МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

Факультет Комп'ютерних наук
Кафедра Програмної інженерії

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

Дослідження керованості англійсько-українського машинного перекладу на основі спеціалізованих корпусів. Набори даних

Виконала: студентка 2 курсу групи ІПЗм-21-2
Сайчишина Н. С.
(прізвище, ініціали)

Спеціальність	121 – Інженерія програмного забезпечення
Тип програми	Освітньо-наукова
Керівник	доцент каф. ПІ. Турута О. П.

(посада, прізвище, ініціали)

Допускається до захисту

Зав. Кафедри _____ З.В. Дудар
(прізвище, ініціали)

2023 р.

Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____

Кафедра _____Програмної Інженерії_____

Рівень вищої освіти _____другий (магістерський)_____

Спеціальність _____121 – Інженерія програмного_забезпечення_____

(код і повна назва)

Тип програми _____освітньо-наукова_____

Освітня програма _____Інженерія програмного забезпечення_____

(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____ _____

(підпис)

«___»_____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентці _____Сайчишиній Наталії Сергіївні_____

(прізвище, ім'я, по батькові)

1. Тема роботи ____«Дослідження керованості англійсько-українського_____ машинного перекладу на основі спеціалізованих корпусів. Набори даних»_____ затверджена указом університету від «29» _березня_ 20_23_р. № _302Ст_____

2. Термін подання студентом роботи до екзамінаційної комісії «18» травня 2023р.

3. Вихідні дані до роботи машинний переклад, набори даних, Python, BERT, пояснювальна записка_____

4. Перелік питань, що потрібно опрацювати в роботі мета роботи, аналіз предметної області, постановка задачі, аналіз наборів даних для методів керованості машинного перекладу, оцінка отриманих результатів_____

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів курсової роботи	Термін виконання етапів роботи	Примітка
1	Аналіз предметної області	04.04.2023	
2	Огляд існуючих рішень	11.04.2023	
3	Аналіз результатів експериментів	21.04.2023	
4	Підготовка пояснювальної записки	22.04.2023	
5	Спецчастина	23.04.2023	
6	Підготовка презентації та доповіді	04.05.2023	
7	Нормконтроль, рецензування	11.05.2023	
8	Занесення диплома в електронний архів	16.05.2023	
9	Попередній захист	16.05.2023	
10	Допуск до захисту у зав. кафедри	17.05.2023	

Дата видачі завдання 29 березня 20 23 р.

Студент _____ ___
(підпис)

Керівник роботи _____ доцент каф. ПІ Турута О.П.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Кваліфікаційна робота магістра містить 101 сторінку, 22 рисунки, 4 таблиці, 7 додатків, 33 джерела.

АНЛГІЙСЬКА МОВА, МАШИННИЙ ПЕРЕКЛАД, НАБІР ДАНИХ, УКРАЇНСЬКА МОВА.

Об’єктом дослідження є набори даних для впровадження керованості машинного перекладу з англійської на українську мови, способи оцінки отриманих машинних перекладів.

Предмет дослідження це вплив наборів даних, що містять дані різного стилістичного забарвлення на керованість отриманого машинного перекладу із англійської мови на українську.

Методи розробки базуються на таких технологіях як Python3, huggingface datasets, мовні моделі BERT і MiniLM.

Практична цінність отриманих результатів полягає в удосконаленні машинного перекладу з англійської мови на українську, а також створення якісних наборів даних, на яких можуть ґрунтуватися майбутні дослідження з області машинного перекладу, генерації текстів, опису зображень, семантичного пошуку.

DATASET, ENGLISH LANGUAGE, MACHINE TRANSLATION, UKRAINIAN LANGUAGE.

The object of the research is data sets for the implementation of controllability of machine translation from English to Ukrainian, methods of evaluating the obtained machine translations.

The subject of the study is the influence of data sets containing data of different stylistic colors on the controllability of the resulting machine translation from English to Ukrainian.

Development methods are based on such technologies as Python3, huggingface datasets, language models BERT and MiniLM.

The practical value of the obtained results lies in the improvement of machine translation from English to Ukrainian, as well as the creation of high-quality data sets on which future research in the field of machine translation, text generation, image description, and semantic search can be based.

Умови публікації пояснювальної записки

Я,
_Сайчишина Наталія Сергіївна_____
(прізвище, ім'я, по батькові)
студентка групи __ІПЗм-21-2__здобувач вищої освіти на другому (магістерському) рівні
кафедра_____програмної інженерії_____,
(повна назва кафедри)
заявляю: моя кваліфікаційна робота на тему
_____ Дослідження керованості англійсько-українського машинного перекладу на основі спеціалізованих корпусів. Набори даних_____,
(назва роботи)
що буде представлена до ЕК для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

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

ЗМІСТ

ВСТУП

Дослідження керованості в області обробки природньої мови являється актуальним і стрімко набираючим розголосу та практичного застосування напрямком машинного навчання. Саме завдяки дослідженням у цій галузі стало можливим генерувати тексти, або отримувати переклади у довільному лексичному стилі. Таким чином, з'явилася можливість змінювати стиль, до прикладу, із повсякденного, тобто розмовного, на офіційно-діловий стиль шляхом додавання специфічної стилістично-забарвленої лексики, та особливих мовних конструкцій.

У 2003 році канадським вченим Йошіо Бенгіо та його командою було оприлюднено перший експеримент [1] отримання машинного перекладу з використанням нейронної мережі. Тодішня нейронна мережа, звичайно, значно відрізнялася від сучасних рішень, проте ця робота стала вагомим відкриттям у області нейронних машинних перекладів.

Аби мав місце подальший розвиток сучасних методів обробки природньої мови, постає нагальна потреба у зборі та упорядкуванні суттєвої кількості наборів даних для подальшого тренування моделей машинного навчання. Крім цього, аби досягти зміни стилістичного забарвлення текстової інформації, такі тренувальні дані мають охоплювати значну кількість прикладів із відповідної сфери [2].

Варто зазначити, що для української мови, на момент дослідження, не існує достатньої кількості наборів даних, які згруповані за певною темою. Через це, українську мову відносять до мов, що є малоресурсними, тобто проведення подібних досліджень значно ускладнено порівняно, до прикладу, із англійською.

Удосконалення якості машинного перекладу, що розглядається у рамках кваліфікаційної роботи, безумовно є актуальною та важливою задачею, до вирішення якої пропонується застосувати сучасні методи обробки природньої мови, які ґрунтуватимуться на підготовлених якісних наборах даних. Важливість якісного машинного перекладу полягає у збільшенні обсягу наявної інформації та ресурсів, доступних іншими мовами. Це сприятиме поширенню та популяризації української мови та культури і зробить інформацію українською мовою більш

доступною і зрозумілою. У результаті збільшуватиметься кількість виданих фільмів, з’являтимуться переклади та локалізації сервісів, ігор, програмних додатків тощо. Також це сприятиме ефективній бізнес-комунікації, полегшуючи підприємствам співпрацю із українськими споживачами і партнерами.

Метою дослідження є покращити якість машинного перекладу для української мови, шляхом збільшення кількості відкритих наборів, дослідження впливу зібраних наборів даних на керованість при машинному перекладі із англійської на українську мову, аналізу існуючих метрик оцінювання якості отриманих перекладів і доречності їхнього використання.

Об’єктом дослідження є набори даних для впровадження керованості машинного перекладу з англійської на українську мови, способи оцінки отриманих машинних перекладів.

Предмет дослідження це вплив наборів даних, що містять дані різного стилістичного забарвлення на керованість отриманого машинного перекладу із англійської мови на українську.

У ході дослідження використано методи аналізу і підготовки якісних наборів даних для машинного навчання, обробки природньої мови для оцінки отриманих результатів,

Практична цінність отриманих результатів полягає в удосконаленні машинного перекладу з англійської мови на українську, а також створення якісних наборів даних, на яких можуть ґрунтуватися майбутні дослідження з області машинного перекладу, генерації текстів, опису зображень, семантичного пошуку.

Перші результати досліджень керованості текстів для української мови на прикладі частини набору даних Multi30k були представлені на конференції CSIT 2022. Подальші дослідження були представлені на заході IST 2022. Демонстрація заключних етапів дослідження і підготовленого набору даних Multi30k відбулася 5 травня 2023 на українському воркшопі UNLP 2023 у рамках EACL 2023 (17th Conference of the European Chapter of the Association for Computational Linguistics).

1 ОПИС ПРОБЛЕМНОЇ ГАЛУЗІ

1.1 Аналіз предметної області

Для того, аби впровадити керованість при машинному перекладі тексту, необхідно мати достатню кількість даних для навчання. Крім того, дані мають бути згруповані та містити саме ті лінгвістичні атрибути, які планується застосувати до вхідного тексту.

Створення високоякісних та великих за розміром наборів даних для таких малоресурсних мов як українська, здатне покращити якість машинного перекладу та надати можливість змінити стиль згенерованого тексту [3]. Останнім часом, методи машинного навчання використовуються для вирішення низки складних проблем, на кшталт перекладу, розпізнавання мовлення і обробки природної мови. Проте для ефективної розробки та навчання моделей, які можуть вирішувати згадані проблеми на достатньо високому рівні, вкрай важливо мати доступ до великих обсягів високоякісних та перевірених даних.

Обмежена кількість даних для навчання означає, що моделям може бути складно правильно інтерпретувати нюанси та особливості мови, що неодмінно призведе до зниження точності та погіршення отриманих результатів. Саме на цьому етапі створення та обробки наборів даних це стає настільки критичною проблемою, яка не дозволяє просуватися у вирішенні задачі навіть маючи складну модель і значні обчислювальні ресурси. Впроваджуючи набори даних, до малоресурсних мов, стає можливим ефективне навчання моделей і досягнення кращих результатів.

Одним з підходів, який нерідко використовується протягом останніх років, є створення і впорядкування мультимодальних наборів даних [4]. Такі набори містять різні типи даних, кожен з яких репрезентує різну модальність інформації. Об’єднуючи текст, зображення, аудіо та інші типи даних, мультимодальний набір даних може охоплювати різний тип та представлення інформації та надати більш точне уявлення про мову, для якої проводиться дослідження.

Залучення мультимодальних наборів даних для подібних досліджень виявилося напрочуд ефективним для малоресурсних мов. Репрезентуючи більш детальне уявлення про мову, такі набори даних сприяють ефективному та змістовному етапу навчання моделей, що так чи інакше призводить до кращої точності та ефективності [5]. Тож, упорядкування високоякісних мультимодальних наборів даних є надважливим кроком у розвитку майбутніх досліджень обробки природньої мови, що також включає у себе задачу керованості.

1.2 Огляд наукової літератури

Задача, що розглядається у рамках кваліфікаційної роботи є актуальною і досліджуваною, що підтверджується низкою публікацій, що стосуються певних аспектів обраної задачі [6]. Таким чином, аби ознайомитися із існуючими напрацюваннями було проведено аналіз наукових публікацій за двома основними напрямками: керованість у задачах перекладу текстів та машинний переклад для малоресурсних мов.

Проблема керованості у задачах машинного перекладу текстів не є вирішеною, проте має суттєві напрацювання щодо можливих методів отримання задовільних результатів. Загалом, у області генерації та перекладу текстів зростає інтерес до контролювання наявності певних лінгвістичних атрибутів у вихідному тексті. Такими атрибутами можуть бути: довжина, ввічливість, емоційна забарвленість, тема тощо [7].

Контроль за такими особливостями зазвичай здійснюється за допомогою керуючих кодів: категоріальних змінних, які представляють необхідну вихідну властивість і попередньо додаються до вхідних даних моделі під час навчання та тестування. Маючи велику кількість якісних наборів даних, відповідні моделі, що призначені для вирішення саме такого типу задач, та потужні обчислювальні ресурси, моделі можуть навчитися достатньо точно та високоякісно підхоплювати

стилістичні особливості текстів із тренувальної вибірки та відтворювати їх при генерації тексту.

Одним із методів для досягнення керованості при генерації тексту є використання підходу, при якому модель навчається створювати текст на основі набору попередньо визначених вхідних даних або підказок. Наприклад, у задачі машинного перекладу модель може бути обумовлена конкретним вихідним текстом.

Інший підхід полягає у використанні таких методів, як передача стилю, коли модель навчається змінювати існуючий текст, щоб відповідати певним вимогам або обмеженням. Цей підхід може бути особливо корисним у завданнях машинного перекладу для досягнення перекладів тексту у різних стилістичних формах.

Однак досягнення високих рівнів керованості при генерації тексту все ще є складною проблемою, і досі тривають дослідження щодо розробки більш ефективних методів і технік для покращення результатів моделей машинного навчання.

