагає сформувати міцну наукову спільноту, здатну вирішувати актуальні задачі і активно просувати галузь українського NLP. У 2023 році відбувся другий UNLP Workshop, де були представлені ключові технології, зокрема UberText2.0 [44] та UA GPT-2 [45].

UberText 2.0 - це оновлена версія корпусу текстів української мови, призначена для різноманітних задач у галузі обробки природної мови (NLP). Цей корпус включає значну кількість даних, а саме: приблизно 2.5 млрд токенів, 8.59 млн текстів, 156 млн речень та 32 гігабайтів тексту [44]. Він створений за допомогою доповнення текстів з попередньої версії UberText 1.0, а також з нових джерел. Тексти збиралися за допомогою спеціального

фреймворку Scrapy, дозволяючи автоматизувати процес вилучення текстів з веб-сторінок, а використання спеціалізованих скрейперів для кожного джерела, дало змогу збирати лише релевантні текстові дані. UberText 2.0 містить підкорпуси з новин, художньої літератури, соціальних медіа, Вікіпедії та судових рішень, кожен із яких має власну колекцію в базі даних MongoDB, яка дозволяє ефективно управляти великим обсягом текстів. Тексти обробляються для сегментації на речення, токенізації, лематизації та POS-тегування, що робить дані корисними для широкого спектру завдань NLP. [44]

Завдяки тому, що UberText 2.0 містить багато якісних даних і знаходиться у відкритому доступі, він є значним ресурсом для досліджень у обробці української мови. Корпус дає можливість розробникам створювати сучасні моделі мови та тренувати їх на великих кількостях даних, сприяючи підвищенню якості різноманітних NLP-завдань. Незважаючи на великий потенціал, підтримка такого масштабного корпусу вимагає значних зусиль у сфері інженерії даних. Проблеми можуть виникати з актуальністю даних, особливо коли джерела даних швидко змінюються або містять багато шуму.

Модель GPT-2 UA - мовний декодувальник (decoder) на основі GPT-2, адаптований спеціально для української мови [45]. Вона створена шляхом попереднього тренування (pre-trained) на корпусі текстів UberText 2.0, використовуючи новітні методи SentencePiece [46] для токенізації, архітектуру Transformer (з оновленнями PyTorch 2.0) та інноваційні методи управління пам'яттю. Основні характеристики виділяють: тренування кількох версій моделі (Small, Medium, Large) з різною кількістю параметрів (124 млн, 355 млн, 774 млн) і шарами, використання оптимізацій для зменшення часу тренування та підвищення ефективності.

Model	Accuracy
Flair LSTM Forward/Backward	**0.979**
UDPipe	0.975
GPT-2 Medium Instr. Parallel **(ours)**	0.964
FastText CBOW (flair)	0.940
FastText CBOW (spacy)	0.825

Рисунок 3.1 - Продуктивність POS [45]

На Рисунку 3.1, модель GPT-2 Medium, показує хорошу продуктивність (performance) у завданні POS-тегування з точністю 0.964. Такий високий показник свідчить про високу ефективність GPT-2 UA у обробці морфологічної структури української мови, незважаючи на те, що ця модель decoder-only (призначена для генерації, де кожний токен має доступ тільки до попередніх, але не наступних), яка є менш ефективною для завдань, потребуючих розуміння двонаправленого контексту (як у PoS tagging). Зважаючи на складність української мови, результати GPT-2 UA є доволі вражаючими, що підкреслює потенціал таких моделей у покращенні обробки природної мови.

Model	F1	Prec	Recall
xlm-roberta-large	**0.92**	**0.92**	**0.91**
xlm-roberta-base	0.89	0.89	0.88
dbmdz/electra-base-ukrainian-cased-discriminator	0.89	0.89	0.89
lang-uk/electra-base-ukrainian-cased-discriminator	0.87	0.87	0.87
youscan/ukr-roberta-base	0.87	0.87	0.86
bert-base-multilingual-cased	0.87	0.88	0.87
Flair LSTM Forward and Back-ward	0.86	0.86	0.86
GPT-2 Large Instruction Data, Constrained Decoding **(ours)**	0.85	0.86	0.84
FastText CBOW	0.83	0.86	0.80
FastText skipgram	0.82	0.83	0.81

Рисунок 3.2 - Продуктивність NER [45].

На Рисунку 3.2, представлена таблиця продуктивності моделей для завдання розпізнавання іменних сутностей (NER). Незважаючи на те, що модель GPT-2 Large не змогла перевершити масштабні моделі, такі як xlm-roberta-large, вона все ж показала хороші результати. Враховуючи високу оцінку F1 та баланс між точністю і повнотою, можна стверджувати, що GPT-2 Large ефективно адаптована до української мови і буде корисною до застосування. Такі моделі особливо важливі для обмежених ресурсами мов (low-resource language), зокрема української, де доступ до великих, якісно-анотованих наборів даних обмежений.

UNLP Workshops дають можливість розробникам ділитись, покращувати та розвивати рішення задач NLP, що забезпечує наявність ефективних моделей для обробки української мови. Це сприяє розвитку

інструментів перекладу, автоматизації відповідей та багато інших аспектів, які залежать від глибокого розуміння мови.

