МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Харківський національний університет радіоелектроніки

Факультет Комп’ютерних наук

Кафедра Програмної інженерії

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів автоматичної відповіді на запитання

Виконав:

Студент 2 курсу, групи ІПЗм-19-2

Дашенков Д. С.

Спеціальність 121 - Інженерія програмного забезпечення

Тип програми освітньо-наукова

Керівник: доц. каф. ПІ Турута О. П.

Допускається до захисту

Зав. кафедри _____ З.В. Дудар

2021 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
Кафедра Програмної інженерії
Рівень вищої освіти - другий (магістерський)

Спеціальність 121-Інженерія програмного забезпечення
Тип програми освітньо-наукова програма
Освітня програма Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)
« 26 » березня 2021 р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студента Дашенкова Дмитра Сергійовича

1. Тема роботи Дослідження методів автоматичної відповіді на запитання
затверджена наказом університету від « 26 » березня 2021 р. № 385 Ст.
2. Термін подання студентом роботи до екзаменаційної комісії
« 08 » травня 2021 р.
3. Вихідні дані до роботи *алгоритми обробки природної мови, алгоритми генерації мови на основі вводу мови, штучні нейронні мережі, пояснювальна записка. Використовувати середовище об'єктно-орієнтованого проектування, мову програмування Python.*
4. Перелік питань, що потрібно опрацювати в роботі *мета роботи, аналіз стану досліджень в галузі, постановка задачі, огляд задач обробки мови, огляд методів вирішення даних задач, огляд специфіки обробки української мови в порівнянні з іншими мовами, методи розуміння мови, методи генерації мови, методики налаштування алгоритмів розуміння та генерації мови задля покращення результатів, швидкодія різних алгоритмів розуміння та генерації мови.*

5 Перелік графічного матеріалу: схематичні зображення архітектур нейронних мереж BERT та BERT для задачі QA, приклад тексту після машинного перекладу, графіки зменшення похибки при тренуванні двох моделей, графік залежності часу роботи моделі від довжини речення, знімок екрану, слайди: титул, вступ, актуальність задачі, обробка різних мов, заповнення маски, час роботи моделей, відповідання на питання — дані, відповідання на питання — тренування моделей, відповідання на питання — точність, подальші дослідження, демо-система, висновки.

6 Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім’я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	доц. каф. ПІ, к.т.н., доц. Турута О. П.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі	20 грудня 2020р.	виконано
2.	Огляд існуючих методів	30 січня 2021р.	виконано
3.	Спецчастина, експеримент	10 березня 2021р.	виконано
4.	Підготовка пояснювальної записки	15 березня 2021р.	виконано
5.	Підготовка презентації та доповіді	20 березня 2021р.	виконано
6.	Нормоконтроль	2 травня 2021р.	виконано
7.	Рецензування	4 травня 2021р.	виконано
8.	Попередній захист	5 травня 2021р.	виконано
9.	Архівування	8 травня 2021р.	виконано
10.	Допуск до захисту у зав. кафедри	8 травня 2021р.	виконано

Дата видачі завдання « 25 » січня 2021 р.
Студент _____

Керівник роботи _____ доц. каф. ПІ, к.т.н., доц. Турута О. П.

РЕФЕРАТ / ABSTRACT

Кваліфікаційна робота містить 70 с., 7 рис., 5 табл., 48 джер.

ШТУЧНІ НЕЙРОННІ МЕРЕЖІ, ОБРОБКА МОВИ, РОЗУМІННЯ МОВИ, ГЕНЕРАЦІЯ МОВИ, ВІДПОВІДАННЯ НА ЗАПИТАННЯ.

Метою роботи є аналіз, порівняння та покращення існуючих методів автоматичного відповідання на запитання в багатомовному контексті. Робота акцентує увагу на обробці української мови.

Методи, описані в роботі, базуються на актуальних дослідженнях в сфері обробки природної мови. В роботі порівнюються результати роботи існуючих нейромережевих моделей для англійської мови та для української мови; демонструються слабкі сторони деяких сучасних моделей в обробці недостатньо представлених мов на прикладі української мови. На базі знайдених недоліків, пропонуються методи покращення результатів обробки мови, і конкретно автоматичного відповідання на запитання.

ARTIFICIAL NEURAL NETWORKS, NATURAL LANGUAGE PROCESSING, LANGUAGE COMPREHENSION, LANGUAGE GENERATION, QUESTION ANSWERING.

The goal of this work is analysis, comparison and enhancement of the existing methods of automatic question answering in multilingual context. The work draws particular attention to processing Ukrainian language.

Methods, described in the work, are based on the up-to-date research in natural language processing. The work compares the outputs of existing neural network models on English and Ukrainian; weaknesses of some modern models in processing underrepresented languages are demonstrated on the example of Ukrainian language. Based on the found shortcomings, the work suggests methods for enhancing language processing, in particular for the question answering task.

Я, Дашенков Дмитро Сергійович, студент групи ІПЗм-19-2, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження методів автоматичної відповіді на запитання», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElAr KhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

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

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

BERT — (англ. Bidirectional Encoder Representations from Transformers) — двонаправлені представлення кодувальників трансформерів — нейронна мережа загального призначення для роботи із мовою.

mBERT — (англ. Multilingual BERT), багатомовна версія BERT.

NLG — (англ. Natural Language Generation), генерація природної мови, один із архетипів завдань NLP.

NLP — (англ. Natural Language Processing), автоматизована обробка природної мови.

QA — (англ. Question Answering), продукування відповідей на запитання, сформовані природною (не формалізованою) мовою.

RoBERTa — (англ. Robustly Optimized BERT), сильно оптимізована версія BERT.

SQuAD — (англ. Stanford Question Answering Dataset) — набір даних, націлений на задачу продукування відповідей на запитання, розроблений в Стенфордському університеті.

ВСТУП

Мова є однією з визначальних характеристик людства. Комунікація між індивідами з доісторичних часів і до сьогодні є основним інструментом розвитку людської цивілізації.

Вважається, що усна мова виникла від 50 тис. до 150 тис. років тому, на ряду із виникненням анатомічно-сучасної людини [1]. Це свідчить про невідривну природу мови та людства.

У наші дні, мови умовно поділені за походженням на 142 мовні мовні сім'ї [2]. Сучасні мови представляють неймовірне розмаїття і подальше класифікуються за ознаками фонетики, писемності та методів писемності, тощо.

Незважаючи на важливість проблеми, автоматична обробка мови за допомогою алгоритмів (NLP) є відносно новою галуззю комп'ютерних наук. Причиною на це є складність задачі розуміння людської мови. Люди будують думки за допомогою послідовностей звуків чи символів. Проте, окремі елементи цієї послідовності несуть мало інформації. Контекст в якому знаходиться окремий звук, слово чи речення є таким же важливим як і самий звук, слово чи речення. При цьому прив'язка складних абстрактних понять до слів вимагає від того хто сприймає мову повного розуміння цих абстрактних понять, а отже значна частина контексту, необхідного для розуміння мови, існує неявно в свідомості того хто її продукує та того хто її сприймає. Іншими словами, для розуміння виразів людської мови довільної складності необхідне володіння та розуміння інформації про світ на рівні людини.

Ще одна складність NLP полягає в розмаїтті мов та в тому наскільки сильно мови відрізняються одна від одної. Носіям мови однієї сім'ї може бути важко навіть уявити всі великі та дрібні відмінності мов що належать до інших сімей. Так, окрім різних слів для позначення понять, мови можуть мати зовсім різні структури та правила, виконання яких може бути обов'язковим або довільним.

Для того аби боротися із складністю обробки мов, сучасні дослідники використовують методи засновані на статистиці. Одним з найважливіших архетипів завдань NLP є генерація мови.

Генерація мови є важливою технологією, спрямованою на встановлення безперешкодного спілкування інформаційних систем з людьми використовуючи природну мову людини. Велика кількість різних завдань в NLP є завданнями генерації мови, і здатність ефективно виконувати ці завдання означає, що системи оснащені знаннями про світ, які можуть вимагати мультимодальної обробки та міркувань (наприклад, текстові, візуальні та слухові входи, або сенсорні потоки даних), і вивчення потужних нових методів машинного навчання, оскільки практично всі найсучасніші моделі NLP базуються на даних. Більше того, людські мови можуть сильно відрізнятися за своєю поверхневою реалізацією та внутрішньою структуру, що свідчить про те, що багатомовність є центральною метою дослідження для безпосередньої генеранії мови.

Технології генерації мови та NLP в цілому мають великий потенціал для застосування як державними, так і приватними установами, що несе великі економічні та соціальні переваги для осіб, які має доступ до цих технологій.

Однією з задач з розуміння та генерації мови є задача з автомагимного відповідання на запитання (QA). Програмні системи розроблені для вирішення цієї задачі мають деякий корпус даних — систематизований набір текстів, по якому здійснюється пошук відповіді на поставлене користувачем запитання. Запитання ставиться у довільній формі і може не мати ключових слів, які б допомагали при пошуці.

Метою цього дослідження є покращення існуючих методів відповідання на запитання українською мовою та в багатомовному контексті. Об’єктом дослідження є автоматична обробка природної мови з метою розуміння та генерації мови, конкретно для QA та в цілому. Предметом дослідження є методи вирішення задачі QA.

В цій роботі проведено декілька типів випробувань існуючих моделей, що вирішують поставлену задачу. Моделі оцінюються за показниками точності

відповіді та за показником F-міри [3]. Також, моделі комбінуються у ансамблі та порівнюється ефективність роботи ансамблів із ефективністю роботи окремих моделей. Нарешті, нові моделі створюються шляхом до-навчання штучних нейромереж загального характеру, тобто застосовується трансферне навчання [4].

В результаті роботи удосконалено якість вирішення задачі QA для української мови. Практична цінність такої якості вирішення цього завдання полягає в потенціалі для покращення всіх комерційних, інформаційних, навчальних, тощо систем пов’язаних із пошуком інформації у великих текстах.