Задачі машинного перекладу для малоресурсних мов [8], до яких належить українська мова, через відсутність значної кількості наборів даних для навчання моделей нейронних мереж, також актуальні і досліджувані. Зокрема методи, якими необхідно оперувати у такій ситуації суттєво відрізняються від тих мов, для яких доступна велика кількість даних. Застосована архітектура зазвичай спирається на великомасштабні паралельні набори даних. Це є перевагою, котра недоступна для малоресурсних мов.

Одним із можливих рішень є – синтетична генерація даних, що стосується методів доповнення даних. Такі методи можна застосовувати незалежно від використовуваної архітектури моделі машинного навчання. У випадку, коли паралельні дані недоступні, також можуть бути використані методи нейронного машинного перекладу без учителя [9]. У випадку, якщо доступні паралельні дані, побудова двомовних моделей між кожною парою мов не є вирішальним способом вирішення задачі. У якості рішення було запропоновано моделі, які полегшують переклад між більш ніж однією мовною парою за допомогою однієї моделі.

Більшість мультимовних моделей базуються на принципі навчання із учителем, хоча доступні також деякі дослідження щодо застосування методики навчання без учителя у багатомовному середовищі. Хоча мультимовні моделі спочатку були запропоновані, щоб уникнути необхідності побудови окремих двомовних моделей перекладу, їх можливість перекладу для пар, де є малоресурсна мова, є насправді багатообіцяючою [10].

1.3 Обґрунтування цілей дослідження

Основною ціллю дослідження є вдосконалення нейронного машинного перекладу для української мови шляхом впровадження нового мультимодального набору даних. Дослідження опосередковано стосується вирішення задачі машинного перекладу та семантичного пошуку для малоресурсних [11] мов, оскільки в ході роботи постає необхідність протестувати набори даних на ширшому колі задач обробки природньої мови.

Головний етап, на якому буде зосереджено увагу – це розробка мультимодального набору даних, який включатиме ряд відмінних типів даних, таких як текст та зображення [12]. Об’єднуючи кілька доменів, набір даних охоплюватиме більш ніж 30 000 речень українською та англійською мовами, що зробить його ефективним і корисним інструментом на етапах навчання і донавчання моделей. Також, дослідження супроводжуватиметься ознайомленням і ретельним аналізом сучасних методів обробки природньої мови та оцінки отриманих результатів.

Окрім створення вище зазначеного набору даних, робота також передбачає дослідження впливу впровадженого набору на задачу керованості під час машинного перекладу для української мови.

Також важливою задачею є оцінка якості отриманого перекладу. У завданнях обробки природньої мови важливо використовувати декілька різних метрик для оцінювання для перевірки ефективності моделі, оскільки різні метрики фіксують

різні показники ефективності моделі. Варто зазначити, що жоден показник не охоплює всі можливі способи оцінки ефективності моделі.

Наприклад, точність замірює загальну правильність прогнозів, що надає модель, тоді як метрика F1 здатна зафіксувати як точність, так і ступінь запам’ятовування прогнозів моделі.

Показники також можуть мати деяке зміщення щодо певних типів помилок. Наприклад, обрахована точність може бути дещо зміщена в бік класів, що мають найбільше екземплярів у незбалансованих наборах даних, у свою чергу точність може бути зміщена в бік негативного класу в наборах даних із високою часткою негативних прикладів.

До того ж, зазвичай метрики є спеціалізованими для різних типів завдань. А саме, різні завдання з області обробки природньої мови вимагають різних метрик для оцінювання. Для машинного перекладу здебільшого використовується метрика BLEU, яка є стандартним показником. Для аналізу емоційної забарвленості текстової інформації використовується метрика F1, що є більш доречною. Метрики також здатні допомогти у виборі відповідної моделі та під час налаштування гіперпараметрів [13].

Тож, важливим етапом дослідження є проаналізувати існуючі метрики, щоби оцінити можливість використання наборів даних для визначених задач дослідження. Використання таких показників зробить можливим ефективно та збалансовано оцінити отриманий результат.

Машинний переклад відіграє важливу роль у забезпеченні доступу до інформації осіб, які не володіють певною мовою. Удосконалюючи машинний переклад, є можливість подолати мовний розрив і адаптувати цінні ресурси, такі як книги, статті та онлайн-контент, і зробити їх більш доступними для ширшої аудиторії, незалежно від її мовних знань. Це сприяє інклюзивності та забезпечує рівний доступ до інформації.

2 ПОСТАНОВКА ЗАДАЧІ

2.1 Методи підготовки наборів даних

Наразі існує низка доступних методів збору наборів даних для задач із обробки природної мови та обробки зображень. Їх можна класифікувати наступним чином:

1) веб-скрапінг – що являється одним із найпоширеніших та найпростіших методів колекціонування даних для завдань із обробки природньої мови. Цей метод передбачає використання різноманітних скриптів для автоматичного колекціонування даних із веб-сайтів, соціальних мереж чи інших інтернет-джерел. Проте, використовуючи даний підхід, важливо упевнитися, що дані є відкритими для використання, не містять конфіденційної інформації та проведення збору дозволене платформою;
2) краудсорсинг - передбачає сумісну роботу над одним джерелом даних. Таким чином, формування набору даних відбувається із залученням великої кількості людей, зазвичай за допомогою онлайн платформ, таких як Amazon Mechanical Turk. Метод може приносити значну користь для завдань, що вимагають людської оцінки, таких як аналіз емоційної забарвленості текстів або розпізнавання об’єктів на зображенні;
3) контент, створений користувачами соціальних мереж чи онлайн-платформ – така інформація, може бути корисним джерелом даних для завдань обробки природньої мови. Такі дані мають містити характеристики цільової групи, за певні проміжки часу та включати відповідні анотації;
4) експертні анотації – це спосіб створення наборів даних, які впорядковуються спеціалістами та мають високу точність та експерність. Спеціалістів долучають при формуванні наборів даних, що стосуються медицини, перекладів тощо.

Ще одним відомим методом для створення наборів та корпусів даних є генерація синтетичного тексту, також відомий як генерація природної мови. Це

метод генерування текстової інформації за допомогою алгоритмів і моделей машинного навчання. Хоча цей спосіб має низку переваг, існує також значна кількість недоліків, особливо під час використання цього способу для малоресурсних мов як українська [14].

Для впровадження цієї технології генерації тексту вимагається велика кількість навчальних даних, для створення якісного результату, який у подальшому можна буде використовувати у дослідженнях. Однак, як відомо, малоресурні мови характеризуються тим, що для них відсутні прийнятні обсяги навчальної інформації. Це означає, що синтезований текст може бути не таким точним та логічним, як це було б для для такої мови як англійська.

Крім того, сучасні моделі генерації тексту спираються на чисельний словниковий запас, щоб створювати текст, що звучить природньо [15].

Для української мови доступний словниковий запас, який може бути опрацьований моделлю, може бути обмеженим, що у підсумку призведе до тавтологій або тексту, який передаватиме некоректний зміст. Також, для генерації текстової інформації потребуються потужні обчислювальні ресурси для належної роботи програмного додатку, що знаходиться у розробці.

Як відомо, для будь-якої мови притаманні певні унікальні культурні та мовні нюанси. Коректна генерація текстової інформації із урахуванням таких особливостей може бути ускладнена через нестачу навчальних даних, які б містили такі особливості.

Окрім культурних та мовних нюансів, особливої уваги також потребують предметно-специфічні знання, що можуть вживатися лише у певній, до прикладу, науковій галузі. Аби тренування відбулося вдало, навчальні дані мають включати таку інфомацію із специфічних сфер [16].

Метод для збору наборів даних безумовно залежить від поставленої задачі, наявних ресурсів, часу тощо. Загальним об'єднуючим фактором є те, що дані мають бути репрезентативними, зрозумілими та коректно анотованими, що у підсумку напряму впливатиме на точність моделей машинного навчання.

Для обраної задачі, з огляду на вище наведені факти, найбільш ефективним методом стане розробка набору даних, що міститиме професійні анотації та збиратиметься вручну. Причиною є те, що процес генерації текстів вимагає значної кількості тренувальних даних, що не доступні для малоресурних мов [17].

2.2 Алгоритми машинного перекладу

Для того, аби визначити метод формування набору даних із зазначених вище у пункті 2.1 для задачі машинного перекладу, необхідно ознайомитися із існуючими алгоритмами та на основі цих досліджень обрати найбільш раціональний спосіб.

Одним із можливих рішень є моделі Seq2Seq [18]. На рисунку 1 наведено типову архітектуру для Seq2Seq моделей машинного навчання.

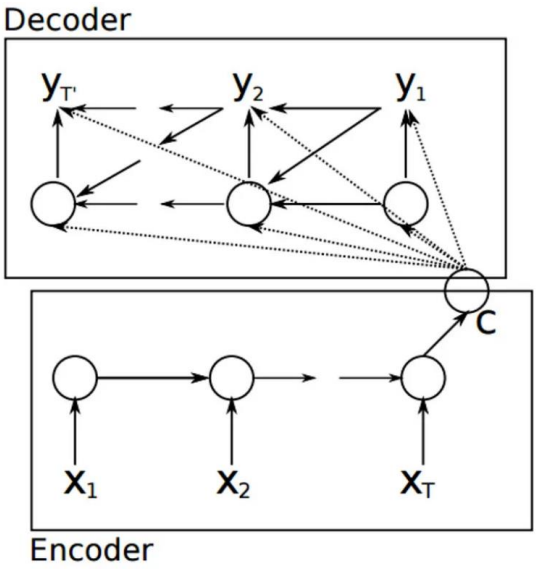

Рисунок 1 – Архітектура моделі Seq2Seq

Такі моделі є найчастіше застосованою архітектурою для подібних задач. Моделі складаються із енкодера та декодера, де у енкодер подається речення вихідною мовою та у результаті отримується представлення у вигляді контекстного вектору.

Наступним етапом декодер генерує вихідне речення цільовою мовою, спираючись на контекстний вектор та генеруючи одне слово.

Encoder зчитує вхідну послідовність і генерує контекстний вектор фіксованого розміру, який представляє семантичний підсумок вхідної послідовності.

Ще одним методом для вирішення подібних задач є трансформери. Такі моделі є відносно новим типом архітектури для нейронного машинного перекладу. Дослідження показують, що трансформери перевершують моделі Seq2Seq за ефективністю навчання та якістю перекладу, що оцінюється стандартними метриками [19].

На відміну від рекурентних нейронних мереж, які використовуються у вище наведених моделях Seq2Seq, трансформери повністю покладаються на алгоритми самоконтролю, аби враховувати контекст послідовності. Архітектура трансформерів складається з двох головних компонентів: енкодера та декодера. Енкодер виконує обробку вихідної послідовністі, в свою чергу декодер генерує цільову послідовність. Енкодер і декодер складаються з кількох рівнів, кожен із яких має два підрівні: механізм самоконтролю і нейронна мережа прямого зв’язку.

В архітектурі трансформерів механізм attention обчислюється інакше, ніж у стандартних моделях Seq2Seq. Він дозволяє моделі звернути увагу на різні частини вхідної послідовності, для того щоб обчислити представлення послідовності на кожному рівні.

Вхідна послідовність подається до послідовності векторів, представленої у формі матриці X форми (L, d), де L — довжина послідовності, а d — розмірність вектору.

Для кожної позиції у вхідній послідовності трансформеру обчислюється три вектори: вектор ключа, вектор запиту та вектор значення. Ці вектори використовуються для подальшого розрахунку ваг уваги. Зазначений механізм уваги для трансформерів розраховується як зазначено у формулі (2.1)

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \quad (2.1)$$

де Q – вектор запитів;

K – вектор ключів;

V – матриця значень;

d_K – розмірність вектору ключів.

Мультилінгвістичні моделі — це моделі нейронного машинного перекладу, які мають змог вирішувати задачі обробки природньої мови для декількох мов одночасно. Такі моделі розроблено для покращення машинного перекладу між кількома мовами, а не з ціллю навчання лише на парі вихідної та цільової мов, як стандартні моделі машинного перекладу [20].

Сильна сторона таких моделей мультилінгвістичного машинного перекладу полягає в тому, що вони мають змогу знизити кількість навчальних даних а також обчислювальних ресурсів, що використовуються, необхідних для досягнення прийнятної якості перекладу між кількома мовами.

Виконуючи навчання однієї модель кількома мовами, стає можливо користуватися схожістю між цими мовами та обмінюватися і порівнювати отриману інформацію. У результаті, це може призвести до більш точних і коректних перекладів, особливо для малоресурсних мов де навчальні дані суттєво обмежені.