3.2 Найкращі досягнення в українському NLP

Протягом останніх років українська спільнота розробників досягла значних успіхів у розвитку інструментів для обробки українських текстів. Окрім UNLP Workshops, важливий внесок у цей напрямок роблять також і інші плафторми, серед яких виділяють Kaggle, відомий своїми змаганнями, і Hugging Face, що забезпечує доступ до широкого спектру моделей та корпусів даних для української мови.

	lang-uk NER (Micro F1)	WikiANN (Micro F1)	UD Ukrainian IU POS (Accuracy)
roberta-base-wechsel-ukrainian	90.81 (1.51)	92.98 (0.12)	98.57 (0.03)
roberta-large-wechsel-ukrainian	**91.24 (1.16)**	**93.22 (0.17)**	**98.74 (0.06)**
roberta-base-scratch-ukrainian*	89.57 (1.01)	92.05 (0.09)	98.31 (0.08)
roberta-large-scratch-ukrainian*	89.96 (0.89)	92.49 (0.15)	98.52 (0.04)
dbmdz/electra-base-ukrainian-cased-discriminator	90.43 (1.29)	92.99 (0.11)	98.59 (0.06)
xlm-roberta-base	90.86 (0.81)	92.27 (0.09)	98.45 (0.07)
xlm-roberta-large	90.16 (2.98)	92.92 (0.19)	98.71 (0.04)

Рисунок 3.3 – Оцінка якості roberta-large-wechsel-ukrainian. [47]

Модель “roberta-large-wechsel-ukrainian” [47] - це адаптована мовна модель RoBERTa, оптимізована для української мови за допомогою методу WECHSEL [48], що дозволяє ефективно переносити великі мовні моделі (LLM) навчені на певній мові, для використання на нових мовах. Її використовують для вирішення завдань NER, PoS tagging та інші. Модель

демонструє високі показники на різних наборах даних, а оцінка якості (Рисунок 3.3) показала високу ефективність у порівнянні з іншими моделями, розробленими спеціально для української мови, що підтверджує здатність roberta-large-wechsel-ukrainian ефективно обробляти українські тексти.

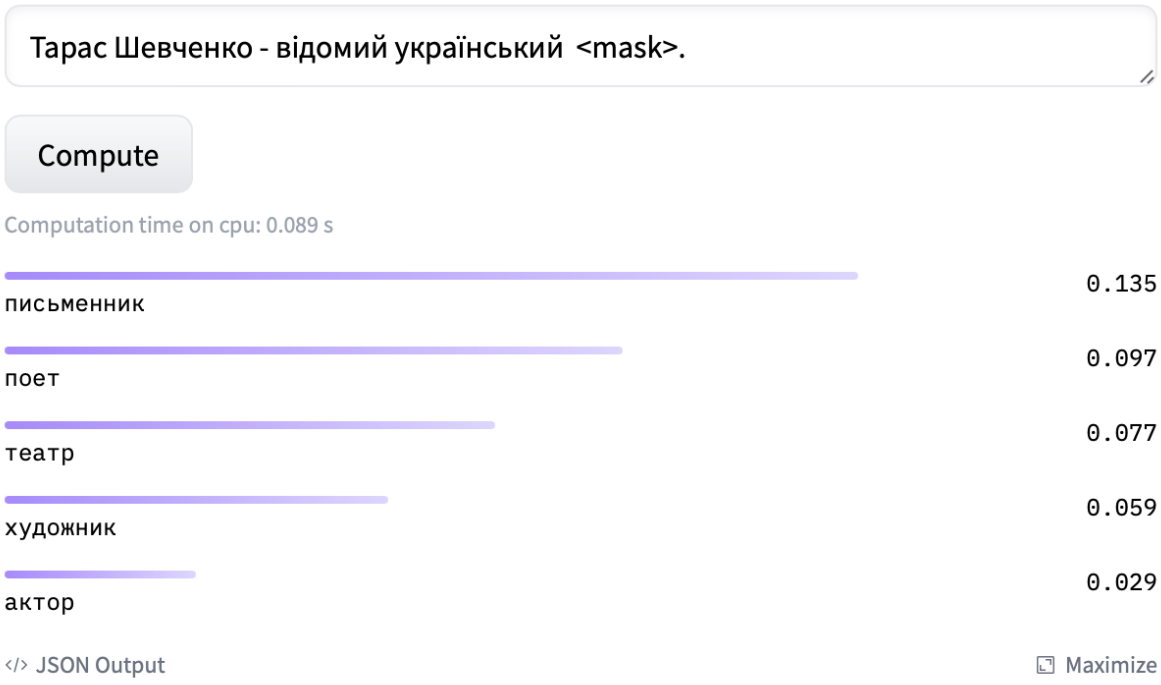

Рисунок 3.4 – Приклад роботи roberta-large-wechsel-ukrainian

На Рисунку 3.4 зображений результат даної моделі виконання завдання заповнення маски (fill-mask), яке є одним зі стандартних для оцінки розуміння контексту. Його суть полягає у передбаченні найімовірнішого слова, яке має бути вставлене замість маскувального токена (зазвичай <mask>). Згенеровані слова доводять глибоке розуміння контексту фрази, що свідчить про здатність моделі ефективно використовувати набуті знання української мови. Дані результати доречні та відображають хороший рівень виконання завдання, що робить roberta-large-wechsel-ukrainian цінним інструментом для використання в українських NLP-системах.

Модель youscan/ukr-roberta-base – це також адаптація для української мови мовної моделі RoBERTa, створена компанією YouScan [49]. Модель базується на архітектурі RoBERTa та спеціально навчена на великому наборі даних українською мовою.