Ця робота є поглибленням досліджень, що проводяться співробітниками кафедри Програмної Інженерії в сфері NLP, таких як дослідження методів семантичного аналізу тексту [5] тощо.

1 ПОТОЧНИЙ СТАН РОЗВ'ЯЗАННЯ ПРОБЛЕМИ

1.1 Завдання обробки мови та методи їх вирішення

1.1.1 Еволюція методів обробки мови

NLP як галузь комп'ютерних наук бере початок у 50-х роках XX століття із символьними алгоритмами. Такі алгоритми беруть на вхід множину правил мови та за їхньою допомогою обробляє та генерує текст. Значущим є дослідження компанії IBM та університету Джорджтауну (США) [6]. В цьому дослідженні було продемонстровано за допомогою невеликої множини правил, машинний переклад декількох речень з російської на англійську мову.

Пізніше, у 90-х роках XX століття, набувають популярності методи обробки мови засновані виключно на статистиці. Одним з значущих методів, який використовується для деяких задач і зараз є марківські ланцюги [7]. Цей метод здатний розпізнавати послідовності слів у реченні, або генерувати речення.

Обмеження марківських ланцюгів, природа яких лежить за межами цієї роботи, призвела до потреби в більш складних алгоритмах NLP. Зараз цю потребу задовольняють алгоритми на основі штучних нейромереж [8].

1.1.2 Сучасні задачі обробки мови

Дослідники NLP працюють над багатьма задачами. Найпоширеніші з них:

— машинне позначення — генерація ключових слів на основі тексту.

— узагальнення тексту — створення короткого резюме довгого тексту з метою збереження смислу та ідеї оригіналу. Метою завдання узагальнення тексту є стиснення тексту, на відміну від, наприклад, стиснення відео [9], узагальнення тексту, отримання значення тексту та видалення зайвих слів.

— машинний переклад — переклад тексту однією мовою на іншу. Це одна з перших та найпоширеніших задач NLP. Машинний переклад, зокрема, слугує допоміжним інструментом для вироблення даних для недостатньо представлених мов [10].

— діалог, включаючи відповіді на запитання — генерація відповідей на запитання людини. У випадку діалогу, завдання передбачає збереження контексту між кількома взаємодіями запитання-відповідь або команда-відповідь.

— розпізнавання іменованої сутності — класифікація сутностей, на які посилається власний іменник, наприклад, ім’я людини, назва компанії, назва книги тощо.

— аналіз настрою — оцінка емоцій людини, яка є автором фрагмента тексту. Наприклад, використовується для виявлення позитивних та негативних відгуків про певний товар чи послугу; семантичний аналіз тексту [5].

Такі задачі, як машинне позначення, узагальнення тексту, відповіді на запитання та аналіз настроїв згруповані в категорію завдань на розуміння — тих, що наслідують людське розуміння тексту.

Ця публікація зосереджена на завданнях на розуміння читання та, зокрема, на відповіді на запитання.

1.2 Стан досліджень задачі QA

1.2.1 Роботи світових дослідників

Методи вирішення завдань із розуміння мови значно відрізняються за складністю. До таких завдань, як аналіз настроїв, можна підходити з такими ж примітивними алгоритмами, як наївний Байєс [11]. У той же час узагальнення тексту та відповіді на запитання вимагають набагато складніших підходів.

Домінуюча архітектура моделі машинного навчання, яка використовується для складних завдань на розуміння читання, називається трансформатором [12].

Архітектура трансформатора використовує підхід кодера-декодера, що означає, що нейронна мережа кодує вхідний текст у вектор високої розмірності, а потім виробляє вихідні дані на основі кодованого подання вхідних даних.

Помітними прикладами моделей, побудованих на архітектурі трансформатора, є BERT (Bidirectional Encoder Representations from Transformer — двонаправлені представлення кодувальників від трансформаторів) [13] та його похідні моделі, такі як RoBERTa (Robustry Optimized BERT Pretraining Approach — надійно оптимізований підхід до перед-тренування BERT) [14], mBERT (багатомовний BERT) [15] та АЛЬБЕРТ (A Lite BERT — полегшений BERT) [16]. BERT та його оптимізації призначені для додавання додаткових вихідних шарів, які дозволяють моделі вирішувати довільні завдання, такі як відповідь на запитання. Однак початковим завданням є заповнення масок — заміщення пропусків у тексті ймовірними пропущеними словами. Загальна схема моделі BERT наведена на рисунку 1.1.

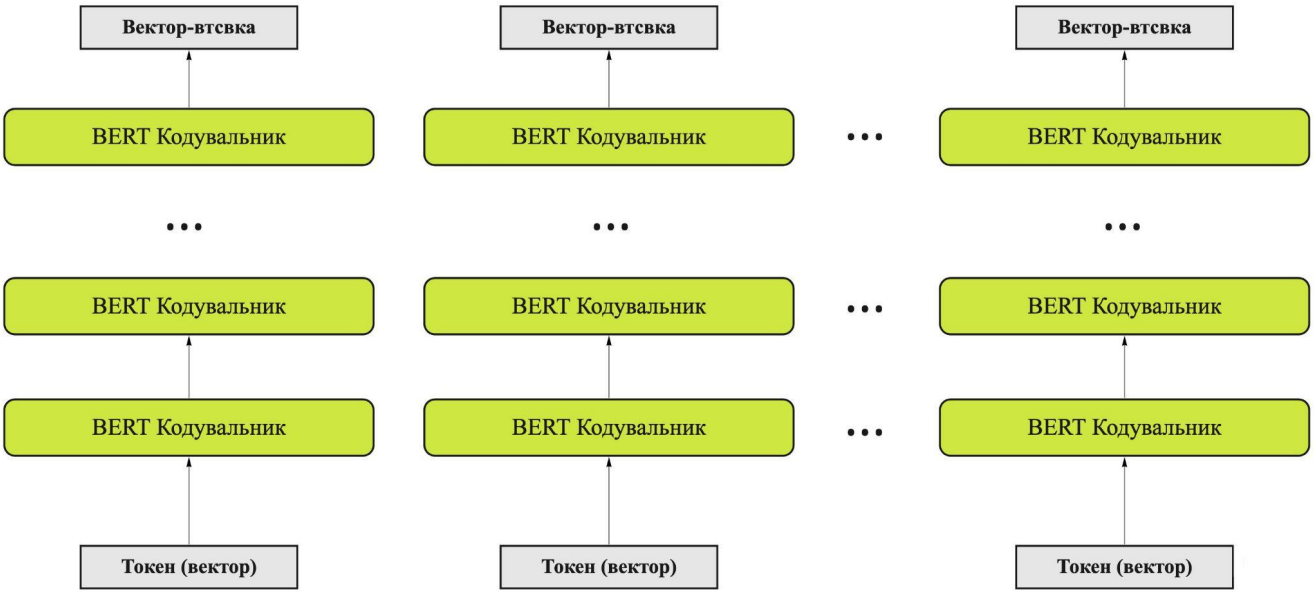

Рисунок 1.1 — Архітектура моделі BERT та подібних моделей

Іншими помітними прикладами трансформаторів є моделі сімейства GPT (генеративний попередньо підготовлений трансформатор), GPT-2 [17] та GPT-3 [18]. На відміну від BERT, GPT видає послідовність слів, одне за одним. Кожне

сформоване слово додається до вхідних даних, щоб слова, що йдуть за ним, були семантично та граматично пов’язані. Моделі GPT, особливо GPT-2 та GPT-3, призначені для роботи з великою кількістю параметрів, до 1,5 млрд для GPT-2, 17 млрд для GPT-3 [17, 18].

Ще один напрям досліджень простежується у використанні архітектури трансформерів для розуміння мовлення, тобто аналізу аудіо з метою виокремити людську мову [19]. Для отримання вхідних даних для такої обробки застосовується перетворення Фур’є — інтегральне перетворення, що також застосовується в комбінації із машинним навчанням для вирішення таких задач як обробка даних із спеціального обладнання, наприклад в геології [20], а також обробки медичних даних, зокрема даних риноманометрії [21][22].

1.2.2 Роботи українських дослідників та особливості обробки української мови

Незважаючи на те, що багатомовні мовні моделі, такі як mBERT, підтримують українську, деякі моделі можуть демонструвати менші показники роботи з українською мовою порівняно з англійською. Існують окремі моделі, навчені для української мови. Одним з них є RoBERTa, який навчався на українських наборах даних [23]. Обраною моделлю для цієї програми є базова версія RoBERTa. Модель складається з 125 мільйонів параметрів. Модель навчається за статтями в українській Вікіпедії (201 мільйон слів), дедуплікаційним набором даних OSCAR (2,25 мільярда слів) [24] та спеціальним набором даних, зібраним із публікацій у соціальних мережах (128 мільйонів слів).

Проте важче знайти спеціальні марковані набори даних для української мови. Один із таких наборів даних, а точніше колекція наборів даних, складає досліджувана спільнота lang-uk [25]. Ці набори даних призначені для вбудовування слів та завдань розпізнавання іменованих сутностей.

Серед останніх робіт є кілька розробок українського NLP загалом. Однією з них є робота з моделювання української мови [26], тобто формальний опис її структури для цілей обробки. Моделювання мови може відігравати важливу роль при вирішенні традиційних задач NLP, таких як відповіді на запитання та інші завдання на розуміння читання, а також виправлення помилок, аналіз настроїв тощо.

Інша заслуговує на увагу стаття представляє корпус українського тексту різних жанрів, коментується з урахуванням регіон або діаспора походження [27]. Корпус складається з понад 50 000 різних текстів сучасних та історичних авторів. Він націлений на лінгвістичні дослідження української мови.

