DOI 10.31891/2307-5732-2023-327-5-200-203
УДК 637.5.02

ЦАП ВЛАДИСЛАВ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-8062-0079
e-mail: vladyslav.b.tsap@lpnu.ua
БРУСЕНЦОВ ГЕОРГІЙ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-3346-0164
e-mail: heorhii.y.brusentsov@lpnu.ua

СИСТЕМА ДЛЯ АНОТУВАННЯ ТЕКСТУ УКРАЇНСЬКОЮ МОВОЮ

Ця наукова робота присвячена розробці ефективної системи, призначеної для анотування текстів українською мовою. Основна мета полягає у створенні системи, здатної аналізувати введені українські тексти, виокремлювати ключові поняття та генерувати інформативні анотації. Дослідження зосереджене на області анотування текстів українською мовою, вивчаючи два різні підходи: один передбачає інтеграцію моделі Pegasus з перекладачем, а інший використовує модель mT5, спеціально адаптовану для завдань анотування українською мовою. У дослідженні здійснено комплексне оцінювання цих підходів з особливим акцентом на їхню швидкість роботи та показники ефективності. Воно підкреслює специфіку роботи спеціалізованих моделей для ефективної обробки деталей української мови. У статті також підкреслюється важливість суб'єктивних оцінок для визначення ефективності системи у передачі основних ідей вихідного тексту. Таким чином, ця наукова стаття робить свій внесок у розвиток української системи обробки природної мови, пропонуючи нові підходи до анотування текстів. Вона підкреслює виклики та можливості в цій галузі, наголошуючи на важливості спеціалізованих моделей та суб'єктивних оцінок для досягнення точного та контекстуально релевантного анотування тексту українською мовою.

Ключові слова: узагальнення тексту, обробка тексту українською, обробка природної мови

TSAP VLADYSLAV, BRUSENTSOV HEORHII
Lviv Polytechnic National University

SYSTEM FOR ANNOTATING TEXT IN UKRAINIAN

This research paper delves into the realm of text annotation within the Ukrainian language, aiming to devise a highly efficient system that can adeptly process Ukrainian texts, extracting salient concepts and crafting informative annotations. Within this context, our study predominantly explores two distinct approaches: one revolves around the fusion of the Pegasus model with a translation mechanism, while the other leverages the mT5 model, fine-tuned to cater specifically to the intricacies of Ukrainian annotation tasks.

The primary objective of this research is to evaluate these approaches in depth, with a particular emphasis on their speed and performance metrics. Speed is of paramount importance in the era of real-time data processing, especially when dealing with large volumes of text.

In the age of NLP (Natural Language Processing), language-specific models are instrumental in ensuring the robust handling of linguistic nuances. Ukrainian, with its unique phonology, grammar, and vocabulary, presents its own set of challenges. It is imperative, therefore, that models tailored to the specifics of the Ukrainian language are developed and integrated into annotation systems.

Furthermore, we emphasize the significance of subjective evaluations in this research. While quantitative metrics provide a valuable foundation for measuring system performance, subjective assessments offer a more holistic view. Human evaluation helps in gauging the system's ability to capture the essence of a text, convey its main ideas, and maintain the contextual relevance of annotations. These subjective evaluations serve as a bridge between machine-driven performance metrics and the actual utility of the system in real-world applications.

In conclusion, this research article not only advances the field of Ukrainian natural language processing but also offers novel methodologies for text annotation. It sheds light on the challenges and opportunities that arise when working with a language as intricate as Ukrainian.

Keywords: text summarization, Ukrainian text processing, natural language processing.

Вступ

Сучасний світ насичений екстенсивною масою інформації, яка безперервно росте, і призводить до виникнення феномену, відомого як "інформаційний вибух." Цей феномен відзначається масштабним накопиченням інформації з плином часу. Додатково, ми пережили такий великий прогрес у сфері Штучного Інтелекту, що не вся інформація, що присутня в Інтернеті, більше створюється виключно людьми. Безліч інструментів тепер може не тільки допомагати створювати контент, але й автоматично генерувати його. Таким чином, виникає необхідність у засобах, що дозволять зменшити надлишковість інформації, виділити основну суть і, можливо, навіть ранжувати важливість інформації, перед тим як представити її для сприйняття людиною. В ідеальному варіанті, цю функцію повинні були б виконувати заголовки, але в сучасному світі клікбейти стали поширеним явищем, тому цей метод більше не ефективний.

Метою даного дослідження є розгляд проблеми узагальнення тексту та застосування наявних алгоритмів до текстів українською мовою. Окрім цього, в рамках роботи передбачено розробку базової моделі для вирішення даної завдання. Очікуваним результатом проведеного дослідження є:

- Покращення опрацювання текстів в існуючих мовних моделях
- Опрацювання української мови великими мовними моделями.

Аналіз літературних джерел

У першій статті [1] автори заглиблюються у тонкощі узагальнення тексту, використовуючи можливості попередньо навчених кодерів. Ця робота є ключовою в контексті розробки нашої власної системи, оскільки

ми також плануємо закласти основу нашої архітектури саме на цьому підході. Основний акцент у цій статті зроблено на використанні можливостей існуючих попередньо навчених кодерів.