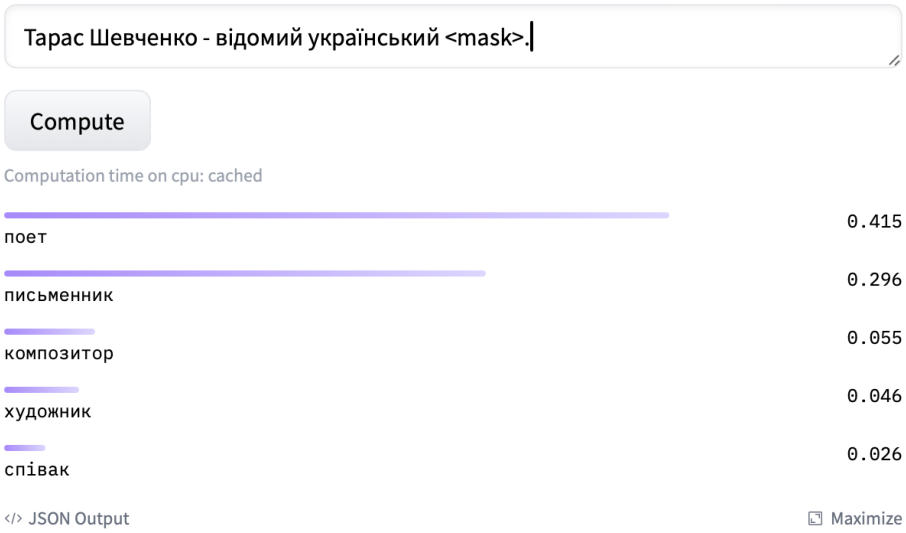

Рисунок 3.5 – Приклад роботи youscan/ukr-roberta-base.

Провівши у приклад таке ж завдання як і для roberta-large-wechsel-ukrainian (Рисунок 3.5), можна помітити, що влучність передбачених слів, у порівнянні із попередньою моделлю, в даному випадку, є кращою. Корпуси даних, що використовує дана модель, перелічують такі: Ukrainian Wikipedia (станом на травень 2020), Ukrainian OSCAR deduplicated dataset, Sampled mentions from social networks (Рисунок 3.6).

Tables	Lines	Words	Characters
Ukrainian Wikipedia - May 2020	18 001 466	201 207 739	2 647 891 947
Ukrainian OSCAR deduplicated dataset	56 560 011	2 250 210 650	29 705 050 592
Sampled mentions from social networks	11 245 710	128 461 796	1 632 567 763
Total	85 807 187	2 579 880 185	33 985 510 302

Рисунок 3.6 – Тренувальні корпуси даних для youscan/ukr-roberta-base. [49]

Як і попередня модель, youscan/ukr-roberta-base доступна для використання та інтеграції на платформі Hugging Face, проте відсутність результатів виконання різноманітних завдань з обробки природної мови (NLP) на сторінці - суттєвий недолік. Дана модель є ефективною (Рисунок 3.2), зокрема для завдання виявлення іменних сутностей (NER), що підкреслює її цінність для спільноти українських NLP.

UA SQuAD - це український переклад Stanford Question Answering Dataset (SQuAD) [50]. Він містить приблизно 13 тисяч запитань, сформованих на основі параграфів з Вікіпедії. Цей корпус даних надзвичайно важливий для української NLP-спільноти, оскільки забезпечує україномовний ресурс для тренування та оцінювання моделей відповідей на запитання. Зокрема, robinhad/ukrainian-qa – адаптована (fine-tuned) версія ukr-models/xlm-roberta-base-uk, натренована саме на UA SQuAD. Модель працює ефективно (Рисунок 3.7) та є хорошими інструментом для застосування.

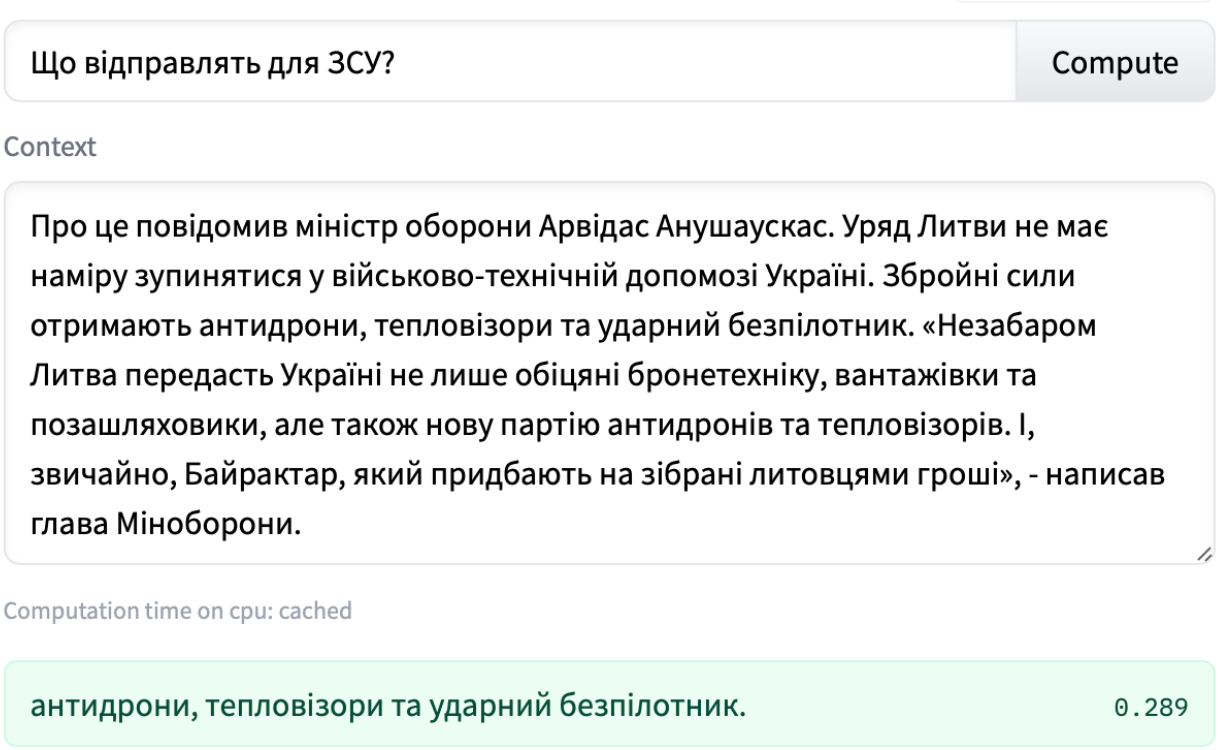

Рисунок 3.7 – Приклад використання robinhad/ukrainian-qa. [51]