Ще одна стаття представляє корпус української та інших мов, спрямованих на багатомовне навчання [28]. Набір даних складається з 64 текстів з літератури та 129 текстів з медицини. Набір даних складається з мільйона слів українською мовою та близько мільйона слів іншими мовами - польською, французькою та англійською.

Ще одна робота описує побудову алгебраїчного механізму у вигляді нейромережі з ціллю аналізу смислу речень [29]. Автори цієї роботи пропонують метод декомпозиції речень на логічні одиниці задля квантифікації та створення математичної моделі, що може бути представлена у вигляді нейромережі. Справедливе також і зворотне перетворення — нейромережі на математичну модель, яка кодує смисл речення.

Нарешті, ще одна стаття досліджує можливості тонкої налаштування mBERT на машинно перекладених даних [9]. У статті пропонується модель, яка складається з двох частин. Перша частина - це попередньо навчений mBERT. Друга частина - це відрегульована на замовлення модель, навчена набору даних питань та відповідей до статей Вікіпедії. Отримана модель націлена на завдання відповіді на питання. Стаття також пропонує порівняльний аналіз ефективності різних варіантів mBERT.

1.2.3 Особливості дослідження автоматизованого одномовного та багатомовного відповідання на запитання

Багато задач NLP, стикаються зі схожими проблемами. Перш за все, моделі NLP здатні на багато речей, але ці речі не можуть відповідати людському розумінню. Модель навчена бачити шаблони в тексті, але вона не може по-справжньому зрозуміти значення тексту.

Ця проблема переходить у область філософії, і визначення того, що може бути зроблено для її вирішення, не є чітким.

Однак більш вирішуваними, практичними проблемами є:

— багатомовна обробка без форми перекладу або з перекладом, який не погіршує початкового значення тексту. Будь-яка форма перекладу, людська чи машинна, часто має певні особливості основного наміру тексту. Тим більше що стосується машинного перекладу, який, як уже згадувалося, не може по-справжньому зрозуміти вхідний текст.

— розуміння в декількох документах. Людина може прочитати кілька документів (фрагментів тексту) за день. Для людини думки, викладені в цих документах, можуть переплітатися і формувати цілісну картину домену. Для алгоритму поєднання документів в єдиному «світогляді» є нетривіальною вправою.

— збір даних. Для багатьох мов, таких як англійська, збір даних не є складним. Існує безліч англійських наборів даних для NLP, таких як SQuAD [30], WikiQA [31], SearchQA [32]. Однак, коли справа стосується інших мов, особливо тих, що мають менше носіїв мови, проблема ускладнюється. Деякі набори даних збирають дані для кількох мов. Багато моделей, наприклад, mBERT, використовують набори даних на основі статей Вікіпедії. Такі набори даних часто складають певну кількість Вікіпедій із найбільшою кількістю статей [33].

2 НАУКОВО-ТЕХНІЧНА ЗАДАЧА ТА ЦІЛІ ДОСЛІДЖЕННЯ

Завданням цієї роботи є тренування моделей нейронних мереж, спеціально орієнтованих на задачу відповіді на питання різними мовами. Мета полягає в тому, щоб надати конфігурацію моделі, яка перевершила б існуючі загальні моделі при багатомовній відповіді на запитання, особливо українською та англійською мовами.

Це завдання вимагає двоступеневого підходу:

— збір набору даних відповідей на запитання цільовими мовами.

— побудова ансамблю моделей, їх навчання та налаштування.

Англійська, одна з наших цільових мов, є достатньо поширеною мовою для дослідів в NLP. Існує безліч наборів даних, призначених для завдання відповіді на запитання англійською мовою. Деякі з них: SQuAD, WikiQA, SearchQA, DeepMind Q&A [34], MS Macro QA [35]. Ці набори даних мають схожу структуру. Кожен запис — це питання та відповідь або ряд прийнятних відповідей, іноді супроводжуваний контекстом — текстовим фрагментом, що містить відповідь на запитання. З іншого боку, українська мова набагато менш вивчена. Досліджень NLP для української мови значно меньше. На даний момент не існує великих публічних наборів даних для української мови, створених для завдання відповіді на запитання. Цей розрив є причиною, яка спонукає нас виконувати поставлену задачу.

В результаті дослідження, окрім отриманих артефактів у вигляді навченої моделі, ми сподіваємося побудувати базу знань, на яку інші дослідники зможуть опиратися у їхній роботі над NLP для української мови.

Метою дослідження є встановлення, за умов невеликої кількості даних, який рівень точності можна очікувати від сучасних моделей, що працюють із задачею QA. Умова що стосується невеликої кількості даних походить із потреби розв’язання задачі QA для мов із обмеженим числом наборів даних. Так, наприклад, для української мови, на момент проведення даної роботи, не існує

відкритих наборів даних із структурою, що дозволяла б навчати моделі з ціллю вирішення саме задачі QA, в той час як для англійської мови таких наборів багато. Найзначущіші із цих наборів наведені вище.

Українська мова є хорошим середовищем дослідження, адже автори цієї роботи володіють українською мовою. Проте, крім цього, українська мова досить типова з точки зору відсутності форматованих промаркованих даних. Домінування досліджень із англійською мовою, спричинене, зокрема, тим що саме ця мова є рідною для багатьох провідних дослідників, може створювати ілюзію того що деякі задачі NLG, зокрема і QA, вже вирішені. Насправді ж, це не так.

Однією з технічних задач цього дослідження є оптимізація задачі QA, тобто збільшення точності передбачень моделей, в умовах невеликої кількості даних.

3 ЕКСПЕРИМЕНТ З ОЦІНЮВАННЯ ІСНУЮЧИХ МОДЕЛЕЙ ГЕНЕРАЦІЇ МОВИ

Для того аби знайти найбільш обіцяючі існуючі моделі для генерації мови, було проведено ряд експериментів, пов’язаних із вирішенням іншої задачі NLP — заповнення маски. Задача заповнення маски полягає у тому, аби підібрати найбільш підходяще слово або словосполучення на місці пропущеного слова в реченні.

Для даного аналізу ми створили невеликий спеціалізований розмічений набір даних. Набір даних складається із законодавчих термінів українською мовою. Кожний елемент — це визначення одного терміна. Це включає і сам термін, і кілька слів, які його визначають. Текст із пропущеним словом називається контекстом, а пропущене слово також називають замаскованим словом.

В наборі даних замасковані слова що найбільш асоціюються із терміном, який визначається. Всі ці слова — іменники.

Ефективність моделей у завданні заповнення маски оцінюватиметься двома чинниками:

— кількість точних прогнозів, коли прогноз повністю співпадає з очікуваним словом.

— кількість прогнозів, коли прогноз та очікуване слово різняться, проте, з точки зору людини, смисл речення зберігається (при умові що немає граматичних помилок).

Другий чинник орієнтований на результати, які є правильними в принципі, але сформульовані інакше, ніж оригінальний текст. Наприклад, якщо модель знаходить синонім до слова, яке було пропущене.

Також, для порівняння, вимірюємо ефективність роботи людини над тим самим завданням. Результати роботи людини, за визначенням, мають абсолютне

значення за оцінкою людини (другий чинник), але не обов'язково за точною оцінкою співпадіння слова (перший чинник).

Всього, набір даних складається із близько 30 тверджень, взятих з Конституції України та інших законів. Усі твердження мають описану структуру.

На цьому наборі даних ми запустили три моделі, mBERT, DistilBERT [36] та українську RoBERTa. Всі три моделі побудовані на основі BERT, і мають однаковий інтерфейс. mBERT — це багатомовна версія BERT, а DistilBERT та RoBERTa — оптимізації BERT.

Приклади елементів набору даних представлені в таблиці 1. Символ зірочка (“*”) відображає позицію на замаскованому слові в реченні.

Таблиця 1 — Приклади записів набору даних для заповнення маски

№ з/п	Контекст	Пропущене слово
1	податковий кредит — *, на яку платник податку на додану вартість має право зменшити податкове зобов'язання звітного (податкового) періоду, визначена згідно з розділом V цього Кодексу	сума
2	дебітор — *, у якої внаслідок минулих подій утворилася заборгованість перед іншою особою у формі певної суми коштів, їх еквіваленту або інших активів	особа
3	податкова вимога — письмова * контролюючого органу до платника податків щодо погашення суми податкового боргу	вимога
4	соціальна справедливість — установлення податків та * відповідно до платоспроможності платників податків	зборів

На вхід моделям подається контекст із замаскованим словом у вигляді токенів. На вихід модель дає, в залежності від конфігурації, одну чи декілька відповідей із вірогідностями їх правильності. Ми використовуємо натреновані моделі доступні в репозиторії Hugging Face [37].

Результати показали, що, хоча mBERT добре впорався з пошуком відповідного граматичного іменника, RoBERTa покращила результат що показав mBERT вдвічі, із показником 27,5% для mBERT та 55,2% для RoBERTa. В той же час, DistilBERT, що показав найгірший результат, в деяких випадках навіть не впорався із класифікацією мови. Це видно з того що деякі передбачення цієї моделі були не українською, а іншими мовами, хоча контекст завдання завжди був лише українською мовою.

У таблиці 2 наведені заміряні результати роботи трьох моделей на представленому наборі даних. У колонці “Правильні за смислом передбачення” наведено відсоток правильних з точки зору людини відповідей, які не співпали початковим словом, але зберігають смисл речення контексту.

Таблиця 2 — Результати прогнозування замаскованого слова

Модель	Точні передбачення	Правильні за смислом передбачення
людина	75,9%	-
mBERT	27,5%	47,3%
RoBERTa	55,2%	68,1%
DistilBERT	13,8%	20,7%

Помітимо, що фактична точність прогнозування, показана відсотком правильних за смислом передбачень, сильно більша ніж формальна точність, показана кількістю передбачень що співпали із початковим словом.

