УДК: 004.77

Радзіховська Анастасія Олегівна
(наук. керівник – д-р фіз.-мат. наук, професор Ніколюк П. К.)
Донецький національний університет імені Василя Стуса, м. Вінниця

ДОСЛІДЖЕННЯ АЛГОРИТМІВ ТА МЕТОДІВ АНАЛІЗУ МОВНИХ КОНСТРУКЦІЙ У КОМП'ЮТЕРНИХ СИСТЕМАХ ОБРОБКИ ПРИРОДНОЇ МОВИ. РОЗРОБКА НЕЙРОМЕРЕЖЕВОГО ТРАНСЛЯТОРА

У сучасному світі розвиток технологій обробки природної мови (Natural Language Processing, NLP) стає все більш важливим завданням для розвитку інформаційного суспільства. Дослідження алгоритмів та методів аналізу мовних конструкцій у комп'ютерних системах обробки природної мови відіграє ключову роль у подальшому розвитку цієї галузі.

Однією з основних проблем у обробці природної мови є складність самої мови та її структури. Мова людини є динамічною та багатозначною, оскільки містить багато варіантів виразності та контекстуальних нюансів. Тому виникає потреба в розробці алгоритмів, які здатні адаптуватися до цієї складності та ефективно аналізувати текстові дані.

Дослідники в галузі обробки природної мови активно вивчають та розробляють методи для розпізнавання й аналізу мовних конструкцій. Одним із ключових напрямів є застосування машинного навчання та штучних нейронних мереж. Ці методи дають змогу комп'ютерним системам навчатися на великих обсягах текстових даних та вдосконалювати свої навички аналізу мови.

Автоматичний переклад – це завдання переведення тексту з однієї мови на іншу без участі людини. Ця задача вимагає розуміння граматики, синтаксису та семантики обох мов.

Традиційні методи автоматичного перекладу ґрунтуються на правилах, які описують, як перекладати слова та фрази з однієї мови на іншу. Ці правила розробляються вручну лінгвістами, що робить процес трудомістким і дорогим.

Нейромережеві транслятори – це нове покоління систем автоматичного перекладу, які використовують штучний інтелект для навчання перекладу. Ці транслятори навчаються на великих наборах даних паралельних текстів, що дає їм змогу генерувати якісні переклади, які враховують контекст і нюанси мови. Тому в цій роботі пропонується розглянути розробку нейромережевого транслятора.

Нейромережевий транслятор (НМТ) – це система штучного інтелекту, яка використовує нейронні мережі для перекладу тексту з однієї мови на іншу [1]. НМТ мають низку переваг перед традиційними системами машинного перекладу:

- краща точність (НМТ можуть краще розуміти контекст тексту та генерувати більш точні переклади);
- більш природний переклад (НМТ можуть генерувати переклади, які звучать більш природно і менш штучно);
- краща адаптація до нових даних (НМТ можуть краще адаптуватися до нових даних, що робить їх більш стійкими до змін у мові).

Моделі глибокого навчання досягли точності людського рівня в багатьох завданнях. Ці моделі здатні відображати вхідні та вихідні дані з чудовою точністю та з меншими зусиллями. Але одна з проблем полягає в тому, щоб досягти точності на рівні людини, щоб зіставити одну послідовність з іншою. Зазвичай це спостерігається під час мовного перекладу або перекладу мовлення, і це відомо як машинний переклад [2].

Для розробки нейромережевого транслятора використовується метод глибокого навчання, зокрема з використанням рекурентних нейронних мереж для обробки послідовностей даних, що дають змогу ефективно моделювати та враховувати залежності між послідовними елементами тексту. Для тренування нейромережевої моделі використовуються набори даних, що містять пари фраз на двох мовах, у даному випадку – португальські та англійські. Ці дані істотні для ефективного навчання та оцінки якості перекладу.