Переклад даного корпусу був проведений достатньо якісно, до перекладу були залучені студенти з “Києво-Могилянської Академії” [52]. Попри всю важливість UA SQuAD, даний корпус все ж має певні недоліки. По перше, оскільки це перекладений корпус з англійських джерел, то дані місцями є занадто орієнтованими на англійську публіку. Крім того, під час аналізу корпусу було виявлено відсутність відповідей до деяких запитань (Рисунок 3.8), що ускладнює процес навчання, бо модель не розуміє як правильно реагувати на запитання. Ще одним недоліком є погано оформлена сторінка на платформі Hugging Face, яка містить лише корпус даних у розділі ресурсів, що може суттєво ускладнити доступність разом із використанням для розробників. Тому необхідно покращити оформлення сторінки, додавши більше інформації, яка б допомогла краще використовувати цей інструмент.

```
{
    "Question": "У якій відомій енциклопедії міститься естонська зворотна транслітерація російської мови?",
    "Context": "Слід зазначити, що естонські слова та імена, що цитуються в міжнародних публікаціях із радянських джерел, часто є зворотною транслітерацією з російської транслітерації. Прикладами є використання \"ya\" для \"ä\" (наприклад, Pyarnu замість Pärnu), \"y\" замість \"õ\" (наприклад, Pylva замість Põlva) та \"yu\" замість \"ü\" (наприклад, Pyussi замість Püssi). Навіть в Британській енциклопедії можна знайти \"ostrov Khiuma\", де \"ostrov\" означає \"острів\" російською мовою, та \"Khiuma\" є зворотною транслітерацією з російської замість \"Hiiumaa\" (Hiiumaa > Хийума(а) > Khiuma).",
    "Answer": ""
},
{
    "Question": "Що таке естонська зворотна транслітерація з Британської енциклопедії?",
    "Context": "Слід зазначити, що естонські слова та імена, що цитуються в міжнародних публікаціях із радянських джерел, часто є зворотною транслітерацією з російської транслітерації. Прикладами є використання \"ya\" для \"ä\" (наприклад, Pyarnu замість Pärnu), \"y\" замість \"õ\" (наприклад, Pylva замість Põlva) та \"yu\" замість \"ü\" (наприклад, Pyussi замість Püssi). Навіть в Британській енциклопедії можна знайти \"ostrov Khiuma\", де \"ostrov\" означає \"острів\" російською мовою, та \"Khiuma\" є зворотною транслітерацією з російської замість \"Hiiumaa\" (Hiiumaa > Хийума(а) > Khiuma).",
    "Answer": ""
},
{
    "Question": "Як перекладається ostrov естонською мовою?",
    "Context": "Слід зазначити, що естонські слова та імена, що цитуються в міжнародних публікаціях із радянських джерел, часто є зворотною транслітерацією з російської транслітерації. Прикладами є використання \"ya\" для \"ä\" (наприклад, Pyarnu замість Pärnu), \"y\" замість \"õ\" (наприклад, Pylva замість Põlva) та \"yu\" замість \"ü\" (наприклад, Pyussi замість Püssi). Навіть в Британській енциклопедії можна знайти \"ostrov Khiuma\", де \"ostrov\" означає \"острів\" російською мовою, та \"Khiuma\" є зворотною транслітерацією з російської замість \"Hiiumaa\" (Hiiumaa > Хийума(а) > Khiuma).",
    "Answer": ""
},
{
    "Question": "Як перекладається Khiuma естонською мовою?",
    "Context": "Слід зазначити, що естонські слова та імена, що цитуються в міжнародних публікаціях із радянських джерел, часто є зворотною транслітерацією з російської транслітерації. Прикладами є використання \"ya\" для \"ä\" (наприклад, Pyarnu замість Pärnu), \"y\" замість \"õ\" (наприклад, Pylva замість Põlva) та \"yu\" замість \"ü\" (наприклад, Pyussi замість Püssi). Навіть в Британській енциклопедії можна знайти \"ostrov Khiuma\", де \"ostrov\" означає \"острів\" російською мовою, та \"Khiuma\" є зворотною транслітерацією з російської замість \"Hiiumaa\" (Hiiumaa > Хийума(а) > Khiuma).",
    "Answer": ""
},
```


Рисунок 3.8. Відсутність відповідей у UA SQuAD.

CulturaX є великим багатомовним корпусом, створеним для підтримки розробки великих мовних моделей [53]. Він охоплює 6.3 трильйони токенів з 167 мов, включаючи українську, яка знаходиться на 20-му місці по загальній кількості токенів (понад 38 млрд) у даному корпусі, що є 0.61% від усіх. CulturaX складений із об'єднання та очищення даних з mC4 та OSCAR і є важливим ресурсом для тренування моделей, особливо для саме таких менш розповсюджених мов як українська. Цей набір даних підвищує можливості українського NLP, забезпечуючи вдосконалення мовних моделей.

Ukrainian StackExchange Dataset, зібраний з Ukrainian StackExchange, представляє собою цінний ресурс текстових даних для задач обробки української мови [54]. Дані включають інформацію таких типів, як запитання, відповіді, коментарі та метадані, зібрані станом на 02 квітня 2023 року.