Цей факт, із врахуванням високого, але, можливо, досяжного цільового значення 75,9% продемонстрованого оператором-людиною, може свідчити про те, що вдосконалюючі моделі можна отримати результат вищий ніж результат встановлений оператором-людиною.

Важливо зазначити, що модель mBERT здатна прогнозувати деякі слова, з якими не впоралася модель RoBERTa. Це дає нам надію зібрати ці моделі в ансамбль, щоб отримати ще кращі показники.

Побудування ансамблю із цих моделей зводиться до комбінування відповідей моделей. Зазначимо, що результати DistilBERT завжди не кращі ніж результати mBERT, тобто з прикладами з якими впорався DistilBERT впорався і mBERT також (хоча зворотнє твердження не справедливе). Відходячи від цього, було прийняте рішення не включати DistilBERT в ансамбль.

Для отримання зведеного результату роботи ансамблю необхідно:

— отримати відповіді кожної з моделей.

— відповіді включають вірогідності того щт відповідь правильна — впевненість моделі у відповіді. Помножимо впевненість кожної відповіді на відсоток правильних відповідей, який модель дала в попередньому експерименті, тобто на 0,68 для RoBERTa та на 0,46 для mBERT.

— якщо деякі відповіді моделей співпадають, складемо отримані значення. Отримаємо впевненість всього ансамблю у відповіді.

Побудувавши ансамбль таким чином, перевіримо його на тестовому наборі даних. Очікується, що результат показаний ансамблем буде кращий ща результат кожної окремої моделі.

Отриманий результат з ансамблю: 62% точних передбачень, 65,5% передбачень, що зберігають смисл речення.

Ще однією важливою відмінністю між моделями є етап токенізації. RoBERTa здатна об’єднувати сусідні токени, створюючи можливість “натякати” моделі на те, яким має бути закінчення слова. Ця здатність може бути корисною для NLG. Оскільки українська граматика має багато форм слова для різних контекстів, можливість допомагати моделі, надаючи правильне закінчення, може бути корисним. Також присутній зворотний ефект. Модель може знайти правильне закінчення слова базуючись його корені.

Для будь-якого практичного застосування нам також потрібно було б знати час за який працюють моделі. Оскільки ми відкинули DistilBERT як не достатньо точний, вимірюємо час роботи лише для RoBERTa та mBERT.

RoBERTa — це оптимізована версія BERT, як зазначено в назві [14]. mBERT, з іншого боку, є варіантом BERT [15]. Завдяки цьому ми припускаємо, що RoBERTa повинна мати можливість швидше обробляти введені дані.

Вимірювання проводяться на чіпі Intel Core i7 з тактовою частотою 2,6 ГГц з 16 ГБ пам'яті DDR4. Оцінки проводились послідовно. Моделі не мали доступу до графічних процесорів. У таблиці 3 зазначений час обробки моделями прикладів різного розміру.

Таблиця 3 — Час оцінки різних моделей на різних розмірах вхідних даних, в секундах

Модель	15 слів	25 слів	50 слів
mBERT	0,21	0,27	0,46
RoBERTa	0,11	0,14	0,23

Як видно з вимірювань, RoBERTa справді швидша за mBERT. На рисунку 3.1 показана залежність розміру речення від часу обробки для моделей RoBERTa та mBERT.

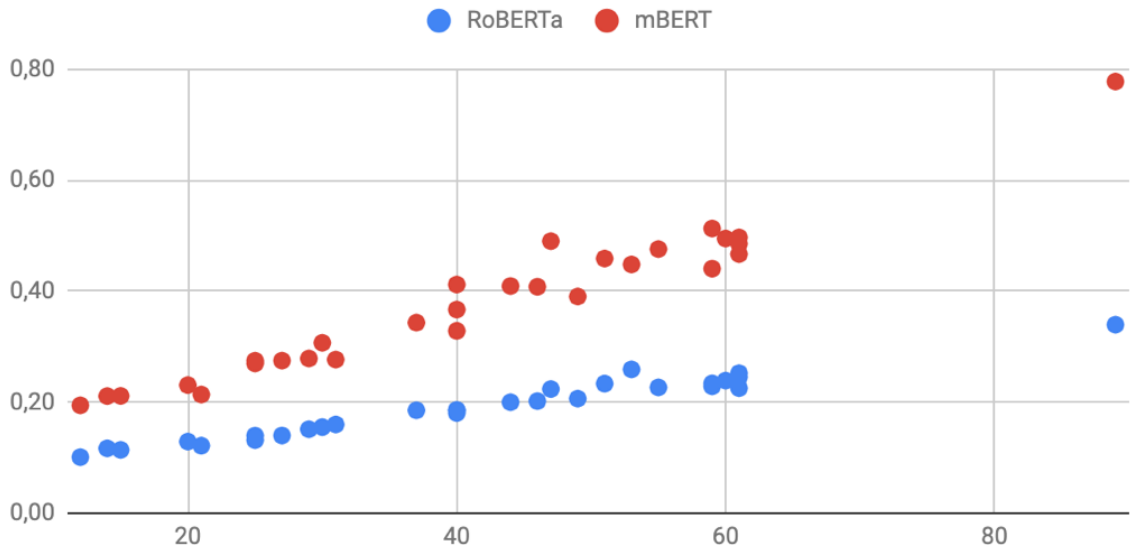

Рис. 3.1 — Залежність часу роботи RoBERTa та mBERT на завданні заповнення маски від кількості слів у реченні

З графіку ясно, що RoBERTa не тільки швидше обробляє введені дані, але й час обробки зростає повільніше, а отже, теоретично на більшому розмірі даних, час обробки не стане значно більшим. Також бачимо, що для обидвох моделей час обробки лінійно залежить від розміру речення лінійно.

Отже ми знайшли що RoBERTa точніша за mBERT у задачі заповнення маски та продукує відповіді швидше. Проте комбінація моделей у вигляді ансамбля все одно надає істотно кращий результат ніж одна модель окремо.

4 ЕКСПЕРИМЕНТ З ТРЕНУВАННЯ МОДЕЛЕЙ ГЕНЕРАЦІЇ МОВИ

4.1 Контрольованість в задачі QA

Контрольованість в QA полягає в механізмах генерації тексту схожого на текст створений людиною за змістом і за стилем. Зміст згенерованої відповіді — це інформація, отримана з необробленого тексту моделлю. Загалом, до завдань генерації мови відноситься широкий спектр моделей — від моделей sequence-to-sequence до трансформаторів.

В нещодавньому дослідженні було знайдено підхід до генерування запитань схожих на запитання людини, керуючи результатами роботи моделі за допомогою модальності іншої моделі [38]. В цьому дослідженні ми перевіряємо багатомовні характеристики ряду моделей із сімейства *BERT (mBERT, RoBERTa). Ми порівнюємо до-налаштовані моделі на різних наборах даних, таких як набір даних SQuAD та невеликий локальний набір даних українською мовою. В якості метрик ми використовуємо Natural Questions [39] та людську оцінку. Ми обмежені необробленим текстом в контекстах до запитання. Моделі сімейства *BERT використовують слова з необробленого тексту для пошуку відповідей. Для експерименту ми використовуємо наступний набір вхідних даних: необроблений текст як контекст українською та англійською мовами, запитання українською та англійською мовами. Отже, в результаті ми забезпечимо точність для чотирьох комбінацій входів.

4.2 Методи збору даних

До збору текстових даних для NLP можна підходити двома способами. По-перше, це ручна праця — група операторів, які виробляють дані, наприклад,

створюючи пари запитання-відповідь із заданого корпусу тексту. Такий підхід вимагає рівня фінансування та управління проектами, які зараз нам недоступні.

Другий підхід — видобуток наборів даних із уже існуючих джерел. Цей підхід вимагає набагато менше фінансування та є надзвичайно автоматизованим.

Перспективним підходом до збору даних є неконтрольоване навчання з використанням мультимодальних даних, таких як різнорідні медичні тексти [40]. Наприклад, можна отримати ознаки для сегментації зображень [41], використовувати методи машинного перекладу для перекладу міток. Ці мультимодальні вихідні дані надають багато можливостей для вдосконалення тренувальних даних для мов із малою кількістю даних, таких як українська.

Для нас джерелами для створення існуючих даних є українські правила дорожнього руху та питання іспитів з водіння, шкільні підручники з літератури, незалежні та державні тести з української та зарубіжної літератури тощо. Кожне з цих джерел окремо не надає достатньо даних, однак у сукупності вони можуть скласти досить великий набір даних, який може дозволити нам відредагувати моделі, навчені для більш простих завдань українською мовою, таких як заповнення маски, для завдання відповіді на запитання — підходу, відомого як транзитивне навчання.

Проте отримання даних таким чином потребує операторської людської роботи. Наприклад, для отримання даних з тестів необхідно копіювати, а інколи, передруковувати текст контексту, питань та відповіді. Також відкритим залишається питання авторського права деяких із цих джерел.

В цій роботі ми використовували дані отримані за допомогою машинного перекладу з англійської. Для цього, було перекладено більш ніж 4 тисячі прикладів набору даних SQuAD за допомогою сервіса IBM Language Translator [42].

Перевагами отримання даних за допомогою машинного перекладу є відсутність великої кількості людської операторської роботи, як, наприклад, при створенні власних наборів даних. безумовною перевагою є доступність даних та структурованість та універсальність даних.

Очевидним недоліком отримання даних за допомогою машинного перекладу є нижча якість даних. Машинний переклад є однією із типових задач NLP та потребує вдосконалення для всіх мов. Тому в наборі даних можуть зустрічатися приклади, в яких контекст або питання або і те й інше втрачають смисл. Від таких прикладів слід позбавлятися під час тренування та оцінювання моделей NLP.

Ще одна проблема машинного перекладу — точний але стилістично некоректний переклад. Використання різних слів в різних мовах може мати нюанси, які втрачаються при перекладі. Приклад речення наведено на рисунку 4.1.

```
Побачивши групу, продюсер R&B Arne Frager
привів їх до своєї студії у Північній
Каліфорнії і помістив їх у Star Search,
найбільшому талант-шоу на національному
телебаченні в той час.
```

Рисунок 4.1 — Приклад перекладеного тексту.