2.3 Формулювання задачі

Спираючись на проведене дослідження у пунктах 2.1 і 2.2, аналіз предметної області, огляд наукової літератури, цілі дослідження та сучасний стан вирішення задачі керованості і створення якісних нейронних машинних перекладів для української мови, було визначено що:

- набори даних, на яких ґрунтується подальше навчання моделей мають істотний вплив на рішення цієї задачі. У рамках роботи необхідно проаналізувати існуючі малочисельні набори даних, та створити власний, враховуючи особливості описаної задачі;
- існує обмежена кількість високоякісних наборів даних, які не дозволяють проводити експерименти із керованістю, тобто зміною лексичного забарвлення для української мови;
- стандартні метрики для оцінювання якості перекладів не орієнтуються на певні мовні особливості, тому ускладнюється задача із оцінкою якості отриманих даних;
- існує невирішеність задачі коректного перекладу слів та словосполучень, що являються фразеологізмами, або передають знаходяться у переносному значенні, що є складовою задачі покращення нейронного машинного перекладу.

Посилаючись на вище наведені положення, що характеризують сучасний стан вирішення проблеми вдосконалення машинного перекладу та впровадження керованості для перекладів, задачею є: розробка та впровадження високоякісного мульмодального набору даних, який дозволить прискорити досягнення у області машинного перекладу для української мови.

У результаті, більша кількість задач обробки природньої мови зможе бути вирішеною, спираючись на створений набір даних, адже планується впровадження мультимодального набору даних, який міститиме текстову інформацію та зображення, що відповідатиме зібраним анотаціям.

2.4 Етапи проведення експерименту

Сформульована у пункті 2.3 задача є комплексною і вимагає детально спланованого експерименту, аби поступово наблизитися до результативного

вирішення задачі впровадження керованості для машинних перекладів для української мови.

Серед основних етапів необхідно зазначити необхідність:

- проаналізувати існуючи паралельні набори даних та корпуси даних, що містять тексти українською мовою;
- ознайомитися зі способом подачі інформації, методами упорядкування наборів даних; зосередитися на мультимодальних наборах даних;
- визначити тематику створюваного набору даних;
- виконати попередню обробку обраного мультимодального набору даних;
- розбити обраний мультимодальний набір даних на декілька груп, аби прискорити коригування та мати змогу протестувати програмні рішення на вже існуючих даних;
- створити мануальні анотації до обраного набору даних, відкоригувати неточності що міститимуться у наборі даних;
- виконати фінальну обробку набору даних;
- підготувати набір даних до публікації на платформі із відкритими наборами даних для досліджень обробки природньої мови;
- застосувати зібраний мультимодальний набір даних до задачі нейронного машинного перекладу, аби перевірити текстовий вміст набору даних;
- застосувати зібраний мультимодальний набір даних до задачі мультимодального семантичного пошуку;
- оцінити існуючі метрики для визначення якості зібраних даних, що містяться у створеному наборі даних;
- запропонувати власну метрику для оцінки якості перекладів для української мови;
- розробити програмний додаток для коригування та менеджменту мультимодальних наборів даних, що значною мірою прискорить час мануальних виправлень і автоматизує процес створення наборів даних, першочергово для української мови;

- підбити підсумки проведеного експерименту для задачі машинного перекладу;
- підбити підсумки проведеного експерименту для задачі мультимодального семантичного пошуку;
- підбити підсумки проведеного експерименту для метрики.

2.5 Метрики оцінки якості

Метрики для оцінювання якості нейронного машинного перекладу зазвичай покладаються на оцінку подібності між гіпотезою, запропонованою застосовуваною моделлю і створеною або перевіреною людиною, мовою перекладу. Стандартні показники базуються на основних функціях, що знаходяться на лексичному рівні, таких як кількість відповідних n-грам між машинним перекладом і людським оцінюванням.

Існує низка показників [21], що зазвичай можуть бути використані для оцінювання точності моделей машинного перекладу. Нижче наведено найбільш поширені показники:

- метрика BLEU — це стандартний показник для оцінювання якості машинного перекладу. Наведена метрика вимірює перекриття між вихідним значенням моделі та одним чи кількома еталонними людськими перекладами. Оцінка знаходиться в діапазоні від 0 до 1, причому вища оцінка вказує на вищу якість перекладу;
- метрика TER — це показник для оцінки якості машинного перекладу, який вимірює відсоток токенів у вихідних даних моделі, які відрізняються від еталонного перекладу, із врахуванням порядку слів;
- метрика METEOR — це показник, який враховує лексичну коректність вихідних даних моделі. Метрика враховує не лише збіг токенів безпосередньо із перекладом, але й синонімію цих токенів, корені та інші морфологічні характеристики;

- метрика NIST — це показник, який порівнює подібність між вихідними даними моделі та еталонними перекладами за допомогою зваженої суми збігів n-грам;
- людська оцінка – це показник, який не є автоматизованим як наведені вище метрики, проте є найбільш важливим і точним, хоча й потребує досить багато часу для його реалізації. Зазвичай проводиться професійними перекладачами чи носіями мови, які вимірюють результат за шкалою, наприклад, від 1 до 5, на основі загальної якості отриманого перекладу.

Вибір правильної метрики має вирішальне значення для оцінки ефективності моделей нейронного машинного перекладу, незалежно від того, чи є мова, що аналізується, малоресурсною.

Крім цього, метрики можуть бути упередженими [22]. Це може проявлятися щодо певних результатів чи характеристик даних, чи навіть мови перекладу, тому критично важливо обирати показники, що із меншою ймовірністю можуть бути упередженими. Як, до прикладу, точність може бути зміщеною у випадку деяких незбалансованих наборів даних, де один клас представлений набагато чисельніше за інші. У такому випадку оцінка F1 може більш відповідним показником.

Вибір коректного показника гарантує, що є змога порівнювати моделі за різними методами оцінки, наборами даних і їхніми результатами [23]. Це є надзвичайно важливим фактором під час порівняння мультилінгвістичних моделей із обмеженою кількістю ресурсів для навчання, оскільки у цьому випадку може бути неможливо залучити великі набори даних для всеохоплюючого оцінювання.

У цілому, вибір необхідної метрики залежить від поставленого завдання машинного перекладу та вхідної інформації. Метрика BLEU є стандартним та найбільш розповсюдженим показником для задач машинного перекладу, проте є сенс порівнювати декілька показників [24].

Людське оцінювання є важливою складової оцінки ефективності моделей обробки природньої мови. Деякі автоматизовані метрики, як оцінка F1 можуть забезпечити кількісне вимірювання вихідних результатів моделі, проте людська

оцінка є важливою з багатьох причин. Будь-яка мова є достатньо складною системою з певними нюансами, і автоматизовані показники не можуть охопити повний набір значень або лексичних чи граматичних тонкощів мови. Людська оцінка беззаперечно в змозі надати більш детальну, проте суб'єктивну оцінку результативності моделі, особливо для таких завдань, як аналіз емоційної забарвленості текстів, генерації текстів або машинного перекладу.

Моделі обробки природньої мови розробляються з метою вирішення практичних задач, в свою чергу людська оцінка здатна надати розуміння того, наскільки добре модель працює в реальних умовах. Людина може надати відгук про якість, актуальність і логічність створеного тексту, що може допомогти в розробці моделі та покращити її продуктивність у реальних умовах.

Вище згадані метрики використовуються у таких програмних системах як чат-боти або перекладачі. Людська оцінка може надати уявлення про те, наскільки добрі має результати модель у програмних додатках, що вкрай важливо аби проаналізувати можливість практичного застосування додатку.

Хоча автоматизовані методи і можуть допомогти в анотуванні великих обсягів даних, людське судження має важливе значення для забезпечення якості, точності та послідовності анотацій. Люди-анотатори беруть участь у початковій фазі створення наборів даних, де вони вручну позначають або анотують дані на основі конкретних вказівок або критеріїв. Наприклад, у наборі даних аналізу настроїв спеціалісти можуть позначати речення чи тексти як позитивні, негативні чи нейтральні на основі свого судження про висловлені настрої.

Після завершення початкової анотації для забезпечення якості використовується людське судження. Це передбачає перегляд підмножини анотованих даних для виявлення та виправлення будь-яких помилок, невідповідностей або двозначностей. Рецензенти перевіряють, щоб анотації відповідали вказівкам, і за потреби вносять необхідні корективи.

Тож, людська оцінка є невід'ємною частиною оцінки результативності моделей машинного навчання.

3 ОПИС ПРОВЕДЕНИХ ТЕОРЕТИЧНИХ І ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

3.1 Аналіз можливих даних для впровадження керованості

У ході кваліфікаційної роботи було проаналізовано більше 40 існуючих наборів даних, які містять речення українською мовою. Одним із найбільш чисельних є корпус під назвою OPUS.

OPUS — це паралельний корпус [25], що доступний українською та англійською мовами. Набір даних є частиною OPUS (Open Parallel Corpus), який є колекцією вільнодоступних паралельних корпусів кількома мовами.

Українсько-англійський набір даних OPUS включає близько 1,5 мільйона пар речень, причому український текст отримано із різних доменів, таких як інформація з новин, субтитри до фільмів, серіалів та іншого відеоконтенту, художньої літератури, а англійський текст – із відповідних перекладів.

Тексти, що містяться у наборі даних, були попередньо перевірені та автоматично оброблені. Корпус доступний у різних форматах, включаючи TXT, XML і TMX.

Загальні характеристики складових наборів даних OPUS, а саме: розмір файлу, кількість строк, середню кількість слів у реченні англійською та українською, кількість токенів; які містять паралельні текстові дані українською та англійською мовами наведено у таблиці 1.

Таблиця 1– Характеристика корпусу OPUS

№	Назва набору даних	Розмір файлу, b	К-сть строк	Сер. к-сть слів (ан.)	Сер. кі-сть слів (ук.)	Токени (ан), од	Токени (ук), од
1	OPUS-ccmatrix-v1-eng-ukr	1 774 032 207	20240 171	8,925	8,261	756 896 374,67	936 342 6 94,73

Кінець таблиці 1

2	OPUS-elrc_2922-v1-eng-ukr	331 823	2734	19,796	17,251	226 772,29	264 119,71
3	OPUS-eubookshop-v2-eng-ukr	216 706	1790	19,200	14,350	144 001,92	143 844,40
4	OPUS-gnome-v1-eng-ukr	2 890	150	2,846	2,880	1 788,71	2 419,20
5	OPUS-gnome-v1-eng_AU-ukr	408 655	15945	5,590	5,760	373 465,38	514 321,92
6	OPUS-kde4-v2-eng-ukr	11 807 550	233611	25,829	24,849	25 282 202,39	32 507 998,54
7	OPUS-kde4-v2-eng_GB-ukr	7 295 980	178091	13,105	13,414	9 778 967,91	13 377 910,97

Загалом, цей набір даних може бути використаний для навчання та оцінювання точності моделей машинного перекладу, розробки нових сучасних мовних моделей та вивчення лінгвістичних властивостей української та англійської мов, що сприятиме вирішенню задачі керованості.

Ще одним великим набором даних, який містить тексти українською мовою є Ukrainian Paracrawl. Це чисельний паралельний корпус текстів українською та анлійською мовами. Наведена колекція даних є частиною мультимовного набору даних Paracrawl, який можна використовувати для вирішення задач машинного перекладу та інших завданнь із обробки природної мови.

Набір даних включає більше 15 мільйонів пар речень. Український текст отримано з різних джерел, таких як новини, веб-сторінки та юридичні документи.

Загальну характеристику складових наборів даних, а саме розмір файлу, кількість строк, середню кількість слів для української та англійської мов та кількість токенів у наборі Paracrawl для української мови наведено у таблиці 2.

Таблиця 2 – Характеристика набору даних Paracrawl.

№	Назва набору даних	Розмір файлу, b	К-сть строк	Сер. к-сть слів (ан.)	Сер. кі-сть слів (ук.)	Токени (ан), од	Токени (ук), од
1	Paracrawl-1_bonus-eng-ukr	16 428 249 907	23570 0383	5,912	4,921	5 838 600 183,40	6 495 336 874,56

Недоліком цього набору даних є недостатня точність перекладів українською мовою, оскільки він не був перевірений вручну, через що містить значну кількість неточних перекладів, дублювань, пропущених речень і лексичні та граматичні помилки. До того ж, середня довжина речення, що становить 5,912 є достатньо малою, для того аби ефективно навчити модель відтворювати залежність між членами речення у текстах. Це свідчить про велику кількість речень що є односкладними і, фактично, можуть бути замінені словниковими перекладами.