Ще одним важливим ресурсом є Браунський корпус української мови, який створений за аналогічним корпусом для англійської мови [55]. Ідея полягає у створенні відкритого та збалансованого за жанрами (в майбутньому проанотованому) корпусу сучасної української мови (БрУК) обсягом 1 млн

слововживань. Важливість цього корпусу полягає у його здатності відображати сучасний лінгвістичний стан української мови, що є необхідним для створення якісних моделей. Завдяки постійному оновленню, Браунський корпус продовжує залишатися актуальним для розробників.

Також варті згадування інструменти, які слід використовувати для українських завдань NLP:

а) tree_stem - інструмент для стемінгу української мови [56].

б) pymorphy2 та pymorphy2-dicts-uk - морфологічний аналізатор і лематизатор для української мови [56].

в) nlp-uk - набір інструментів для нормалізації текстів, токенізації, лематизації, розпізнавання частин мови та виправлення двозначностей [56].

Протягом останніх років українська спільнота у галузі NLP досягла значних успіхів у розвитку інструментів для обробки мови. Моделі, такі як youscan/ukr-roberta-base та roberta-large-wechsel-ukrainian, показали високу ефективність у розумінні і генерації українських текстів, а UA SQuAD, CulturaX та БрУК розширили можливості для тренування моделей. Необхідну підтримку для лінгвістичних операцій надають інструменти tree_stem, pymorphy2, pymorphy2-dicts-uk та nlp-uk. Попри певні недоліки, усі ці досягнення свідчать про неспинний розвиток NLP для української мови, що надихає на подальші дослідження і неодмінно приведе до ще стрімкішого вдосконалення інструментів та методів.

3.3 Оцінка стану та проблеми українського NLP

Інноваційні інструменти UberText 2.0, GPT-2 UA, roberta-large-wechsel-ukrainian, youscan/ukr-roberta-base, UA SQuAD, CulturaX та БрУК ефективні та знаходяться у відкритому доступі, тому можна стверджувати, що стан для NLP для української мови знаходиться на хорошому рівні. Діяльність UNLP

Workshops, Hugging Face і Kaggle значно сприяє розвитку української спільноти у сфері NLP. Проте, українська мова залишається мовою з обмеженими ресурсами (low-resource language) і її розвиток ще далекий до рівня в англійському NLP.

Дослідивши стан NLP для української мови, було виявлено декілька ключових проблем:

а) *Складність української мови* – основний виклик для розробників через її величезну кількість суфіксів, способів і форм дієслів, довільний порядок слів у реченні, наявність семи відмінків і окремого роду для кожного слова [57]. Труднощі виникають також із наголосами, бо в українських словах він може падати на будь-який склад, змінюючи значення слова (наприклад, "зАмок" і "замОк") [57] . Така багата морфологія значно ускладнює процес аналізу українських текстів. Крім того, існує багато синонімів та омонімів, які призводять до двозначностей у розумінні тексту. Через суржик та русизми, дані робляться сильно зачумленими, що у свою чергу вимагає складного фільтрування. Звідси, моделі можуть неправильно інтерпретувати нюанси або навіть відносити текст до неправильної мови. Щоб вирішити ці проблеми необхідно продовжувати розробляти спеціалізовані методи очищення даних, за для ефективного виявлення і коректування впливу суржика або русизмів.

б) *Проблема обмеження в корпусах даних*, зокрема в українських аналогах таких як UA SQuAD, є значним викликом. Хоча UA SQuAD включає значну кількість даних, його зміст походять з англомовної версії, що не відображає різноманіття використання української мови. Це призводить до того, що моделі не ефективно справлятимуться з особливостями мови. Також, недостатня кількість україномовних даних у специфічних доменах обмежує можливість моделей генерувати релевантні відповіді на питання, що стримує застосування NLP. З вирішенням даної проблеми нам допомагають такі

провідні компанії як Grammarly, яка у 2021 році опублікувала перший корпус даних, спрямований для виправлення граматичних помилок (GEC) та корекції вільного володіння для української мови – UA-GEC [58]. Проте, обмеження в корпусах для специфічних доменів досі існує, що вимагає від спільноти створення більшої кількості ресурсів, які були б спеціально зібрані та анотовані з врахуванням українських лінгвістичних та культурних особливостей.

в) *Недоступність великих обчислювальних кластерів* є серйозною проблемою для розвитку NLP для української мови. Україна має значно меншу кількість обчислювальних кластерів порівняно з США, де існують масштабні обчислювальні центри з величезними потужностями. Також у Європі діють аналоги, як PLGrid [59] у Польщі, що об'єднють декілька суперкомп'ютерів (Helios, Athena, Ares, Prometheus, Zeus) і пропонують великі обчислювальні ресурси для наукових досліджень. Даний обчислювальний кластер безкоштовний для польських наукових проектів, що надає розробникам потужний інструмент. Аналог для України – це NG-Cloud від De Novo [60], який забезпечує необхідну інфраструктуру та ресурси для тренування NLP моделей і інших AI/ML завдань, особливо тих, що потребують великих обчислювальних потужностей. Проте, їхні послуги платні, що обмежує доступ для тренування і тестування моделей для місцевих дослідників. Вирішення проблеми обмеженості в доступності великих обчислювальних кластерів, потребує інвестицій у місцеву інфраструктуру та підтримку міжнародного співробітництва для забезпечення необхідних ресурсів.

г) Українська мова використовує кирилицю як основний алфавіт, проте *у текстах часто можна зустріти елементи латиниці* (назви компаній або URL), що ускладнює розробку, оскільки необхідно ефективно обслуговувати тексти з обидвох алфавітів. Присутність латинських слів у текстах, написаних

кирилицею, може впливати на точність токенізації, морфологічного аналізу та інших NLP-завдань. Токенізатори, оптимізовані виключно для кирилиці, можуть неправильно ідентифікувати межі слова або ігнорувати латинські вставки як нерелевантні, що призводить до помилок у подальшому аналізі. Відповідно, NLP-системи потребують додаткових алгоритмів, які б могли адекватно обробляти змішані текстові дані, а це в свою чергу є великим ускладненням.