Смисл речення зберігається проте стилістика видає буквальний переклад. За можливості, від таких прикладів також слід позбавлятися. В нашому дослідженні, через брак інших даних, подібні приклади використовувалися при тренуванні моделей.

4.3 Метрики і тести

Стандартна метрика для завдань NLP відома як F-міра [3]. F-міра вимагає, щоб вихід моделі під час навчання був двійковим. В формулі 1 наведено спосіб розрахунку F-міри.

$$F_1 = \frac{tp}{tp + \frac{1}{2}(fp + fn)}, \quad (1)$$

де *tp* — кількість істинних позитивних відповідей,

fp — кількість помилково позитивних відповідей,

fn — кількість помилково негативних відповідей.

З цієї причини моделі, які виводяться під час навчання, можуть бути зведені до відповіді на запитання чи правильно надана відповідь на заявлене питання?. F-міра вимірює співвідношення між кількістю істинних позитивних відповідей і сумою кількості істинно позитивних відповідей і половиною кількості хибних відповідей.

Для задачі QA F-міру можна використовувати при наявності в наборі даних питань, на які немає відповіді. Оскільки такі дані поза межами цієї роботи, тут використана лише точність — відношення правильно спрогнозованих відповідей до загальної кількості відповідей.

Крім того, для оцінки моделі порівняно з іншими моделями можна використовувати низку мір-орієнтирів — фіксовані завдання та набори даних, розроблені спеціально для складання моделей NLP одна проти іншої. Такими орієнтирами є SuperGLUE [43] - для загальних завдань на розуміння читання, BLEU [44] - оцінка багатомовних моделей та XTREME [45] — завдань на розуміння читання на багатомовних даних.

В цій роботі використані міри SQuAD та Natural Questions.

4.4 Навчання моделей задачі QA та оцінка результатів

Для вирішення задачі QA було натреновано декілька моделей, що засновуються на існуючих моделях, а саме на mBERT та RoBERTa.

Для тренування було зібрано набір даних українською мовою з перекладеного з англійської мови набору SQuAD. Після підчищення даних від перекладів поганої якості, в набір складається із близько 1300 прикладів. Приклади об’єднані за тематиками, тобто у декількох питань є один контекст. Контексти являють собою невеликі тексти (приблизно 200-300 слів в кожному). Для кожного питання є лише один очікуваний правильний варіант відповіді.

В наборі даних SQuAD також існують питання на які немає відповідей. Такі приклади були виключені із навчання на стадії підготовки даних. В цьому дослідженні ми припускаємо що для будь-якого питання в тексті має бути відповідь. Якби довелося вирішувати проблему ідентифікації питань на які неможливо дати відповідь з тексту, необхідно було б отримувати від моделі додатково впевненність у відповіді і перевіряти її по граничному значенню. Таке граничне значення можна встановлювати емпірично або за допомогою ще одного шару нейронів в нейромережі.

На рисунку 4.2 зображена архітектура мережі, що вирішує задачу QA.

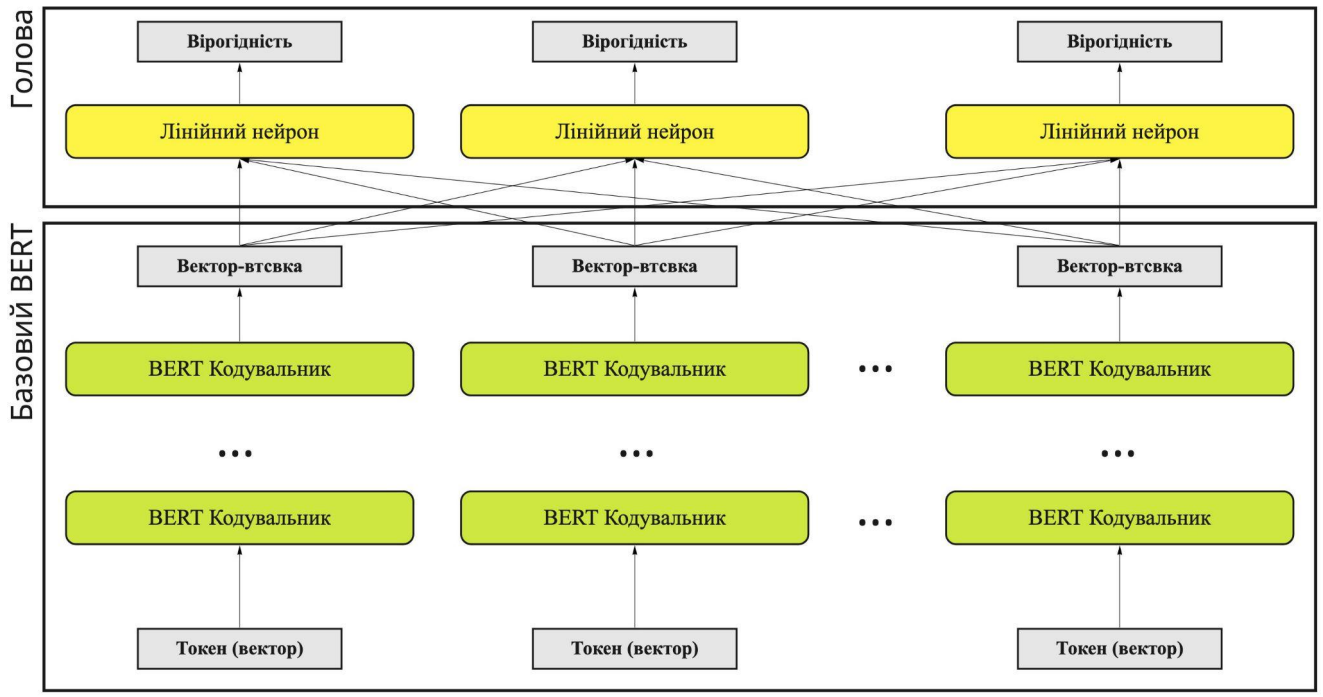

Рисунок 4.2 — Архітектура моделі для вирішення задачі QA

Навчання моделей побудованих на *BERT-подібних архітектурах для відповідей на питання здійснюється за допомогою додавання двох голів кожна по одному шару нейронів до виходів початкової нейромережі. Кожна голова генерує відповідно індекс початку та кінця відповіді. Архітектура голів однакова. Шар нейронів щільний, тобто кожен нейрон отримує ввід від кожного з нейронів попереднього шару, тобто вивід базової моделі. Кожен з нейронів відповідає за один із токенів в контексті. Найбільший вивід означає що саме цей токен має бути першим або останнім з точки зору моделі.

Цей підхід був випробуваний і його ефективність доведена в попередньому дослідженні [46]. В цьому дослідженні автори використовують навчену на англійській мові модель BERT і досягають на метриці SQuAD показника F-міри у 77,827. Це свідчить про потенціал підходу, тому саме така архітектура була обрана для цього дослідження.

Навчання проводилося за допомогою бібліотеки Transformers [47].

Спочатку, до-навчалася модель RoBERTa. На рисунку 4.3 зображений графік зменшення похибки під час навчання. По абсцисі — номер пакета прикладів від першого до 198.

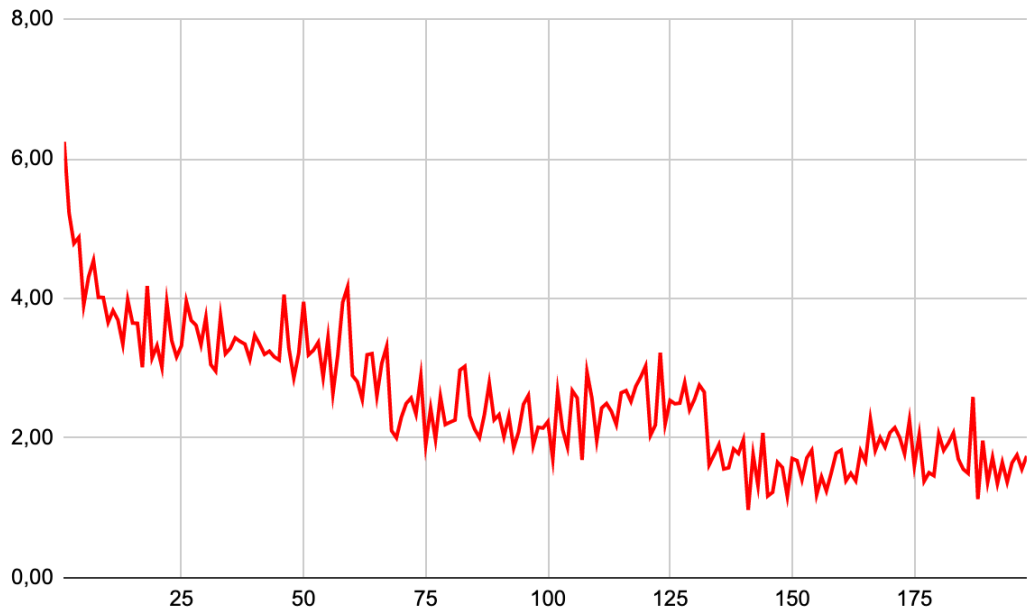

Рисунок 4.3 — Графік зменшення значення похибки під час навчання на моделі RoBERTa

Ця модель спеціально натренована виключно на українській мові. Навчання проводилося в три епохи. Всього було використано 1053 приклади. 351 приклад був збережений для валідації, а отже не брав участі в навчанні. Дані оброблялися пакетами по 16 прикладів.

На відміну від наведеного вище дослідження, в цьому дослідженні ми використовуємо та порівнюємо одну спеціалізовану українську та одну багатомовну модель.