3.2 Огляд набору даних Multi30k

Набір даних Multi30K [26] є модифікацією відомого набору Flickr30K [27] який у свою чергу є еталонним набором даних, який використовується для створення описів до зображень, дослідженнях у галузі машинного перекладу та інших задачах машинного навчання, які передбачають обробку текстової та візуальної інформації. Він включає 31 783 зображення та понад 158 000 анотацій, зібраних із Flickr - популярної платформи для обміну фотографіями. Цей набір даних вперше був представлений у 2014 році і відтоді став одним із

найрозповсюдженіших наборів даних у сфері комп'ютерного зору та обробки природної мови, через мультимодальний тип данних, що він містить.

Кожне зображення у згаданому наборі даних Flickr30k пов'язане з п'ятьма анотаціями англійською мовою, що були зібрані із використанням платформи Amazon Mechanical Turk. Підписи знаходяться у відповідності до зображеннь, також кожен такий підпис є коротким реченням або словосполученням, що відображає візуальний вміст зображення. Представлені зображення в цьому мультимодальному наборі даних достатньо різнопланові та містять різноманітні суб'єкти, сцени, предмети які виконують певний набір дій. До поширених категорій належать люди, домашні тварини, будівлі, спортивні події тощо.

Перевагою набору даних Flickr30k є те, що він є достатньо чисельним і різноманітним, а отже його можна використовувати для тренування та оцінювання складних моделей. Окрім цього, згаданий набір даних нерідко використовується в дослідженнях, пов'язаних із обробкою природньої мови і комп'ютерним зором, що робить його еталоном для визначення ефективності різноманітних алгоритмів і моделей.

У свою чергу Multi30k включає зображення із Flickr30k і лише одну відповідну анотацію, як показано на рисунку 2.

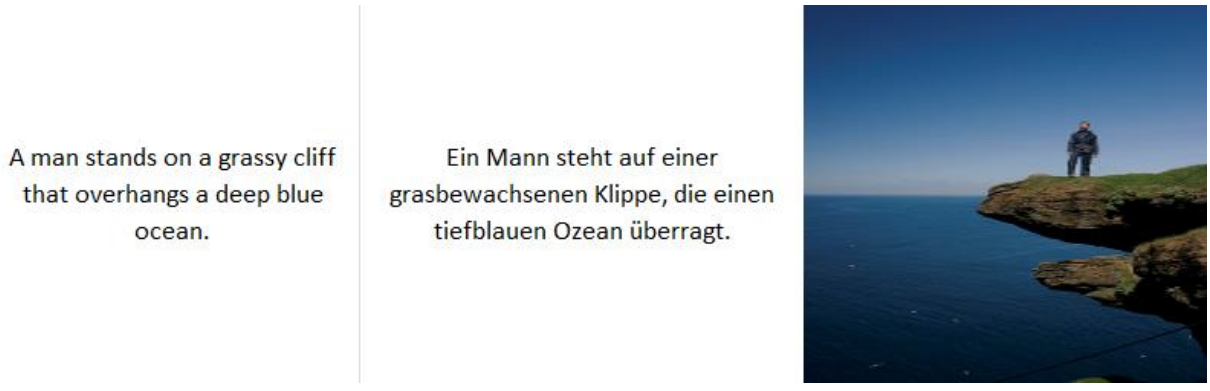

Рисунок 2 – Екземпляр із набору Multi30k

Цей набір даних редагувався професійними перекладачами, а також неодноразово перекладався такими мовами, як французська, чеська, турецька. Це є головною причиною, чому згаданий мультимодальний набір даних є

перспективним, корисним, придатним для подальшої роботи за обсягом, змістом і темами, які він охоплює.

Вище згаданий набір даних Multi30k унікальний тим, що містить підписи кількома мовами, що робить можливим мультилінгвістичні дослідження [28]. Він активно використовується для навчання моделей нейронного машинного перекладу, які можуть перекладати текстові дані з однієї мови на іншу, а також ефективно працювати із завданнями генерації текстів базуючись на зображеннях.

3.3 Підготовка набору даних Multi30k

Для того, аби вирішити задачу керованості та покращити машинний переклад з англійської на українську мови, у якості основного набору даних обрано Multi30k. Вихідна версія містить описи англійською та німецькою мовами. Анотації англійською мовою можна побачити на рисунку 3.

Рисунок 3 – Огляд даних із набору Multi30k

Першим етапом обробки вихідного набору даних, прибираємо з нього анотації німецькою мовою, залишаючи лише описи англійською мовою, які точно описують предмети, кольори, локації та дії.

Наступним кроком, за допомогою сервісу Google Translator було перекладено наявні описи українською мовою з англійської. Необхідно зауважити, що результат, отриманий від сервісу Google Translator, слугував лише основою для наступного мануального перекладу. Такий підхід заощадив час і дозволив вносити зміни за допомогою невеликих правок.

Тож увесь процес підготовки включив завантаження та попередню обробку набору даних Multi30k, вибір відповідних описів англійською мовою та переклад їх українською мовою для створення мультимодального набору даних для машинного перекладу. Екземпляр із набору даних зображено на рисунку 4.

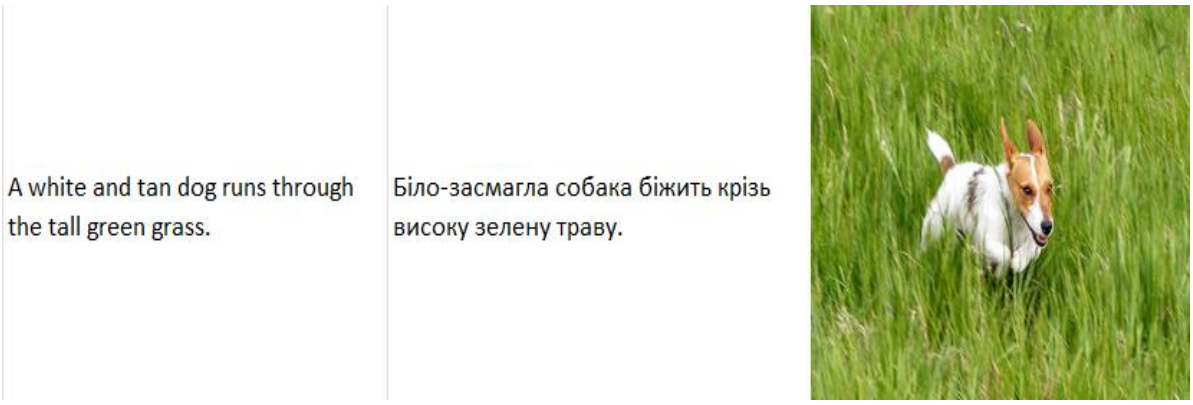

Рисунок 4 – Екземпляр набору даних Multi30k

У процесі автоматичного перекладу дані були розподілені на групи по 2000 речень. Після цього кожна партія піддавалася мануальній перевірці, а саме, виправленню граматичних і лексичних неточностей.

Наступним і найважливішим кроком було вручну виправити дані, що були отримані від сервісу Google Translator. На цьому етапі було знайдено та виправлено лексичні помилки, граматичні помилки та неправильне форматування. Важливо зазначити, що виправлення були зроблені для повної відповідності англійському значенню та зображенню

Приклад виправлення лексичних, граматичних та помилок форматування для речення «A young girl in a red shirt hitting a tennis ball with a pink racket» - «Маленька дівчинка і червоній сорочці б’є тенісний м’яч рожевою ракеткою» наведено на рисунку 5.

Рисунок 5 – Приклад виправлення помилок

Під час ручного перегляду машинних перекладів було виявлено, що автоматичні переклади не завжди є точними, особливо коли мова йде про вирази у переносному значенні. Наприклад, в одному випадку англійський вислів «make a face» було перекладено як «зробити обличчя», хоча точніший український переклад — «кривляється», що показано на рисунку 6.

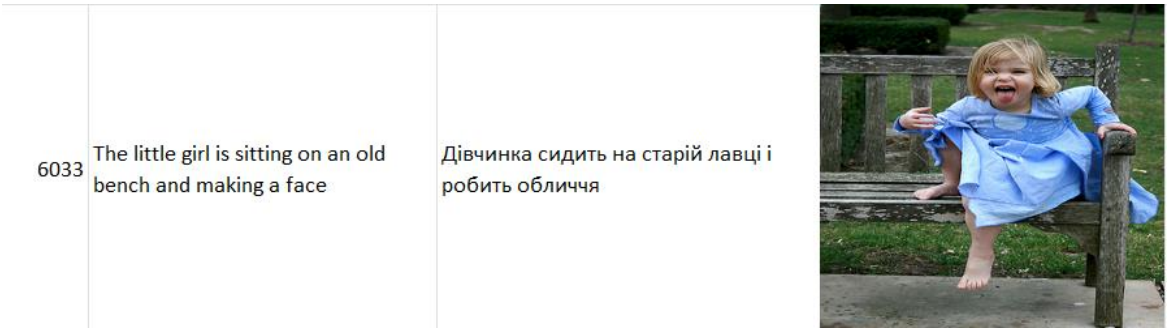

Рисунок 6 – Екземпляр перекладу

Аби покращити якість перекладів, такі помилки було виправлено вручну. Також було видалено всі некоректні символи чи знаки пунктуації.

У результаті було створено якісний і мануально відредагований набір даних англо-українських перекладів, який можна використовувати для подальших досліджень і розробки моделей машинного перекладу для української мови і мультилінгвістичних досліджень в цілому.

Чисельну характеристику набору даних наведено у таблиці 3.

Таблиця 3 – Чисельна характеристика набору даних

Мова анотацій	Кількість речень	Кількість токенів
Англійська	31 014	357 172
Німецька	31 014	333 833
Українська	31 014	276 520

Також було побудовано кластеризацію [29] отриманих даних для кращого ознайомлення з результатами. Для цієї потреби було використано модель MiniLM, а також виконано топічне моделювання. Кластеризацію для англійської мови наведено на рисунку 7. Побудовану кластеризацію для української мови наведено на рисунку 8.

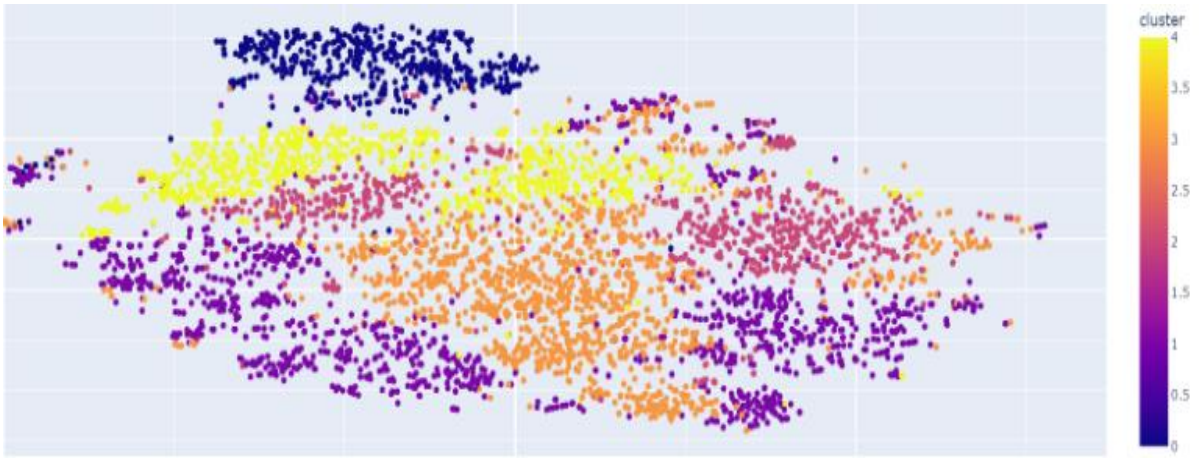

Рисунок 7 - Кластеризація для англійської мови

Іншими словами, отримані дані українською та англійською мовами було розподілено за темами. Загалом речення розподілилися на 384 теми. З цього можна зробити висновок, що українська мова є морфологічно багатою мовою.

Іменники і українській мові відмінюються, тобто змінюють закінчення за відмінком, числом і родом. Усього визначається сім відмінків: називний, родовий, давальний, знахідний, орудний, місцевий і кличний. До того ж, кожен відмінок має

кілька форм, і моделі відмінювання змінюються залежно від роду та групи відмінювання іменника.