д) Існує проблема *низької репрезентації української мови* у багатомовних корпусах для тренування генеративних моделей і паралельних корпусах для машинного перекладу. Великі багатомовні корпуси сформовані з переважаючою більшістю даних з англійської та інших широко вживаних мов, залишаючи українську мову позаду. Така репрезентація призводить до того, що натреновані на них моделі, не можуть досягти такої ж продуктивності у перекладі з англійської на українську, як для інших мов, оскільки, вони не мають достатньої кількості українських текстових даних для ефективного навчання та генерації якісних перекладів. Також через це знижується якість корпусів створених машинним перекладом з англійської на українську або за допомогою машинної генерації тексту, оскільки вони будуть менш точними та містити помилки. Для вирішення цієї проблеми, необхідно збільшувати частку українськомовних даних у масштабні мовні ресурси.

е) *Недостатність спеціалізованих орієнтирів (benchmark)* - ще одна критична проблема. Великі англомовні орієнтири GLUE та SuperGLUE забезпечують розуміння якості моделі, у той час як для української мови подібні комплексні орієнтири відсутні, що ускладнює оцінку разом із порівнянням моделей. Проте, на UNL Workshop 3 (відбудеться у травні 2024 року) буде презентовано роботу Eval-UA-tion 1.0: Benchmark for Evaluating

Ukrainian (Large) Language Models [61], що потенційно може розв’язати дану проблему.

Загальний стан NLP для української мови можна описати як прогресивний, оскільки у відкритому доступі є великі мовні моделей і якісні великі корпуси даних. Проте, ця сфера також зіштовхується з рядом проблем, вирішення яких, вимагає зусиль з боку наукової спільноти та інвестицій в інфраструктуру та ресурси, для того, щоб обробка української мови могла досягти свого потенціалу на рівні з провідними світовими мовами.

Висновки

Під час виконання даної курсової роботи, було охарактеризовано NLP разом із основними завданнями, проаналізовано дану сферу для найбільш розвиненої мови – англійської, визначено стан для обробки української мови та успішно виявлено ключові проблеми, які стримують розвиток у цій сфері, серед яких можна визначити:

а) Складність української мови.

б) Обмеження в корпусах даних.

в) Недоступність великих обчислювальних кластерів.

г) Елементи латиниці в текстах.

д) Низька репрезентація української мови у багатомовних корпусах.

е) Недостатність спеціалізованих орієнтирів.

Потенційні рішення проблем визначені як розвиток спеціалізованих методів для очищення даних, підвищення доступності та різноманітності українськомовних корпусів через співпрацю спільноти, інвестиції в місцеву інфраструктуру для забезпечення обчислювальних ресурсів та залучення міжнародної підтримки.

Для виявлення нових проблем, рекомендується переглянути третю конференцію UNLP, яка запланована на 25 травня 2024 року. На ній очікуються представлення нових передових моделей, корпусів даних та орієнтири (benchmarks), які у свою чергу можуть здійснити значний поштовх у сторону вирішення існуючих проблем і визначення нових напрямків розвитку для обробки української мови, що допоможе описати нові, найактуальніші виклики для розробників.

Список використаних джерел

1. IBM. What is NLP? [Електронний ресурс] / IBM – Режим доступу до ресурсу: https://www.ibm.com/topics/natural-language-processing.

2. TechTarget. The evolution of natural language processing [Електронний ресурс] / TechTarget – Режим доступу до ресурсу: https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP.

3. Ganesan K. Text Preprocessing for Machine Learning & NLP [Електронний ресурс] / Kavita Ganesan – Режим доступу до ресурсу: https://kavita-ganesan.com/text-preprocessing-tutorial/#.Xi2BhhczZTY.

4. Ganesan K. nlp-in-practice [Електронний ресурс] / Kavita Ganesan – Режим доступу до ресурсу: https://github.com/kavgan/nlp-in-practice/blob/master/text-pre-processing/Text%20Preprocessing%20Examples.ipynb.

5. POS(Parts-Of-Speech) Tagging in NLP [Електронний ресурс] – Режим доступу до ресурсу: https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/.

6. Goyal C. Step by Step Guide to Master NLP – Syntactic Analysis [Електронний ресурс] / Chirag Goyal. – 2021. – Режим доступу до ресурсу: https://www.analyticsvidhya.com/blog/2021/06/part-11-step-by-step-guide-to-master-nlp-syntactic-analysis/.

7. Shift Reduce Parser in Compiler [Електронний ресурс] – Режим доступу до ресурсу: https://www.geeksforgeeks.org/shift-reduce-parser-compiler/.

8. Constituency Parsing and Dependency Parsing [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://www.geeksforgeeks.org/constituency-parsing-and-dependency-parsing/.

9. Semantic Analysis: What Is It, How & Where To Works [Електронний ресурс] – Режим доступу до ресурсу: https://www.questionpro.com/blog/semantic-

analysis/#:~:text=Semantic%20analysis%20is%20a%20crucial,sentences%20in%20a%20given%20context.

10. Karatas G. Named Entity Recognition (NER): What It Is & How It Is Used in \'24 [Електронний ресурс] / Gulbahar Karatas – Режим доступу до ресурсу: https://research.aimultiple.com/named-entity-recognition/.

11. Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance. [Електронний ресурс] / [S. Naseer, M. G. Mudasar, A. Kiran та ін.] – Режим доступу до ресурсу: https://pjmr.org/pjmr/article/view/150.