Потім, до-навчалася модель mBERT. Ця модель натренована багатьох мовах, одна з яких — українська, а ще одна — англійська.

Навчання так само проводилося в три епохи. Всього було використано ті самі 1053 приклади, що і для моделі RoBERTa. Дані також оброблялися пакетами по 16 прикладів.

На рисунку 4.4 зображений графік зменшення похибки під час навчання.

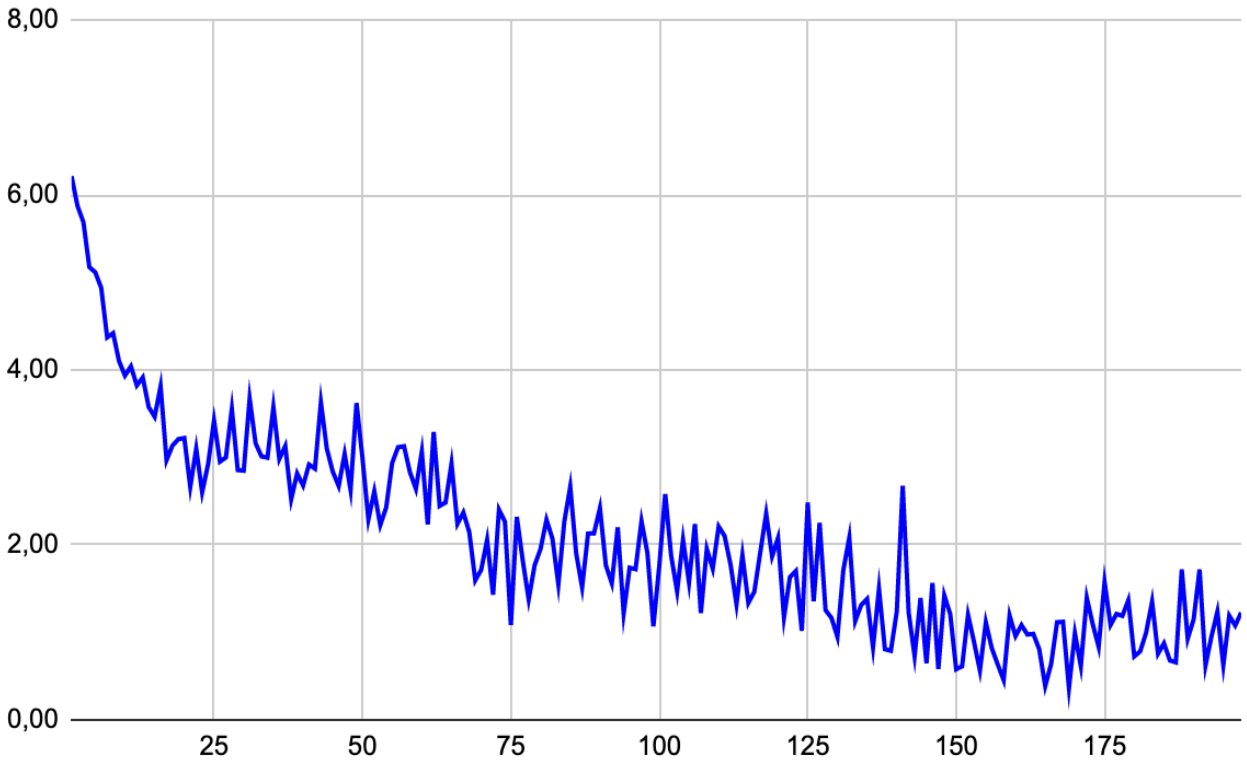

Рисунок 4.4 — Графік зменшення значення похибки під час навчання на моделі mBERT

Виходячи з того що RoBERTa заточена під роботу з українською мовою, а mBERT більш загальний, було поставлено гіпотезу про те що RoBERTa повинна показати кращий результат ніж mBERT.

В результаті перевірки на валідаційних даних (351 приклад), гіпотезу було спростовано. На практиці, точність передбачень RoBERTa — 13,96%, тоді як точність передбачень mBERT — 34,19%.

Зазначимо що ці числа виглядають невеликими, проте треба пам’ятати що набір даних передбачає лише одну можливу правильну відповідь, тоді як практично, їх може бути декілька.

Щодо більшої точності mBERT над RoBERTa, можна зробити припущення що mBERT здатен впоратися з подібними задачами краще. Також, з рисунків 4.2 і 4.3 видно що значення похибки mBERT падає трохи швидше ніж RoBERTa. В цілому, для встановлення справжніх причин, потрібні подальші дослідження.

У наступному експерименті до-навчали моделі на похідному, англомовному, наборі даних SQuAD.

Набір даних SQuAD містить близько 100 тис. прикладів, виключаючі ті, яких питання не мають відповідей. 50 тис. прикладів відводиться на навчання і стільки ж на валідацію.

Кожна з моделей, і mBERT, і RoBERTa, навчалася в однакових умовах. Ідея до-навчати україномовну RoBERTa на англомовних даних досить наївна, але потребує практичної перевірки. Моделі навчалися по три епохи з пакетами по 8 приладів. Через велику кількість даних, навчання проходило на графічній карті, що ніяк не відображається на результатах навчання, а лише на часі.

Гіпотези цього експерименту:

— mBERT отримає точність, вищу за точність RoBERTa.

— на англомовних даних mBERT отримає точність вищу за точність mBERT навченого на українському наборі даних.

— тоді як при тесті на україномовному наборі даних, mBERT навчений на україномовних даних отримає вищу точність ніж mBERT навчений на англомовних даних.

При тесті на англомовних даних, mBERT показав точність 26,57%. На україномовних — 33,9%. При тесті на англомовних даних, RoBERTa показала точність 18,86%. На україномовних — 10,26%.

Навчені на українському наборі даних, mBERT показав точність 10,57% на англомовних тестах, а RoBERTa — 2,28%.

Отже гіпотези виправдалися. mBERT показав в цілому кращі результати ніж RoBERTa. Обидві моделі роблять помітно точніші прогнози на мовах, на яких були навчені.

Нарешті, для ще одного експерименту, ми взяли вже до-навчену модель mBERT, яку навчали на англомовних даних і пропустили її через навчання ще раз, на цей раз на україномовних даних. Ціллю цього експерименту є перевірити, чи покращує точність прогнозів багатомовної моделі змішування мов при тренуванні.

Гіпотеза: після до-навчання на українській, модель проявлятиме вищу точність на українській та на англійській мовах.

Навчання проводилось за тією ж методикою, що і навчання моделей для попередніх експериментів. Всього три епохи. Як і в попередніх експериментах, використовуються ті самі 1053 приклади, розбиті на пакети по 16 прикладів на пакет.

В результаті, отримали модель із такими показниками:

— на україномовних тестах, точність складає 48,14%.

— на англомовних тестах, точність складає 29,14%.

Гіпотеза підтвердилася, дійсно після навчання точність прогнозів моделі виросла на 2,57% на англомовних даних і на 14,24% на україномовних даних.

Бачимо, що модель натренована на англійській і на українській мовах показала кращий результат ніж інші моделі на тестах на обох мовах.

Також, отримані натреновані моделі було перевірено на мірі Natural Questions. Ця міра надає великий набір даних англійською мовою. Відмінність даних Natural Questions від даних SQuAD у меншій вичіщеності.

В таблиці 4 наведені результати всіх описаних випробувань.

Таблиця 4 — Точність прогнозів моделей на різних наборах даних

Мова навчання	Модель	Точність в тестах українською	Тести в тестах англійською
Українська	mBERT	34,19%	10,57%
	RoBERTa	13,96%	2,28%
Англійська	mBERT	33,9%	26,57%
	RoBERTa	10,26%	18,86%
Англійська та українська	mBERT	48,14%	29,14%

Цей набір даних отриманий через обробку веб-сторінок [33]. Через це, в контекстах зустрічаються HTML-теги, заголовки, посилання, тощо. До того ж, набір Natural Questions містить питання з декількома можливими відповідями. Деякі з цих відповідей — довгі, тобто являють собою приблизно абзац тексту. Тоді як інші — короткі, і являють собою одне чи декілька слів. Ідея такого розділення полягає в тому що не на будь-яке питання можна знайти коротку відповідь, наприклад питання що вимагають пояснення подій чи явищ. В той час як знаходження лише довгих відповідей на деякі питання не дає всіх переваг від системи пошуку відповідей на питання, адже користувачеві доведеться вишукувати коротку відповідь в довгому тексті.

Тому міра Natural Questions зазвичай оцінює окремо короткі та довгі відповіді.

В цій роботі наведено сумарні результати оцінювання моделей QA на наборі даних Natural Questions, адже моделі не налаштовані давати кілька варіантів відповідей.

В наборі даних також присутні питання на які немає відповідей. В такому випадку, моделі давали відповідь і при оцінюванні точності такі питання рахувалися як помилки.

Результати оцінювання наведені в таблиці 5.

Таблиця 5 — Точність моделей оцінена на наборі даних Natural Questions

Мова навчання	Модель	Точність
Українська	mBERT	19,76%
	RoBERTa	21,83%
Англійська	mBERT	12,22%
	RoBERTa	22,8%
Англійська та українська	mBERT	11,2%

З результатів видно що RoBERTa впоралася із задачами краще за mBERT незалежно від мови на якій навчалася модель. Це може бути пов’язано із великою кількістю специфічних значень у наборі даних, таких як числа, дати, географічні назви тощо. Подібні сутності легко розпізнаються як моделями так і людьми і простіше вгадуються.

Також відзначимо, що модель mBERT натренована на англійській мові, представлена в цій роботі в двох варіантах: виключно англомовна та англо-україномовна — показала досить низький результат порівняно із іншими моделями. В той же час, модель mBERT натренована виключно на українській мові показала результат який можна порівняти із результатами моделі RoBERTa. Це може бути пов’язане, як і порівняний успіх RoBERTa, із розпізнаванням специфічних сутностей в тексті.