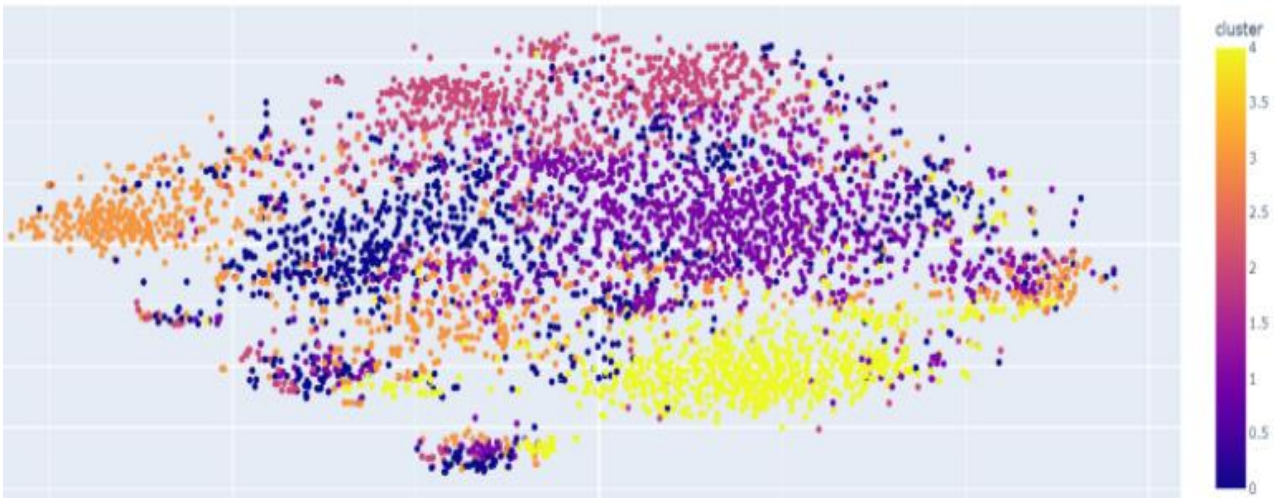

Рисунок 8 – Кластеризація для української мови

Застосована модель машинного навчання MiniLM — це компактна версія популярної мовної моделі BERT, яка розроблена для забезпечення якісної обробки мови зі значно меншою кількістю параметрів. Модель розроблена дослідниками Microsoft і вперше представлена у 2020 році. Порівняно з оригінальною моделлю BERT, MiniLM має лише чверть параметрів, при цьому зберігаючи таку ж саму ефективність у низці завдань обробки природної мови, таких як класифікація тексту, відповіді на запитання тощо [30].

MiniLM досягає цього зменшення розміру за допомогою різноманітних методів, таких як обмін параметрами, дистиляція знань і структуроване скорочення. Модель також включає техніку нормалізації шарів, щоб підвищити ефективність обчислень. Незважаючи на менший розмір, модель MiniLM показує свою високу ефективність у багатьох дослідженнях для ряду завдань обробки мови та має потенціал бути корисним для програмних додатків, які потребують обробки даних природної мови в реальному часі, таких як чат-боти та віртуальні помічники.

Вище було зазначено, що одному слову в англійській мові відповідає декілька слів в українській мові. З цього можна зробити припущення, що наявних даних може бути недостатньо для виконання класичного навчання з

тренувальними та тестовими наборами. Отже, є можливість використовувати підхід One-hot навчання.

One-hot навчання — це підхід у машинному навчанні, який часто використовується в обробці природної мови для представлення текстових даних у числовому форматі, який можна використовувати як вхідні дані для моделі машинного навчання [31].

У такому режимі навчання кожне слово в текстовому наборі даних представлено у вигляді вектора нулів, де одна одиниця вказує на позицію слова в словниковому списку. Такий спосіб надає простий і ефективний підхід до представлення текстових даних для машинного навчання, оскільки він дозволяє моделі розрізняти різні слова та їхню частоту в тексті. Однак він не дає змоги зафіксувати зв’язки між словами чи контекстом, у якому вони використовуються.

One-hot навчання часто використовується в поєднанні з іншими техніками, такими як word2vec, які можуть зафіксувати складніші зв’язки між словами та покращити ефективність моделей обробки природньої мови.

Word2Vec — це популярний алгоритм, який використовується для завдань обробки природної мови, зокрема для створення векторних представлень слів. Отримані значення — це щільні, низьковимірні векторні представлення слів, які фіксують семантичні та синтаксичні зв’язки між словами на основі їхнього контексту [32].

Алгоритм Word2Vec використовує техніку під назвою негативна вибірка, щоб зробити процес навчання ефективнішим. Негативна вибірка передбачає випадковий вибір невеликої кількості неконтекстних слів (негативних вибірок) під час навчання, що допомагає моделі навчитися розрізняти контекстні та неконтекстні слова. Після того, як модель Word2Vec навчена, отримані векторні представлення слів можна використовувати для різних завданьобробки природньої мови. Ці векторні представлення можуть фіксувати семантичні та синтаксичні зв’язки між словами, дозволяючи виконувати такі завдання, як вимірювати схожість слів, шукати аналогії до слів тощо.

4 АНАЛІЗ РЕЗУЛЬАТАТІВ

4.1 Задані умови експерименту

Із урахуванням опису проведених теоретичних і експериментальних досліджень було обрано скористатися методом контрольованого експерименту. Перед цим задаємо певний набір характеристик середовища:

- CPU: Intel Core i5;
- RAM: 8 Гб;
- ОС: Ubuntu 20.04.

Для проведення експерименту було використано середовище Google Colab із вище наведеними характеристиками.

Для експерименту в рамках задачі покращення машинного перекладу було обрано перші 5000 рядків із набору даних Multi30k та доданих вручну анотацій до нього. Для задачі мультимодального семантичного пошуку і впровадження метрики для оцінювання якості отриманих перекладів було використано повний набір даних, тобто 30 014 речень.

4.2 Машинний переклад

Дані, отримані після перекладу, були ретельно проаналізовані. Було вирішено обчислити косинусну подібність за допомогою багатомовної моделі Distiluse-base-multilingual cased-v2 для оригінального перекладу та перекладу за допомогою Google Translator.

Косинусна подібність — це міра подібності між двома векторами n-вимірного простору, яка зазвичай використовується при пошуку інформації чи обробці природньої мови. Косинусна подібність обчислюється як косинус кута між двома векторами, спроектованими в багатовимірному просторі. У контексті обробки природної мови вектори зазвичай створюються з текстових документів або слів, де кожен вимір представляє функцію або частоту термінів.

Було отримано достатньо високе значення косинусної подібності для всіх речень для обох мов. У результаті 4997 речення мають високу оцінку, а 3 — досить низьке значення. У цьому експерименті враховуються значення вище 0,4, що вважається достатнім для загального розуміння змісту речення або словосполучення. Вичерпну порівняльну характеристику можна побачити у таблиці 4.

Таблиця 4 – Порівняння значень косинусної подібності

Значення косинусної подібності	Вихідний набір даних Multi30k	Відкоригований набір даних Multi30k
0,9	1516	1546
0,8	3700	3763
0,7	4616	4675
0,6	4912	4919
0,5	4997	4977
0,4	4999	4999
0,3	5000	5000
0,2	5000	5000
0,1	5000	5000

На прикладі 5000 записів, значення косинусної подібності за допомогою описаної вище моделі, можна відслідкувати ефективність запропонованого методу. У стовпчиках «Вихідний набір даних Multi30k» і «Відкоригований набір даних Multi30k» вказано кількість речень, які перевищують відповідне значення косинусної подібності.

Таким чином, в результаті коригування перекладу, додаткові 30 значень вийшли за межі діапазону 0,9, 63 значення вийшли за межі діапазону 0,8 і так далі. Це відображає результативність виконаної роботи у ході мануальної перевірки обраних даних.

Для мультимодальних задач можливе використання таких моделей як CLIP і StableDiffusion — це дві нові моделі, які стали відомі завдяки високій ефективності в ряді завдань обробки природної мови та комп'ютерного зору.

CLIP — являється розробкою OpenAI, яка використовує контрастний підхід до навчання для попереднього навчання нейронної мережі на чисельному наборі даних із тексту та зображень [33]. Модель передбачає, чи обрана пара зображень та текстів семантично пов'язані чи ні. Описаний метод навчання дозволяє CLIP вивчати складні представлення, які представляють зв'язок між візуальною інформацією та текстом, не вимагаючи великих обсягів розмічених даних. Архітектуру моделі CLIP наведено на рисунку 9.

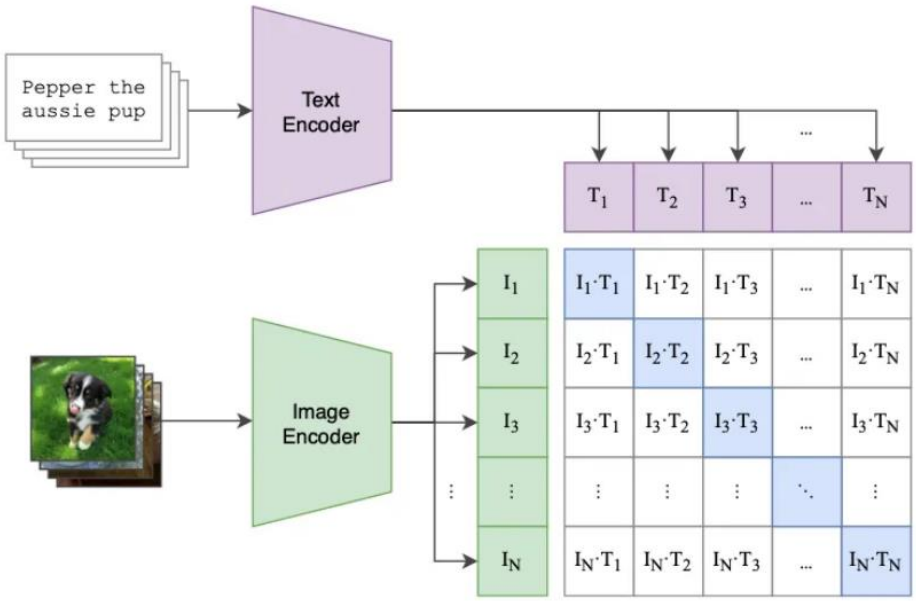

Рисунок 9 – Архітектура моделі CLIP

Модель CLIP відображає ефективність у ряді завдань, таких як класифікація зображень, обробка природної мови та відповіді на запитання, базуючись на візуальній інформації. Однією із необхідних харакетристик CLIP є його здатність добре узагальнювати нові завдання та домени.

Модель StableDiffusion — це ще одне мультизадачне можливе рішення, що застосовується на мультимодальному наборі даних, яке зазвичай демонструє результати в задачах генеративного моделювання, таких як синтез зображень і

генерація текстів до них. StableDiffusion — це генеративна модель, яка використовує принцип дифузії для ітеративного вдосконалення генерації зображень або тексту. Модель ітеративно видаляє шум у вхідному сигналі за допомогою дифузії, зберігаючи високорівневу структуру даних.

Однією з основних переваг моделі StableDiffusion є її здатність генерувати високоякісні та різноманітні текстові та візуальні дані, в той самий час демонструючи ефективність з боку обчислень. Модель має достатньо високі результати в таких задачах, як синтез зображень, генерація тексту та синтез аудіо, тобто на мультимодальних даних.

Для розуміння отриманих результатів на створеному мультимодальному наборі даних Multi30k розглянемо побудовані SHAP візуалізації.

Загалом, SHAP — це структура для розуміння вихідних даних моделей машинного навчання шляхом призначення значень важливості різноманітним вхідним функціям. Значення SHAP надають спосіб пояснити, як вхідні функції впливають на вихідні дані моделі для конкретного вхідного екземпляра. Програмний код для побудови візуалізації наведено на рисунку 10.

```
token = 'general'
text = f"[{token}] {df['input'][INDEX]}"

shap_values = explainer([text], fixed_context=1)
print('ORIGINAL TEXT:', df['input'][INDEX])
print('GROUND TRUTH TRANSLATION:', df['target'][INDEX])
INDEX += 1

shap.plots.text(shap_values)
```

Рисунок 10 – Код для побудови SHAP-візуалізацій

Такі візуалізації є потужним засобом для інтерпретації отриманого вихідного значення для машинного перекладу, а саме дозволяють відслідкувати керованість перекладів. Вони надають графічне представлення значень SHAP, дозволяючи розробникам зрозуміти, як кожна із застосований функція впливає на кінцевий результат моделі.

Одним із найбільш поширених типів візуалізації SHAP є зведений графік, який відображає значення важливості ознак для кожної вхідної функції в усіх екземплярах у наборі даних. Графік також можна відсортувати за важливістю ознак.

На рисунку 11 наведено токен «власності» та токени англійською, які повпливали на його переклад.