12. Rosidi N. Machine Learning: What is Bootstrapping? [Електронний ресурс] / Nate Rosidi. – 2023. – Режим доступу до ресурсу: https://www.kdnuggets.com/2023/03/bootstrapping.html#:~:text=Bootstrapping%20is%20a%20resampling%20technique,same%20size%20as%20the%20original.

13. Sintayehu H. Named entity recognition: a semi-supervised learning approach [Електронний ресурс] / H. Sintayehu, G. Lehal – Режим доступу до ресурсу: https://link.springer.com/article/10.1007/s41870-020-00470-4.

14. Raj N. Guide to Sentiment Analysis using Natural Language Processing [Електронний ресурс] / Nikhil Raj. – 2024. – Режим доступу до ресурсу: https://www.analyticsvidhya.com/blog/2021/06/nlp-sentiment-analysis/.

15. Thakkar H. Approaches for Sentiment Analysis on Twitter: A State-of-Art study [Електронний ресурс] / H. Thakkar, D. Patel – Режим доступу до ресурсу: https://arxiv.org/abs/1512.01043.

16. Pragmatic Analysis [Електронний ресурс] – Режим доступу до ресурсу: https://www.askhandle.com/glossary/pragmatic-analysis.

17. Nath M. Topic modeling algorithms [Електронний ресурс] / Madhurima Nath – Режим доступу до ресурсу: https://medium.com/@m.nath/topic-modeling-algorithms-b7f97cec6005.

18. Van Otten N. Top 5 Ways To Implement Question-Answering Systems In NLP & A List Of Python Libraries [Електронний ресурс] / Neri Van Otten – Режим доступу до ресурсу: https://spotintelligence.com/2023/01/20/question-answering-qa-system-nlp/.

19. Mungalpara J. Evaluation Methods in Natural Language Processing (NLP): Part-1 [Електронний ресурс] / Jaimin Mungalpara. – 2023. – Режим доступу до ресурсу: https://jaimin-ml2001.medium.com/evaluation-methods-in-natural-language-processing-nlp-part-1-ffd39c90c04f.

20. Santhosh S. Understanding BLEU and ROUGE score for NLP evaluation [Електронний ресурс] / Sthanikam Santhosh – Режим доступу до ресурсу: https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb#:~:text=In%20the%20field%20of%20NLP%20evaluation%2C%20BLEU%20and%20ROUGE%20scores,used%20for%20text%20summarization%20tasks.

21. How can you measure natural language processing success? [Електронний ресурс] – Режим доступу до ресурсу: https://www.linkedin.com/advice/1/how-can-you-measure-natural-language-processing.

22. Kumar A. The Future of NLP in 2023: Opportunities and Challenges [Електронний ресурс] / Akash Kumar – Режим доступу до ресурсу: https://singhakash8190.medium.com/the-future-of-nlp-in-2023-opportunities-and-challenges-23779df8eb7d.

23. Dasgupta S. The gaffes and biases of Google Gemini [Електронний ресурс] / Shougat Dasgupta. – 2024. – Режим доступу до ресурсу: https://www.codastory.com/newsletters/the-gaffes-and-biases-of-google-gemini/.

24. Yao M. 10 Leading Language Models For NLP In 2022 [Електронний ресурс] / Mariya Yao – Режим доступу до ресурсу: https://www.topbots.com/leading-nlp-language-models-2020/#language-models-2022-1.

25. Yao M. What Every NLP Engineer Needs to Know About Pre-Trained Language Models [Електронний ресурс] / Mariya Yao. – 2019. – Режим доступу до ресурсу: https://www.topbots.com/ai-nlp-research-pretrained-language-models/.

26. Capital One Tech. Transformer model in NLP: Your AI and ML questions, answered [Електронний ресурс] / Capital One Tech. – 2023. – Режим доступу до ресурсу: https://www.capitalone.com/tech/machine-learning/transformer-nlp/.

27. Rogers A. How the Transformers broke NLP leaderboards [Електронний ресурс] / Anna Rogers – Режим доступу до ресурсу: https://hackingsemantics.xyz/2019/leaderboards/.

28. Yao M. 10 Leading Language Models For NLP In 2022 [Електронний ресурс] / Mariya Yao – Режим доступу до ресурсу: https://www.topbots.com/leading-nlp-language-models-2020/#language-models-2022-1.

29. Lutkevich B. BERT language model [Електронний ресурс] / B. Lutkevich, C. Hashemi-Pour – Режим доступу до ресурсу: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model#:~:text=BERT%2C%20which%20stands%20for%20Bidirectional,calculated%20based%20upon%20their%20connection.

30. Schulze J. What Is GPT? GPT-3, GPT-4, and More Explained [Електронний ресурс] / Jessica Schulze – Режим доступу до ресурсу: https://www.coursera.org/articles/what-is-gpt.

31. Language Models are Unsupervised Multitask Learners [Електронний ресурс] / [A. Radford, J. Wu, R. Child та ін.] – Режим доступу до ресурсу: https://paperswithcode.com/method/gpt-2.

32. Cooper K. OpenAI GPT-3: Everything You Need to Know [Електронний ресурс] / Kindra Cooper – Режим доступу до ресурсу: https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/.

33. Schreiner M. GPT-4 architecture, datasets, costs and more leaked [Електронний ресурс] / Maximilian Schreiner – Режим доступу до ресурсу: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

34. SQuAD 2.0 [Електронний ресурс] – Режим доступу до ресурсу: https://rajpurkar.github.io/SQuAD-explorer/.

35. SQuAD (Stanford Question Answering Dataset) [Електронний ресурс] – Режим доступу до ресурсу: https://h2o.ai/wiki/squad/#:~:text=SQuAD%2C%20short%20for%20Stanford%20Question,text%20within%20the%20corresponding%20article.