5 АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ ТА ПОТЕНЦІАЛ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

З отриманих результатів можна отримати декілька важливих висновків.

По-перше, зважаючи на порівняно невеликі об’єми тренувальних даних, слід звернути увагу на високу ефективність та роботоздатність моделей у виконанні задач NLG в цілому та конкретно задачі QA. Окремі моделі показали адекватні результати на двох мірах: SQuAD та Natural Questions.

Відзначимо, що на різних завданнях різні моделі показують різні рівні точності. Так, наприклад, модель mBERT показали кращі результати на наборі даних SQuAD, тобто на питаннях більш схожих на ті, на якій тренувалася, в той час, як модель RoBERTa показала кращі результати на наборі даних Natural Questions, тобто на менш знайомих типах запитань. Було підтверджено що комбінація навчання mBERT на двох мовах: англійській та українській — покращує точність прогнозів моделі для обох мов, порівняно із моделями що навчалися лише на англійській чи лише на українській.

Це свідчить про необхідність комбінування різних архітектур, мов тренуванн та наборів даних для тренування для отримання найбільш точних передбачень на найбільшому обсязі різноманітних питань та в різних умовах. Про це також свідчать результати експерименту із більш простою задачею заповнення маски, в якому було отримане значно вище значення точності прогнозів ансамблю двох моделей, ніж прогнозів кожної з моделей окремо.

Подальші дослідження в QA та в NLG в цілому, особливо в контексті української мови, мають широкий спектр потенційних покращень та нововведень. на думку автора цієї роботи, такі дослідження обов’язково мають починатися із створення великих україномовних наборів даних, які можна було б використовувати для вирішення різноманітних завдань. Одним з таких наборів даних має бути аналог SQuAD для української мови. Саме такий набір може

розпочати великий хвилю інтересу та привернути увагу дослідників до задачі QA в українській мові та генерації мови в цілому.

Подальші дослідження засновані на спеціалізованих україномовних та на багатомовних моделях мають показати та встановити нові рівні точності прогнозів у вирішенні цієї задачі. Зокрема, засновуючись на результатах цієї роботи, дослідники зможуть ефективно комбінувати *BERT-подібні моделі в ансамблі задля покращення результатів у вирішенні задачі відповіді на питання.

Ще одним цікавим напрямом дослідження є тренування *BERT-подібних моделей для ідентифікації питань на які неможливо знайти відповідь. Подібні приклади залишилися за межами цієї роботи.

6 ПОДАЛЬШИЙ РОЗВИТОК ТА ВПРОВАДЖЕННЯ ДОСЛІДЖЕНЬ

6.1 Можливі сфери використання

Алгоритм що вирішує задачу QA, по суті, є системою що дозволяє здійснювати пошук по не форматованому тексту за допомогою простих запитань замість ключових слів або фільтрів. Така система має безліч можливих застосувань, як для бізнесів, так і для простих користувачів у відкритих програмних системах. Прикладами таких застосувань для бізнесів є:

— інформаційні системи для співробітників, які дозволяють знаходити інформацію за запитом клієнтів.

— універсальні помічники у вигляді чат-ботів, що видають інформацію з закритої бази знань компанії.

Прикладами застосувань алгоритмів QA для пересічних користувачів є:

— чат-боти на сайтах або на стендах в публічних місцях, які допомагають знайти потрібну інформацію в умовах незнайомого інтерфейсу.

— в комбінації із програмним забезпеченням що перетворює мовлення в текст і текст у мовлення, публічні точки допомоги в містах, які орієнтують громадян по місту, наприклад в аеропортах та вокзалах.

— більш корисні пошукові системи, які видають інформацію за запитом без потреби відкриття посилань на сайти-першоджерела.

6.2 Демонстраційна програмна система

Для демонстрації роботи моделей було розроблено демонстраційну програмну систему, яка дозволяє користувачеві задавати питання до введеного тексту.

Система складається із веб-застосунку та серверної частини. У серверній частині міститься програмний код що здійснює обробку тексту та запускає натреновану модель.

Користувач, за допомогою графічного інтерфейсу, може обрати модель, за допомогою якої буде здійснюватися обробка.

На рисунку 6.1 представлений знімок екрану користувацького інтерфейса програмної системи.

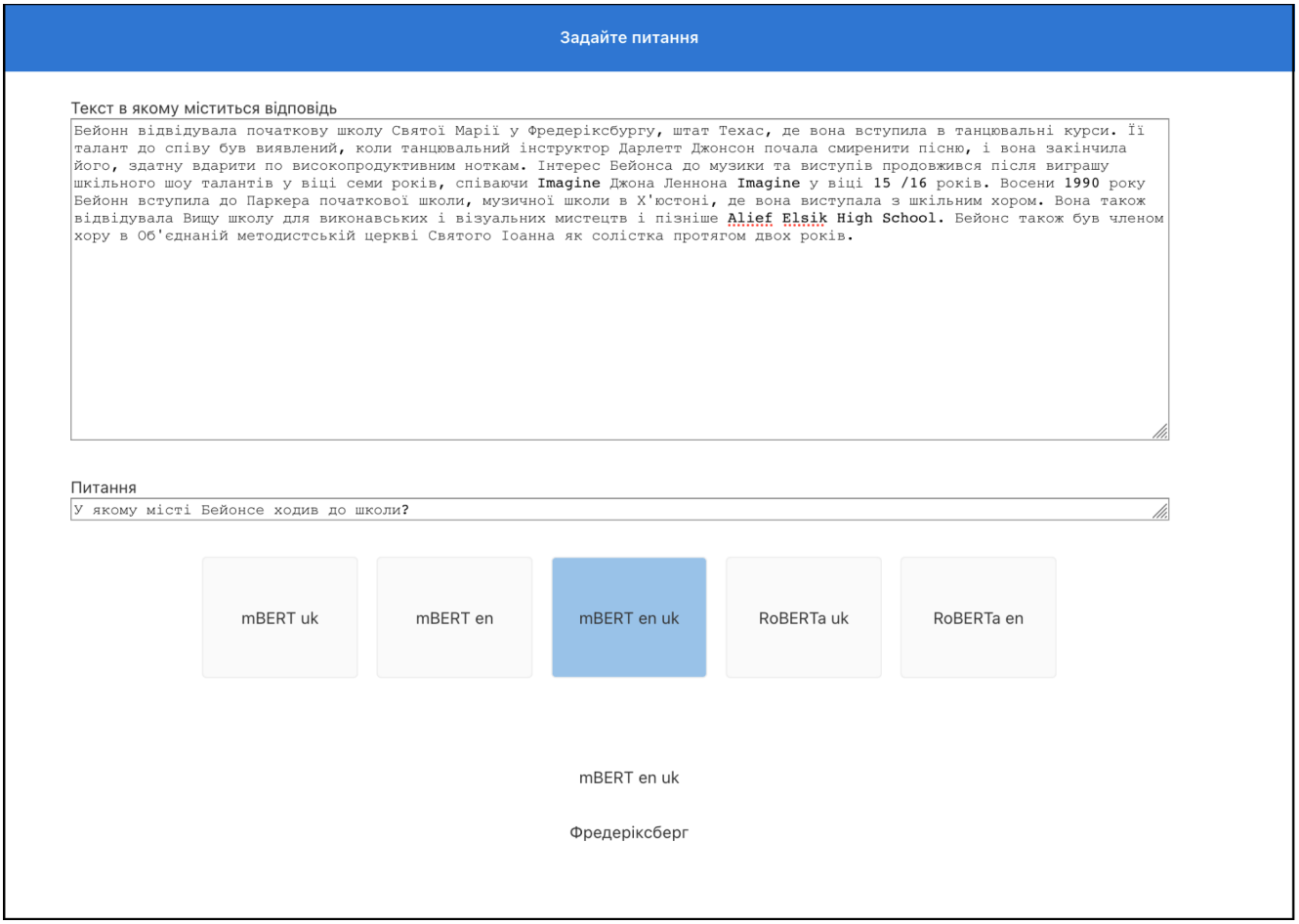

Рисунок 6.1 — Користувацький інтерфейс програмної системи

Клієнтська частина системи реалізована на мові JavaScript з використанням бібліотеки Angular.

Серверна частина системи реалізована на мові програмування Python 3-ї версії з використанням бібліонеми Django.

Програмна реалізація прогнозування за допомогою навчальних моделей створення на основі бібліотек Transformers та PyTorch.

Система не зберігає дані про минулі виклики і не має бази даних.

Як показано на рисунку 6.1, користувач вводить контекст питання в велике поле для вводу. Далі, користувач вводить питання в поле для вводу, що знаходиться безпосередньо нижче поля для вводу контексту питання. Користувач обирає модель, якою бажає обробити питання шляхом натискання на кнопку із назвою відповідної моделі. Коди uk позначає модель навчену на даних українською мовою. Код en позначає модель навчену на даних англійською мовою. По натисканню на кнопку із назвою моделі, відправляється запит на сервер. Всі моделі завантажені в пам’ять на сервері. По запиту, сервер обирає необхідну модель та отримує прогноз з її допомогою. Прогнозована відповідь відправляється на веб-застосунок, який відображає її поруч із вхідними даними та назвою обраної моделі.

ВИСНОВКИ

В ході роботи було проведено аналіз методів вирішення задачі QA у контексті обробки української та англійської мов. Немає причин вважати що отримані результати не є актуальними і для інших європейських мов.

NLP — відносно нова галузь комп'ютерних наук, яка несе в собі значний потенціал і цікава як з академічної, так і з прикладної точки зору. Кількість та якість досліджень українських та зарубіжних авторів вказує на значний інтерес, як зі сторони незалежних дослідників, так і зі сторони великих компаній та урядів держав, які, вбачаючи в результатах дослідження NLP потенціал, впроваджують значні інвестиції в дану сферу.