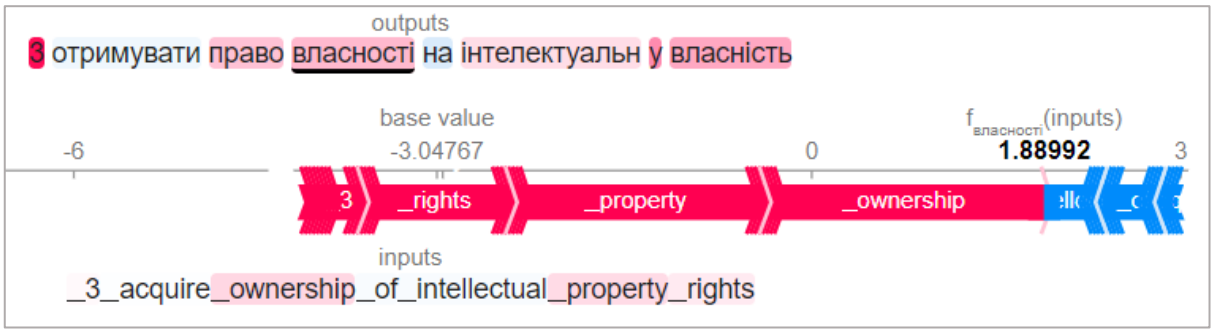

Рисунок 11 – SHAP візуалізація

На рисунку 12 наведено ще один токен «власність» та токен перекладу англійською мовою «property».

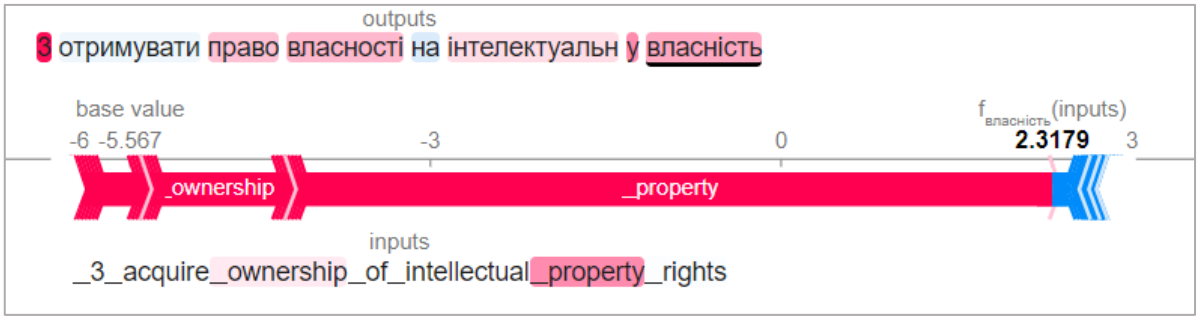

Рисунок 12 – SHAP візуалізація

Ще одним типом візуалізації SHAP є графік залежності, який показує зв’язок між конкретною вхідною ознакою та виходом моделі.

На рисунку 13 наведено приклад токенізації слова «знервовані» за його морфологічними ознаками. Вирішальний вплив на переклад має корінь «нер».

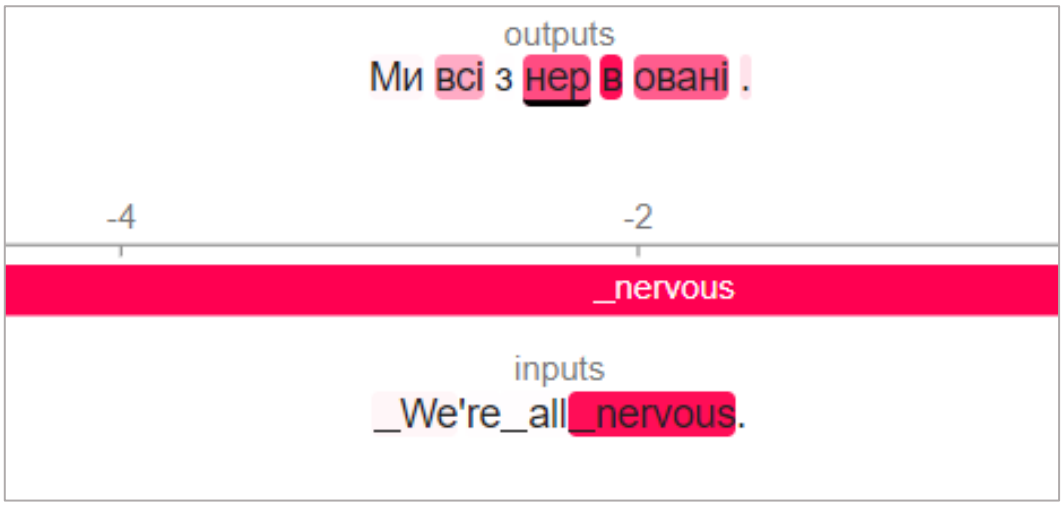

Рисунок 13 – SHAP візуалізація

На рисунку 14 наведено приклад токенізації займенника «те» та відсутність токенів у англійській мові, що мають вплив на це слово, оскільки при перекладі воно не використовується.

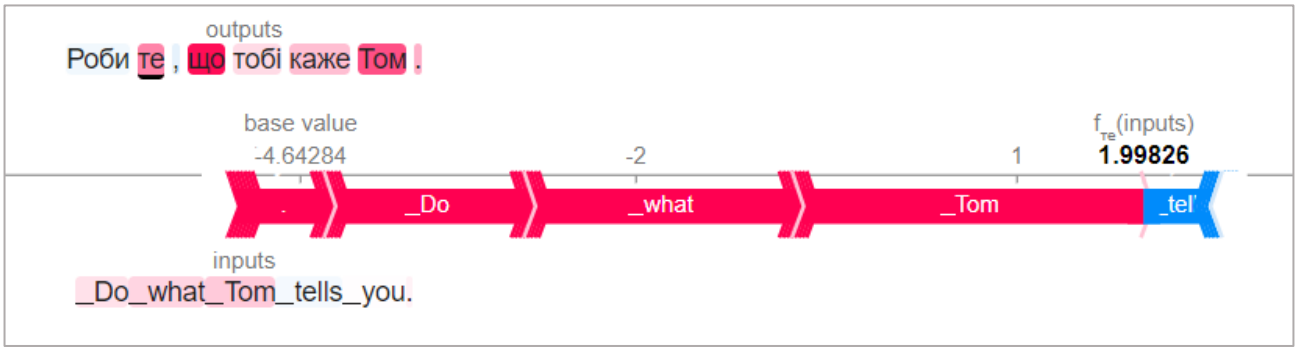

Рисунок 14 – SHAP візуалізація

На рисунку 15 наведено програмний код на мові програмування Python, що дозволяє виконати попередню обробку вхідної інформації, аби застосувати наведені SHAP візуалізції, та відстежити влив кожного конкретного токена на його переклад. Токенізація відбувається за морфологічними ознаками слова. Зазвичай саме коренева морфема виражає основне лексичне значення слова, та є основою лексичного значення слова.

Функція «preprocess_function» приймає наведені задані параметри та токенізує вхідні дані за допомогою пакету для обробки природньої мови.

```
prefix = ""
max_input_length = 128
max_target_length = 128
source_lang = "en"
target_lang = "uk"


def preprocess_function(examples):
    inputs = examples['en']
    targets = examples['fixed_uk']
    model_inputs = tokenizer(inputs, max_length=max_input_length, truncation=True)

    with tokenizer.as_target_tokenizer():
        labels = tokenizer(targets, max_length=max_target_length, truncation=True)

    model_inputs["labels"] = labels["input_ids"]
    return model_inputs
```

Рисунок 15 – програмний код для попереднього опрацювання даних

За допомогою діаграм було відображено значення SHAP для деяких екземплярів в мультимодальному наборі даних Multi30k для англійської та української мов у вигляді діаграми зі значенням вхідної функції на осі X і значенням SHAP на осі Y. Це дозволяє ознайомитися із отриманими результатами із керованістю та покращенням машинних перекладів для української мови.

4.3 Мультимодальний семантичний пошук

Проаналізуємо підготовлений мультимодальний набір даних Multi30k на прикладі задачі мультимодального семантичного пошуку. Це метод пошуку та отримання інформації, який поєднує інформацію з кількох модальностей, таких як текст, зображення, відео та аудіо. Кінцевою метою мультимодального семантичного пошуку є забезпечення користувачів більш точними і відповідними результатами пошуку шляхом використання додаткової інформації, наявної в наявних модальностях інформації.

У широко вживаних пошукових системах користувацький запит зазвичай представляється набором ключових вагомих слів, і тоді пошукова система повертає документи, які містять ці ключові слова. Однак цей метод часто не може охопити

наміри користувача, оскільки він не бере до уваги контекст і змістовне значення запиту. Мультимодальний семантичний пошук, на відміну від вище згаданого методу, прагне зрозуміти запит користувача всеоб'ємлюючим шляхом аналізу вмісту кількох модальностей.

Мультимодальний семантичний пошук імплементує декілька кроків, таких як виділення ознак, мультимодальне злиття та порівняння подібності. Етап виділення функцій передбачає вилучення функцій із кожної наявної модальності.

Наступний етап мультимодального злиття об'єднує ці функції для створення спільного представлення користувацького запиту. Після чого, відбувається етап порівняння подібності, який зіставляє представлення запиту з представленнями проіндексованих документів для отримання найбільш актуальних результатів.

У роботі було використано комбінацію CLIP і Siamese DistilBERT (Sanh et al., 2019) у нашому дослідженні як модель мультимодального семантичного пошуку. Архітектуру використаної мережі можна побачити та рисунку 16.

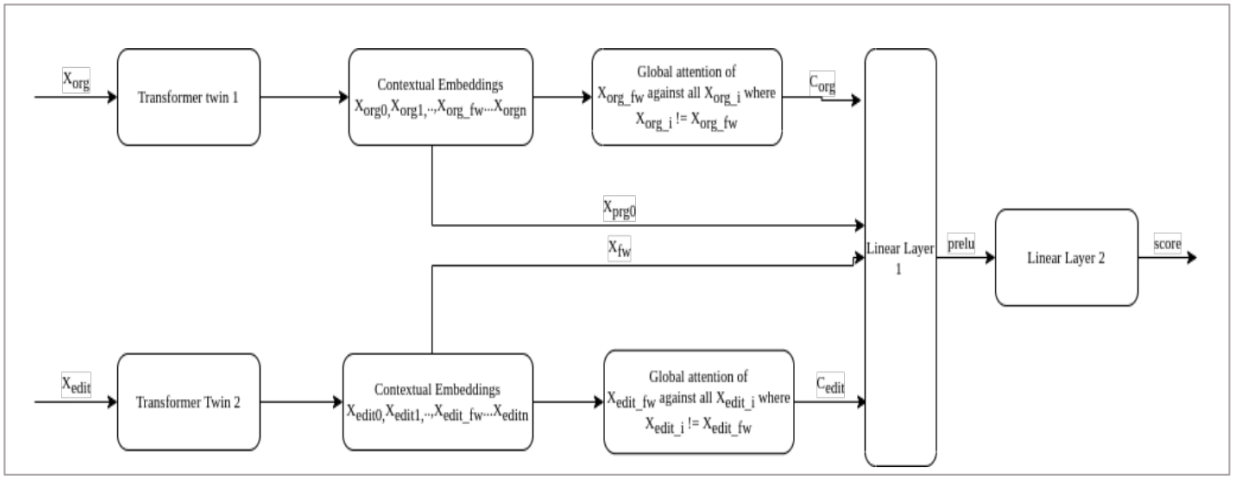

Рисунок 16 – Архітектура сіамських нейронних мереж

Для "clip-ViT-B-32-multilingual-v1" було використані ваги із Huggingface hub як початкова контрольна точка для перевірених моделей. Спочатку необхідно було візуалізувати вбудовані зображення, вихідні англійські тексти та їх виправлення вручну українські переклади.

Моделі повертають вектори, які складаються з 512 елементів зі значеннями від -1 до 1. Перший крок полягав у тому, аби навчити модель таку як оригінальний BERT, для мови з великими ресурсами, як-от англійська.

Тоді DistilBERT має спробувати відтворити цей вектор для перекладу оригінальних текстів максимізувати косинусну подібність між однаковими текстами різними мовами. Застосовується той самий процес до CLIP, щоб відтворити подібний простір для вбудовування відповідні зображення. Тож зображення було закодовано і для текстів обома мовами було використано алгоритм t-SNE для створення 2D проекцій оригінальних векторів вбудовування.

На рисунку 17 можна побачити проекцію, отриманих за допомогою моделі, векторів. Українські переклади майже ідеально повторюють форму англійського тексту. Проекція доводить, що фіксовані переклади мають бути близькими до оригінальних описів і може використовуватися для деяких реальних завдань.