36. Allen Institute for AI, Datasets [Електронний ресурс] – Режим доступу до ресурсу: https://allenai.org/data.

37. Kaggle [Електронний ресурс] – Режим доступу до ресурсу: https://www.kaggle.com.

38. Hugging Face [Електронний ресурс] – Режим доступу до ресурсу: https://huggingface.co.

39. Benchmarking [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://www.larksuite.com/en_us/topics/ai-glossary/benchmarking#.

40. Завдання GLUE [Електронний ресурс] – Режим доступу до ресурсу: https://gluebenchmark.com/tasks.

41. GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING [Електронний ресурс] / [A. Wang, A. Singh, J. Michael та ін.] – Режим доступу до ресурсу: https://openreview.net/pdf?id=rJ4km2R5t7.

42. Таблиця лідерів GLUE [Електронний ресурс] – Режим доступу до ресурсу: https://gluebenchmark.com/leaderboard.

43. UNLP [Електронний ресурс]. – 2021. – Режим доступу до ресурсу: https://unlp.org.ua.

44. Chaplynskyi D. Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale [Електронний ресурс] / Dmytro Chaplynskyi. – 2023. – Режим доступу до ресурсу: https://aclanthology.org/2023.unlp-1.1/.

45. Kyrylov V. GPT-2 Metadata Pretraining Towards Instruction Finetuning for Ukrainian [Електронний ресурс] / V. Kyrylov, D. Chaplynskyi – Режим доступу до ресурсу: https://aclanthology.org/2023.unlp-1.4.pdf.

46. Kudos T. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing [Електронний ресурс] / T. Kudos, J. Richardson. – 2018. – Режим доступу до ресурсу: https://aclanthology.org/D18-2012/.

47. Minixhofer B. benjamin/roberta-large-wechsel-ukrainian [Електронний ресурс] / Benjamin Minixhofer – Режим доступу до ресурсу: https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian?text=Тарас+Шевченко+-+відомий+український+%3Cmask%3E.

48. Minixhofer B. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models [Електронний ресурс] / B. Minixhofer, F. Paischer, N. Rekabsaz. – 2022. – Режим доступу до ресурсу: https://aclanthology.org/2022.naacl-main.293/.

49. Radchenko V. youscan/ukr-roberta-base [Електронний ресурс] / Vitalii Radchenko – Режим доступу до ресурсу: https://huggingface.co/youscan/ukr-roberta-base?text=Тарас+Шевченко+-+відомий+український+%3Cmask%3E.

50. FIdo-AI/ua-squad [Електронний ресурс] / [B. Ivanyuk-Skulskiy, A. Zaliznyi, O. Reshetar та ін.] – Режим доступу до ресурсу: https://huggingface.co/datasets/FIdo-AI/ua-squad.

51. Paniv Y. ukrainian-qa [Електронний ресурс] / Yurii Paniv – Режим доступу до ресурсу: https://huggingface.co/robinhad/ukrainian-qa?context=Про+це+повідомив+міністр+оборони+Арвідас+Анушаускас.+Уряд+Литви+не+має+наміру+зупинятися+у+військово-технічній+допомозі+Україні.+Збройні+сили+отримають+антидрони%2C+тепловізори+та+ударний+безпілотник.+«Незабаром+Литва+передасть+Україні+не+лише+обіцяні+бронетехніку%2C+вантажівки+та+позашляховики%2C+але+також+нову+партію+антидронів+та+тепловізорів.+І%2C+звичайно%2C+Байрактар%2C+який+придбають+на+зібрані+литовцями+гроші»%2C+-+написав+глава+Міноборони.&text=Що+відправлять+для+ЗСУ%3F.

52. ua-datasets [Електронний ресурс] / [B. Ivanyuk-Skulskiy, A. Zaliznyi, O. Reshetar та ін.] – Режим доступу до ресурсу: https://github.com/fido-ai/ua-datasets.

53. CulturaX [Електронний ресурс] – Режим доступу до ресурсу: https://huggingface.co/datasets/uonlp/CulturaX.

54. Ukrainian StackExchange Dataset [Електронний ресурс] – Режим доступу до ресурсу: https://huggingface.co/datasets/zeusfsx/ukrainian-stackexchange.

55. Корпусна група БрУК [Електронний ресурс] / [В. Старко, А. Рисін, О. Гавура та ін.] – Режим доступу до ресурсу: https://r2u.org.ua/corpus.

56. awesome-ukrainian-nlp [Електронний ресурс] – Режим доступу до ресурсу: https://github.com/osyvokon/awesome-ukrainian-nlp.

57. Пулатова К. Названо 8 найскладніших мов для вивчення: українська поміж них [Електронний ресурс] / Катерина Пулатова. – 2023. – Режим доступу до ресурсу: https://www.unian.ua/curiosities/samye-slozhnye-yazyki-mira-dlya-

izucheniya-top-8-12404412.html#:~:text=Українська%20мова%20вважається%20досить%20складною,що%20вже%20є%20великою%20складністю.

58. Syvokon O. Announcing UA-GEC: A Grammatical Error Correction Dataset for the Ukrainian Language [Електронний ресурс] / Oleksiy Syvokon. – 2021. – Режим доступу до ресурсу: https://www.grammarly.com/blog/engineering/announcing-ua-gec/.

59. PLGrid [Електронний ресурс] – Режим доступу до ресурсу: https://www.plgrid.pl.

60. De Novo [Електронний ресурс] – Режим доступу до ресурсу: https://denovo.ua/en.

61. UNLP WORKSHOP PROGRAM AND ACCEPTED PAPERS [Електронний ресурс] – Режим доступу до ресурсу: https://unlp.org.ua/program/.