Роботи українських та зарубіжних дослідників становлять міцний фундамент для проведення нових досліджень, зокрема і для цієї роботи. При цьому, зважаючи на великий обсяг можливих не проведених експериментів, брак альтернативних рішень, особливо для української мови, та загальну корисність артефактів досліджень, подальші дослідження обіцяють бути плідними.

В попередній роботі було висвітлено особливості обробки української мови за допомогою алгоритмів машинного навчання [48].

В цій роботі було досліджено тонкощі застосування попередньо навчених моделей із сімейства *BERT до завдання заповнення маски, проведено навчання та продемонстровано точність роботи цих моделей у вирішенні задачі QA для української та англійської мов.

На прикладі задачі заповнення маски видно як на основі загальної інформації про мову, що вбудована в нейромережу за рахунок процесу навчання, можлива подальша генерація мови. При цьому, модель, спираючись лише на дані контексту, переважно генерує навантаженні смислом словосполучення.

На прикладі задачі QA показано як за допомогою невеликого до-налаштування моделі шляхом додавання нових шарів нейронів можна вирішувати складні задачі із генерації мови.

Експеримент із заповнення маски показав потенціал використання евристичних підходів із комбінації моделей у ансамблі для отримання значно вищої точності передбачень. Особливість цього підходу полягає в різних, хача і схожих архітектурах використаних моделей: mBERT та RoBERTa. Різноманіття архітектур дозволяє комбінувати різні підходи до навчання та інформацію про мову, здобуту з різних наборів даних.

Успіх підходу створення ансамблів при вирішенні більш простої задачі вказує на потенціал використання цього підходу при вирішенні більш складної задачі, а саме QA.

Для подальших дослідженнях в сфері NLG в український мові в цілому та дослідженнях конкретно задачі QA, необхідне створення великого, очищеного, промаркованого відкритого набору даних, яким могли б користуватися дослідники представники різних інституцій.

В цілому, задача відповіді на запитання потребує подальшого дослідження та збільшення точності роботи відкритих моделей. Досвід зарубіжних колег, особливо в дослідженні англійської мови, показує, що колаборація академічних установ та приватних дослідників є дуже ефективним шляхом і призводить до покращення результатів на підвищеного інтересу до галузі. Необхідною умовою для такої колаборації є створення і розвиток відкритих наборів даних та архітектур моделей.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Perreault C., Mathew S. Dating the origin of language using phonemic diversity // PLOS ONE, 2012. № 7.

2. What are the largest language families? // Ethnologue. [Електронний ресурс] URL: https://www.ethnologue.com/guides/largest-families (дата звернення 03.03.2020).

3. Sokolova M., Japkowicz N., Szpakowicz S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. // AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science. 2006. T. 4304. C. 1015-1021.

4. Bozinovski S., Fulgosi A. The influence of pattern similarity and transfer learning upon the training of a base perceptron B2. // Proceedings of Symposium Informatica 3-121-5, Bled, 1976.

5. Gruzdo I., Kyrychenko I., Tereshchenko G., Cherednichenko O. Application of paragraphs vectors model for semantic text analysis // CEUR Workshop Proceedings, 2020, № 2604, C. 283-293.

6. Hutchins, J. The history of machine translation in a nutshell [Електронний ресурс] URL: https://docplayer.net/18621555-The-history-of-machine-translation-in-a-nutshell (дата звернення 10.03.2021)

7. Gagniuc P.A. Markov Chains: From Theory to Implementation and Experimentation: USA, NJ: John Wiley & Sons., 2017. C. 132.

8. Goodfellow I, Bengio Y., Courville A. Deep Learning. Boston, USA: MIT Press, 2016.

9. Tiutiunnyk S., Dyomkin V. Context-Based Question-Answering System for the Ukrainian language // CEUR Workshop proceedings. 2019.

10. Kuchuk H., Kovalenko A., Ibrahim B.F., Ruban I. Adaptive compression method for video information // International Journal of Advanced Trends in Computer Science and Engineering. 2019. T. 8, №. 1.2, C. 66-69.

11. Aggarwal C.C., Zhai C.A. Mining Text Data. Springer: Survey of Text Classification Algorithms. Boston, USA. 2012.

12. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention is all you need // NIPS Proceedings. 2017.

13. Devlin, J., Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // NAACL-HLT, 2019.

14. Yinhan L., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach // ArXiv [Електронний ресурс] URL: http://arxiv.org/abs/1907.11692 (дата звернення 02.12.2020).

15. Shijie W., Dredze M.. Are All Languages Created Equal in Multilingual BERT? // ArXiv [Електронний ресурс] URL: http://arxiv.org/abs/2005.09093 (дата звернення 02.12.2020).

16. Zhenzhong L., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations // International Conference on Learning Representations [Електронний ресурс]. 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS

17. Radford, A., Jeffrey W., Child R., Luan D., Amodei D., Sutskever I. Language Models are Unsupervised Multitask Learners. 2019 [Електронний ресурс]. URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (дата звернення 02.12.2020).

18. Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A. Language Models Are Few-Shot Learners. 2020. [Електронний ресурс] URL: https://arxiv.org/abs/2005.14165 (дата звернення 02.12.2020).

19. Martin Radfar, Athanasios Mouchtaris, Siegfried Kunzmann. End-to-End Neural Transformer Based Spoken Language Understanding // Interspeech 2020, Shanghai, China. 2020.

20. A. Yerokhin, A. Babii and O. Turuta. Geoscience Laser Altimeter System sparse ICESat data processing based on F-transform // IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL). 2019. C. 553-556.

21. A. Yerokhin, A. Nechyporenko, A. Babii and O. Turuta. Usage of F-transform to finding informative parameters of rhinomanometric signals // Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. C. 129-132.

22. A. Yerokhin, A. Nechyporenko, A. Babii and O. Turuta. Processing and analysis of rhinomanometric signals by F-transform approximation // IEEE First International Conference on Data Stream Mining & Processing (DSMP). 2016.

23. We Trained the Ukrainian Language Model. [Електронний ресурс] URL: https://youscan.io/blog/ukrainian-language-model (дата звернення 01.02.2020).

24. Suarez O., Javier P., Romary L., Sagot B. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020. C 1703-1714.

25. Homepage: lang-uk. [Електронний ресурс]. URL: https://lang.org.ua/en/models (дата звернення 01.02.2020).

26. Khaburska A., Tytyk I. Toward Language Modelling for the Ukrainian Language // CEUR Workshop proceedings. 2019.

27. Shvedova M. The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality // CEUR Workshop proceedings. 2020.

28. Grabar N., Hamon T. Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploration. // Colins AI, Kharkiv, Ukraine. 2017.

29. З. В. Дударь, Д. Е. Шуклин. Семантическая нейронная сеть, как формальный язык описания и обработки смысла текстов на естественном языке. // Радиоэлектроника и информатика. №3(12). 2000. С. 10-25.

30. Rajpurkar P., Jia R., Liang P.. Know What You Don’t Know: Unanswerable Questions for SQuAD // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. T. 2.

31. Microsoft Research WikiQA Corpus. [Електронний ресурс]. URL: https://www.microsoft.com/download/confirmation.aspx?id=52419 (дата доступу 28.11.2020).

32. Dunn M., Sagun L., Higgins M, Guney V. U., Volkan Cirik, Kyunghyun Cho. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine // ArXiv [Електронний ресурс] URL: http://arxiv.org/abs/1704.05179 (дата доступу 28.11.2020).

33. List of Wikipedias. [Електронний ресурс]. URL: https://en.wikipedia.org/wiki/List_of_Wikipedias (дата доступу 02.12.2020).

34. DMQA [Електронний ресурс]. URL: https://cs.nyu.edu/~kcho/DMQA (дата доступу 02.12.2020).

35. Bajaj Payal, Campos D., Craswell N., Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset // CEUR Proceedings. 2016. T. 1773.

36. Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter // EMC^2: 5th Edition Co-located with NeurIPS'19, Vancouver , Canada. 2019.

37. The AI community building the future. [Електронний ресурс] URL: https://huggingface.co/ (дата доступу 05.04.2021).

38. Google's Natural Questions. [Електронний ресурс] URL: https://ai.google.com/research/NaturalQuestions (дата доступу 03.04.2021).

39. Bang Liu, Haojie Wei, Di Niu, Haolan Chen, Yancheng He. Asking Questions the Human Way: Scalable Question-Answer Generation from Text Corpus. // Proceedings of The Web Conference 2020 (WWW '20). Association for Computing Machinery, New York, USA. 2020.

40. Yerokhin A., Turuta O., Babii A., Nechyporenko A. Intelligent information system of heterogeneous medical data analysis. // Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT. 2017.

41. Smelyakov K., Ponomarenko O., Chupryna A., Tovchyrechko D., Ruban I. Local Feature Detectors Performance Analysis on Digital Image. // IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), Kyiv, Ukraine. 2019.

42. Getting started with Language Translator. [Електронний ресурс] URL: https://cloud.ibm.com/docs/language-translator?topic=language-translator-gettingstarted (дата доступу 12.01.2021).

43. SuperGLUE Benchmark. [Електронний ресурс] URL: https://super.gluebenchmark.com/, (дата доступу 12.04.2021).

44. bleu.dvi. [Електронний ресурс] URL: https://www.aclweb.org/anthology/P02-1040.pdf (дата доступу 12.01.2021).

45. XTREME. [Електронний ресурс] URL: https://sites.research.google/xtreme (дата доступу 12.01.2021).

46. Yuwen Zhang, Zhaozhuo Xu. BERT for Question Answering on SQuAD 2.0 [Електронний ресурс] URL: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15848021.pdf (дата доступу 27.02.2021).

47. Transformers: State-of-the-art Natural Language Processing for Pytorch and TensorFlow 2.0. [Електронний ресурс] URL: https://github.com/huggingface/transformers (дата доступу 25.04.2021).

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