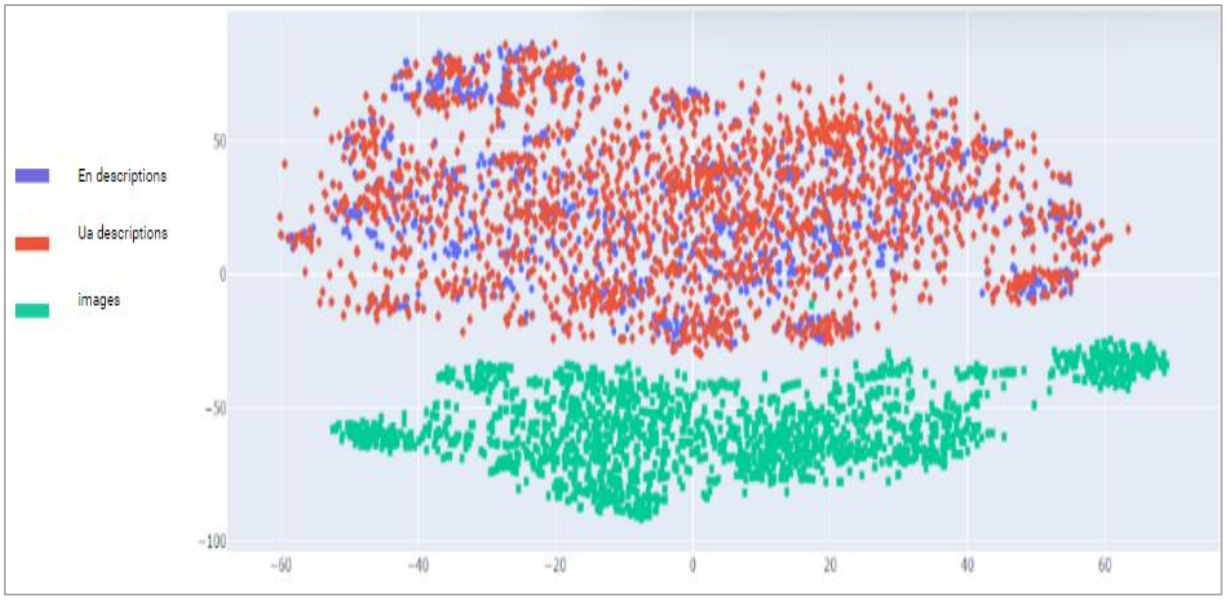

Рисунок 17 – Проекція отриманих даних

Проекцію було створено використовуючи алгоритм t-SNE — це широко застосовувана техніка машинного навчання, яка зазвичай використовується для зменшення розмірності та візуалізації даних великої розмірності. Це критично необхідно для візуалізації складних даних у нижчих вимірах, таких як двовимірні

чи тривимірні, аби отримати коректне уявлення про основні закономірності та зв’язки для даних, що розглядаються. Алгоритм t-SNE реалізується шляхом відображення високовимірних точок даних у низьковимірному просторі таким чином, що відстань між точками в цьому просторі відображає подібність між відповідними високовимірними точками.

Алгоритм використовує ймовірнісний підхід для відображення точок даних, при цьому більша ймовірність призначається найближчим точкам, а менша – віддаленим точкам.

Однак зображення на знаходяться зовсім окремо у цьому просторі векторних представлень. Вони намагаються відтворити форму текстових хмар англійською українською мовами, проти вони знаходяться далеко від текстів і не зовсім відповідають їхнім описам судячи із зображень. У підсумку, 44,5% зображень здебільшого відповідають їх реальному опису. Для української мови це значення дорівнює 29,55%.

Візьмемо для прикладу наступне речення: «Молодий бородатий чоловік у білій безрукавці сидить за барабанною установкою». Модель семантичного пошуку повертає відповідне зображення, але з дещо іншими деталями. На зображенні, що повернулося у якості результату, є насправді людина з бородою, яка сидить, але він малює щось на своєму планшеті, а не грає на барабанах.

Модель дуже добре знаходить згадувані предмети, більшість помилок які було виявлено, пов’язані з дією чи деякими деталями оточення чи фоновими об’єктами. Тестові зображення можуть бути відрізняються деякими дрібними деталями, наприклад кількістю людей у човні, типом локації (печера, гірські породи, певний тип лісу).

Однак алгоритм зазвичай вловлює лише найважливіші деталі, насамперед суб’єкті, що виконують дії. Таким чином, модель пропускає всі ці дрібні деталі, як випадку, що проілюстровано на рисунку 18.

Тож, можна зробити висновок, що для української мови можна застосовувати наведений алгоритм із мульмодального семантичного пошуку, проте із деякими

обмеженнями, оскільки він не все ще допускає деякі некритичні помилки, які можуть бути виправлено шляхом подальшого тонкого налаштування.

Рисунок 18 – Приклад роботи семантичного пошуку

Саме у такому випадку, як наведено вище, запропонований мультимодальний набір даних Multi30k може бути корисним, що сприятиме покращенню показників для української мови для задачі мультимодального семантичного пошуку. Такі комбінації зображень та їх описів можуть використовуватися під час подальших досліджень.

4.4 Метрика для оцінки якості

Кожен класичний підхід до вимірювання якості перекладів залежить від певної правдивості в цільовій мові значення. Однак нерідко трапляються випадки, коли немає такого еталонного тексту. Сучасні багатомовні мовні моделі можуть створювати подібні векторні представлення для тексту кількома мовами.

На рисунку 19 наведено частину програмного коду, що демонструє спосіб обчислення стандартних метрик, що застосовувався вирішуючи попередні

завдання із впровадження керованості для української мови та мультимодального семантичного пошуку.

```
def compute_metrics(eval_preds):
    preds, labels = eval_preds
    if isinstance(preds, tuple):
        preds = preds[0]
    decoded_preds = tokenizer.batch_decode(preds, skip_special_tokens=True)
    # Replace -100 in the labels as we can't decode them.
    labels = np.where(labels != -100, labels, tokenizer.pad_token_id)
    decoded_labels = tokenizer.batch_decode(labels, skip_special_tokens=True)
    # Some simple post-processing
    decoded_preds, decoded_labels = postprocess_text(decoded_preds, decoded_labels)

    bleu_result = bleu_metric.compute(predictions=decoded_preds, references=decoded_labels)
    meteor_result = meteor_metric.compute(predictions=decoded_preds, references=decoded_labels)
    rouge_result = rouge_metric.compute(predictions=decoded_preds, references=decoded_labels)

    print(meteor_result)
    print()
    print(rouge_result)
```

Рисунок 19 - Код для обрахування стандартних метрик

Розглянемо результат стандартних для мультимодального набору даних Multi30k, що наведено на рисунку 20, за п’ять епох тренування.

Epoch	Training Loss	Validation Loss	Bleu	Gen Len
1	No log	2.221202	17.900600	46.632600
2	No log	2.125790	18.869600	47.116100
3	2.475000	2.085327	19.609700	46.948700
4	2.475000	2.068400	19.681600	47.233100
5	1.984400	2.060794	19.592100	47.466300

Рисунок 20 - Результат стандартної метрики Bleu

У ході виконання та аналізу завдання було розраховано косинусну подібність векторів для англійської та української мов, щоб перевірити, як зовнішня модель (сіамський DistilBERT у реалізації distiluse base-multilingual-cased-v2) зможе

оцінити отримані переклади. На рисунку 21 представлена гістограма розподілу значень косинусної подібності.

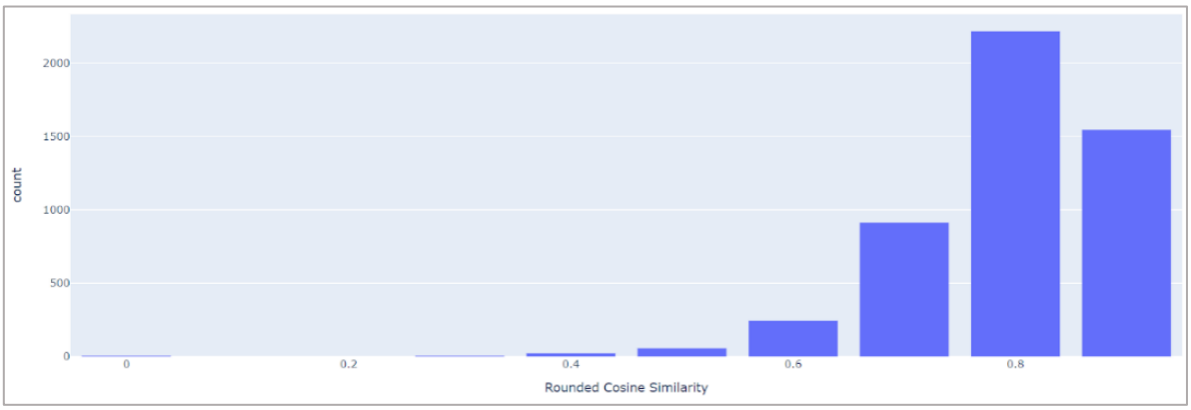

Рисунок 21 – Гістограма розподілу косинусної подібності між векторами

Більшість текстів потрапляє в розподіл 0,6 і вище, що є дійсно прийнятним результатом, оскільки він вказує на те, що отримані переклади вловлюють оригінальний зміст. Загалом, 98,38% текстів знаходиться у цьому проміжку. Такий результат – досягнення для цього показника, оскільки на перший погляд, він майже повторює людське судження. Однак також наявні декілька проміжків, які цікаво дослідити детальніше. Розглянемо переклади, що знаходяться в діапазоні [0,4, 0,6).

Було досліджено тексти, які належать до цих проміжків і здебільшого вони складаються з випадків, коли англійська фраза або словосполучення перекладається як одне слово українською мовою, що також є рідкістю і завжди вживається таким чином. Як, до прикладу, фраза «horse shoes» перекладається на слово "підковки", що є правильним перекладом, але цей переклад не є достатньо поширеним і може дещо неправильно впливати на результат мовної моделі.

Було виявлено ще один схожий приклад: англійська фраза “give high-fives” отримує переклад як «дає п’ять». Переклад, насправді, є коректним, а сама фраза схожа на англійську, але модель її не розпізнає відповідним чином, тому що достеменно не розуміє значення фрази.

Цей випадок наочно показує, що навіть правильний переклад подібного словосполучення в переносному значенні значно знижує оцінку точності моделі.

Тепер ознайомимося із нижчими сегментами, які лежать у діапазоні [0, 0,4). У цьому сегменті присутні лише 3 тексти і всі вони містять тексти з певним сленгом або словосполучення, що транслюють переносне значення. Наприклад іменник «радіоприймач» розмовною англійською мовою перекладається як «walkie talkie».

Українська версія вважає єдиним правильним варіантом лише слово «рація». Це словосполучення отримало оцінку косинусної подібності 0,32, хоча модель не змогла зрозуміти цей розмовний стиль мовлення. Інший приклад містить рідкісне слово «волосінь». Модель не зрозуміла це слово оскільки, не зустрічала його, іншого подібного слово під час проведення етапу навчання.

Крім того, було здійснено спробу зробити те саме, використовуючи siamese DistilBERT із зображеннями, які було використано в завданні із мультимодального семантичного пошуку. Результати відрізняються, оскільки допускається введення візуальної інформації нейронної мережі, аби краще відслідковувати такі фрази, як «дай п'ять» або «рація». Виходить, що додатковий домен зміг надати достатньо контексту для мережі, щоб порівняти ці речення точнішим способом. Речення, яке містить слово «рація» цього разу отримало оцінку 0,7456. Також, відсутні переклади із оцінкою нижче 0,5. На рисунку 22 зображено побудовану гістограму.

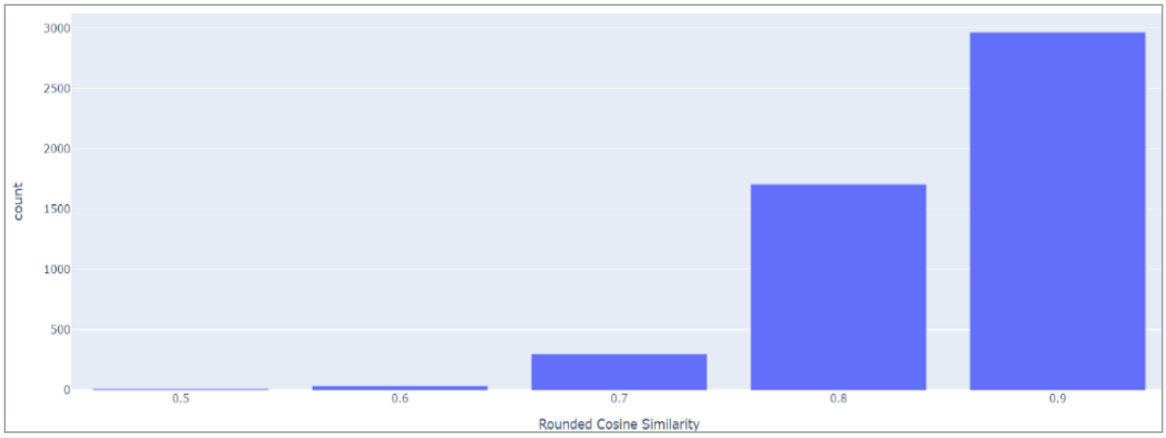

Рисунок 22 – Гістограма розподілу косинусної подібності між векторами

Ця сфера потребує подальших досліджень, але проведені випробування та експерименти демонструють, що такі моделі можна використовувати й надалі, щоб

відобразити деякі фрази або сленги в поєднанні з деякими загальновживаними метриками, наприклад, на основі токенів.