Код нейромережевого транслятора реалізовано з використанням мови програмування Python та доступний для огляду на платформі GitHub [3]. Цей код складається з різних функціональних модулів, кожен із яких відповідає за певний аспект обробки тексту або реалізацію алгоритму машинного навчання. Важливо зазначити, що реалізація використовує сучасні бібліотеки та інструменти для роботи з нейронними мережами та обробки даних.

Аналізуючи структуру коду, можна виділити кілька основних етапів, кожен із яких відповідає за певну частину процесу перекладу. Зокрема:

1. Зчитування та підготовка даних:

● Код використовує бібліотеку Pandas для зчитування даних із файла "por.txt", який, здається, містить текст для тренування моделі перекладу.

● Зчитані дані розділяються за допомогою табуляції та зберігаються у форматі DataFrame, де перша колонка містить вхідний текст, друга – цільовий текст, а третя – інші дані (які, судячи з коду, не використовуються).

● Виводяться перші декілька рядків зчитаних даних, щоб переконатися, що дані зчиталися правильно.

2. Підготовка даних для навчання моделі:

● Визначаються розмір пакета (batch_size), розмірність латентного простору (latent_dim) та кількість вибірок для тренування (num_samples).

● Після зчитування даних створюються списки input_texts та target_texts для зберігання вхідних і цільових текстів відповідно.

● Унікальні символи вводу та виводу зберігаються у множинах input_characters та target_characters відповідно.

3. Побудова моделі кодера-декодера з використанням TensorFlow:

● Використовується бібліотека TensorFlow для побудови моделі кодера-декодера.

● Визначаються вхідні та вихідні дані для кодера та декодера.

● Використовуються шари GRU для моделі кодера та декодера.

● Функція активації softmax використовується на виході декодера для отримання розподілу ймовірностей символів.

4. Тренування моделі:

● Використовується функція втрати categorical_crossentropy та оптимізатор rmsprop для компіляції моделі.

● Модель тренується на тренувальних даних, розділених на пакети.

● Кількість епох тренування та розмір валідаційного набору визначаються параметрами.

5. Декодування тестових послідовностей:

● Використовується навчена модель для декодування вхідних послідовностей.

● Після закінчення навчання моделі виконується декодування 20 тестових послідовностей.

● Для декодування використовується алгоритм із використанням вихідних станів кодера для ініціалізації станів декодера та використанням власних передбачень моделі як вхідних даних для наступних кроків декодування.

● Декодовані послідовності виводяться на екран разом з оригінальними вхідними послідовностями для порівняння.

Цей алгоритм створює повний конвеєр для навчання та використання моделі перекладу. Кожен з етапів виконує свою функцію для підготовки та використання даних у моделі.

Розроблений нейромережевий транслятор здатен генерувати якісні переклади текстів із португальської мови на англійську. Транслятор враховує контекст і нюанси мови, що робить переклади більш точними й природними. Далі можливі дослідження включають розширення функціональності та оптимізацію продуктивності нейромережевого транслятора.

Список використаних джерел

1. Як відбувається НМТ*? REPORTER. Information portal*. URL: https://reporter.zp.ua/yak-vidbuvayetsya-nmt.html (date of access: 24.03.2024).

2. How to Implement Seq2seq Model. *cnvrg.io.cnvrg*. URL: https://cnvrg.io/seq2seq-model/ (date of access: 24.03.2024).

3. GitHub – AnastasiaRadzikhovska/Translator: *Translation based on a machine learning. GitHub.* URL: https://github.com/AnastasiaRadzikhovska/Translator (date of access: 24.03.2024).

Анотація. У цій роботі було розглянуто розробку нейромережевого транслятора, який є перспективним напрямом у вдосконаленні інструментів розробки програмного забезпечення. Він може покращити якість і продуктивність перекладу, що сприятиме швидкому розвитку індустрії програмування.

Ключові слова: нейромережевий транслятор, Natural Language Processing, машинне навчання, штучні нейронні мережі.