Використання мультидоменної моделі для вимірювання якості перекладу також має великий потенціал, оскільки її результати виявилися значно кращими, ніж у перевіреної текстової моделі. Такий підхід вирішив основні проблеми, з якими із реченнями, що знаходяться у непрямому значенні.

Було виконано декілька додаткових перевірок із деякими випадковими фразами. Англійською мовою такі речення звучать як «Murder will out» і «Keep the change». Українські відповідники мають вигляд: "Правди не сховаєш", "Здачі не треба". Текстова модель показала такі оцінки відповідно: 0,2495 і 0,2599.

Текстова модель, налаштована так, щоб розуміти речення, що знаходяться у непрямому значеннями, є фразеологізмами або сленговими виразами, дала такі бали: 0,9569 і 0,9497. Результати значно перевершують попередні показники.

Однак, як вже було зазначено раніше, теорія про те, що додавання візуальної модальності для оцінки якості машинного перекладу на мультимодальних даних може дійсно наблизитися до успішного вирішення цієї задачі, що було продемонстровано під час проведеного експерименту.

Оцінювання слів і словосполучень у непрямому значенні може бути складною задачею, оскільки вимагає глибокого розуміння контексту та наміру мовця. Непряму мову зазвичай використовують для передачі значення натяком або непрямим способом, частіше за все покладаючись на культурні чи ситуативні знання аби тлумачити правильно сенс.

У непрямому мовленні значення фрази може бути не відразу очевидним, і для її розуміння може знадобитися додатковий аналіз тону мовця, та контексту, у якому була використана така фраза. Через цю особливість може бути занадто складно оцінити вхідні дані, слова та словосполучення в непрямому значенні, оскільки не завжди однозначно зрозуміло, що мовець намагається транслювати. Ще одна суттєва проблема з оцінюванням мови у переносному значенні полягає в тому, що її можна неправильно тлумачити.

ВИСНОВКИ

У ході кваліфікаційної роботи було виконано ряд експериментальних і теоретичних задач на тему «Дослідження керованості англійсько-українського машинного перекладу на основі спеціалізованих корпусів. Набори даних». Результатом роботи над удосконаленням машинного перекладу для української мови на основі зібраного вручну набору даних є опрацювання теоретичної інформації, наукових досліджень та сучасних підходів до вирішення задачі.

Створений мультимодальний набір даних Multi30k був визнаний високоякісним і репрезентативним для української мови та завдань машинного перекладу, які були поставлені. Він включає різні модальності даних, як текстова та візуальна інформація.

Можна зробити висновок, що робота над керованістю машинного перекладу для української мови на основі зібраного вручну набору даних була успішною та продемонструвала значущість високоякісних наборів даних для вдосконалення моделей машинного навчання для мов із невеликою кількістю наявних наборів даних, як українська мова.

Окрім цього, моделі машинного навчання, навчені на цьому наборі даних, показали значне покращення точності перекладу порівняно із іншими наборами даних. Тож задача машинного перекладу, розроблена з використанням цього набору даних.

Ще одним із результатів роботи стало створення вирішення задач із мультимодального семантичного пошуку та впровадження власної метрики яка не потребує еталонного тексту у якості порівняння. Ці задачі були опрацьовані за допомогою створеного набору даних Multi30k.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Bengio Y., Senecal J. S. Adaptive Importance Sampling to Accelerate Training of a Neural Probabilistic Language Model. *IEEE Transactions on Neural Networks*. 2008. Т. 19, № 4. С. 713–722. URL: https://doi.org/10.1109/tnn.2007.912312.

2. Niu X., Carpuat M. Controlling Neural Machine Translation Formality with Synthetic Supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020. Т. 34, № 05. С. 8568–8575. URL: https://doi.org/10.1609/aaai.v34i05.6379 (дата звернення: 11.04.2023).

3. Imageability- and Length-Controllable Image Captioning / M. A. Kastner та ін. *IEEE Access*. 2021. Т. 9. С. 162951–162961. URL: https://doi.org/10.1109/access.2021.3131393 (дата звернення: 01.04.2023).

4. Multimodal Sarcasm Detection: A Deep Learning Approach / S. K. Bharti та ін. *Wireless Communications and Mobile Computing*. 2022. Т. 2022. С. 1–10. URL: https://doi.org/10.1155/2022/1653696 (дата звернення: 11.05.2023).

5. Unsupervised Cross-lingual Representation Learning at Scale [Електронний ресурс] / [А. Конно, К. Ханделвал, Н. Гоял та ін.]. – 2020. – Режим доступу до ресурсу: https://arxiv.org/abs/1911.02116.

6. Daniiel Y., Turuta O. Collection of questionnaire results, received by using the visual analog scale method, for its further processing in the medical web application. *ScienceRise*. 2017. Т. 5, № 2. С. 27–30. URL: https://doi.org/10.15587/2313-8416.2017.102296 (дата звернення: 11.05.2023).

7. Evaluation and Analysis of the NLP Model Zoo for Ukrainian Text Classification [Електронний ресурс] / [Д. Панченко, Д. Максименко, О. Турута та ін.] // ICTERI 2022. – 2022. – Режим доступу до ресурсу: https://link.springer.com/chapter/10.1007/978-3-031-20834-8_6.

8. Multilingual Neural Machine Translation for Low-Resource Languages / S. M. Lakew та ін. *Italian Journal of Computational Linguistics*. 2018. Т. 4, № 1. С. 11–25. URL: https://doi.org/10.4000/ijcol.531 (дата звернення: 11.05.2023).

9. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis [Електронний ресурс] / [Є. Джіа, Ю. Джанг, Р. Вейс та ін.] // 32nd Conference on Neural Information Processing Systems. – 2019. – Режим доступу до ресурсу: https://arxiv.org/pdf/1806.04558.pdf.

10. Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning / E. Erdem та ін. Journal of Artificial Intelligence Research. 2022. Т. 73. С. 1131–1207. URL: https://doi.org/10.1613/jair.1.12918 (дата звернення: 11.05.2023).

11. Neural Machine Translation for Low-Resource Languages: A Survey / S. Ranathunga та ін. *ACM Computing Surveys*. 2022. URL: https://doi.org/10.1145/3567592 (дата звернення: 11.05.2023).

12. A. Yerokhin, A. Babii and O. Turuta, "Geoscience Laser Altimeter System sparse ICESat data processing based on F-transform," 2019 IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL), Sozopol, Bulgaria, 2019, pp. 553-556, doi: 10.1109/CAOL46282.2019.9019463.

13. Improving Language Understanding by Generative Pre-Training [Електронний ресурс] / А.Редфорд, Н. Картік, Т. Саліманс, І. Сутскевер. – 2018. – Режим доступу до ресурсу: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

14. Leveraging Rule-Based Machine Translation Knowledge for Under-Resourced Neural Machine Translation Models [Електронний ресурс] / [Д. Торрегроса, Н. Пасріча, Б. Раджа та ін.] // MT Summit XVII, volume 2. – 2019. – Режим доступу до ресурсу: https://aclanthology.org/W19-6725.pdf.

15. Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease / G. N. Ahamad та ін. *Processes*. 2023. Т. 11, № 3. С. 734. URL: https://doi.org/10.3390/pr11030734 (дата звернення: 11.05.2023).

16. Dmytro Dashenkov, Kirill Smelyakov, Oleksii Turuta, "Methods of Multilanguage Question Answering", 2021 IEEE 8th International Conference on

Problems of Infocommunications, Science and Technology (PIC S&T), pp.251-255, 2021.

17. Mezzoudj F., Benyettou A. An empirical study of statistical language models: n-gram language models vs. neural network language models. *International Journal of Innovative Computing and Applications*. 2018. Т. 9, № 4. С. 189. URL: https://doi.org/10.1504/ijica.2018.095762 (дата звернення: 11.05.2023).

18. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation [Електронний ресурс] / [Й. Ву, М. Щустер, Ж. Чен та ін.]. – 2016. – Режим доступу до ресурсу: https://arxiv.org/pdf/1609.08144.

19. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [Електронний ресурс] / [К. Раффел, Н. Шазір, А. Робертс та ін.]. – 2019. – Режим доступу до ресурсу: https://arxiv.org/pdf/1910.10683.

20. ZENNAKI O., SEMMAR N., BESACIER L. A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages. *Natural Language Engineering*. 2018. Т. 25, № 1. С. 43–67. URL: https://doi.org/10.1017/s1351324918000293 (дата звернення: 11.05.2023).

21. Agreement on Target-Bidirectional Recurrent Neural Networks for Sequence-to-Sequence Learning / L. Liu та ін. *Journal of Artificial Intelligence Research*. 2020. Т. 67. С. 581–606. URL: https://doi.org/10.1613/jair.1.12008 (дата звернення: 11.05.2023).

22. Clegg B. A., DiGirolamo G. J., Keele S. W. Sequence learning. *Trends in Cognitive Sciences*. 1998. Т. 2, № 8. С. 275–281. URL: https://doi.org/10.1016/s1364-6613(98)01202-9 (дата звернення: 11.05.2023).

23. F-transform 3D Point Cloud Filtering Algorithm / A. Yerokhin та ін. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), м. Lviv, 21–25 серп. 2018 р. 2018. URL: https://doi.org/10.1109/dsmp.2018.8478581 (дата звернення: 11.05.2023).

24. Multi-level Distillation of Semantic Knowledge for Pre-training Multilingual Language Model [Електронний ресурс] / [М. Лі, Ф. Дінг, Д. Жан та ін.]. – 2022. – Режим доступу до ресурсу: https://arxiv.org/pdf/2211.01200.

25. Bleu: a Method for Automatic Evaluation of Machine Translation [Електронний ресурс] / К.Папіені, С. Рукос, Т. Ворд, В. Жу. – 2002. – Режим доступу до ресурсу: https://aclanthology.org/P02-1040.pdf.

26. King M. Evaluating natural language processing systems. *Communications of the ACM*. 1996. Т. 39, № 1. С. 73–79. URL: https://doi.org/10.1145/234173.234208 (дата звернення: 11.05.2023).

27. Lapata M., Keller F. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*. 2005. Т. 2, № 1. С. 3. URL: https://doi.org/10.1145/1075389.1075392 (дата звернення: 11.05.2023).

28. Takahashi S., Tanaka-Ishii K. Evaluating Computational Language Models with Scaling Properties of Natural Language. *Computational Linguistics*. 2019. Т. 45, № 3. С. 481–513. URL: https://doi.org/10.1162/coli_a_00355 (дата звернення: 11.05.2023).

29. OPUS – an open source parallel corpus. URL: https://opus.nlpl.eu/ (дата звернення: 24.03.2023).

30. V. Golian, N. Golian, I. Afanasieva, K. Halchenko, K. Onyshchenko and Z. Dudar, "Study of Methods for Determining Types and Measuring of Agricultural Crops due to Satellite Images," 2022 XXXII International Scientific Symposium Metrology and Metrology Assurance (MMA), Sozopol, Bulgaria, 2022, pp. 1-8, doi: 10.1109/MMA55579.2022.9992568.

31. Pre-trained Word Embedding and Language Model Improve Multimodal Machine Translation: A Case Study in Multi30K / T. Hirasawa та ін. *IEEE Access*. 2022. С. 1. URL: https://doi.org/10.1109/access.2022.3185243 (дата звернення: 11.05.2023).

32. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models / B. A. Plummer та ін. *International Journal of Computer Vision*. 2016. Т. 123, № 1. С. 74–93. URL: https://doi.org/10.1007/s11263-016-0965-7 (дата звернення: 11.05.2023).

33. Improving the Machine Translation Model in Specific Domains for Ukrainian Language / [Д. Максименко, Н. Сайчишина, О. Турута та ін.]. // CSIT 2022. – 2022.