Друга стаття [2] також присвячена складному завданню узагальнення тексту, але її основна увага зосереджена на ранжуванні речень, яке слугує визначником їхньої важливості. У цьому підході автори утримуються від перефразування речень, а натомість визначають їхню важливість, зберігаючи їх у первісному вигляді. Ця методика має свої переваги та недоліки, і її всебічне вивчення може виявитися дуже корисним.

Третя робота [3] ретельно розбирає фундаментальну теорію, що лежить в основі узагальнення тексту, одночасно надаючи основні принципи роботи з обробкою природної мови (Natural Language Processing, NLP). Помітним аспектом цієї роботи є акцент, зроблений на семантичному аналізі. Примітно, що цей підхід продемонстрував високу ефективність у тестових сценаріях, що робить інтригуючим розгляд його включення в нашу власну розробку системи.

Стаття [4] заглиблюється в тонкощі обробки англійської та української мов, спираючись на семантику та синтаксис. Ця інформація має значну цінність, оскільки всі попередні дослідження в галузі узагальнення текстів та обробки природної мови переважно базуються на англійській мові. Враховуючи, що наша система призначена для роботи з українською мовою, розуміння нюансів обробки обох мов є важливим аспектом нашої роботи.

П’ята стаття [5], яку ми розглянемо, присвячена комплексному методу виявлення кореферентних пар в україномовних текстах на основі нейронної мережі BiLSTM. Ця робота вирізняється своєю важливістю, оскільки в ній пропонується новий інтегрований метод ідентифікації кореферентних кластерів, що має велике значення в сфері обробки природної мови. Особливо цінною є той факт, що аналіз виконується для української мови, що робить цю роботу важливою для нашого контексту.

Шоста стаття [6] впроваджує ідею створення узагальненого тексту за допомогою гібридного підходу, який об'єднує абстрактний та екстрактивний методи узагальнення. Основною пропозицією є створення нової гібридної моделі, яка використовує вбудовування слів BERT та методи навчання з підкріпленням. Це цікавий підхід, і варто детально розглянути результати, які він надає.

Методи та моделі

Для здійснення анотації тексту в українській мові, використовуючи абстрактивний метод, ми вибрали високоефективну модель Pegasus XSum від Google. Ця модель працює на принципі енкодер-декодер та нейронних мереж, які визначають контекст та основну ідею тексту.

Для виконання анотацій оберемо абстрактивний метод. Цей підхід ґрунтується на здатності моделі розуміти документ, обробляючи токени через кодувальний та декодувальний шари нейронної мережі, таким чином виділяючи контекст та основну ідею документа. Для нашої системи анотування ми обрали модель Pegasus XSum від Google. Ця модель ініціює процес анотування, спочатку розглядаючи контекст усього вхідного тексту і кодуючи його в числове представлення, відоме як вектор контексту. Однак варто зазначити, що ця модель працює лише англійською мовою. Щоб адаптувати її до українського тексту, ми використовуємо сервіс перекладу DeepL для перетворення вхідного документа з української на англійську мову. Цей етап перекладу є ключовим для забезпечення сумісності з моделлю. Після завершення перекладу ми застосовуємо токенізатор до отриманого документа і використовуємо попередньо навчену модель для генерації вихідного тексту. Щоб оцінити точність анотації, ми перекладаємо документ назад українською мовою і суб'єктивно оцінюємо кінцевий результат, перевіряючи, що основна ідея оригінального тексту точно збережена в процесі анотування. Такий багатокроковий підхід дозволяє нам використовувати можливості великих мовних моделей та адаптувати їх для стислого викладу тексту українською мовою.

Розглянемо ще альтернативний підхід до анотування текстів, який передбачає використання поширених багатомовних попередньо навчених моделей, які підтримують українську мову. Цей підхід дає змогу оцінити ефективність використання машинного перекладу та попередньо навченої моделі на англійській мові, порівняно з моделлю, навченою спеціально для української мови. Для цього експерименту я обрав універсальну багатомовну модель T5 XLSum. Для проведення цього експерименту з веб-сайту BBC було отримано масив даних українською мовою, що містить загалом 54 тисячі записів. Цей набір даних слугує цінним ресурсом для нашого дослідження, сприяючи проведенню ретельних експериментів та отриманню надійних результатів.

Розвиток української системи обробки природної мови (NLP) є критичним імперативом у поточній роботі. Щоб досягти цієї мети, ми впроваджуємо комплексну стратегію, спрямовану на підвищення точності нашої існуючої мовної моделі. Вона починається зі створення ретельно курованого набору даних. Використовуючи перший метод, що передбачає переклад, ми створимо набір даних, який складатиметься з новин українською мовою, кожна з яких супроводжуватиметься автоматично згенерованою анотацією. Цей процес переслідує подвійну мету. По-перше, він надає нам значний набір даних, який інкапсулює суть українського тексту, а по-друге, забезпечує нас набором анотацій, закладаючи основу для навчання нашої моделі. Після етапу створення набору даних ми зосередимося на навчанні нашої цільової мережі, яка підготовлена багатим українським контентом набору даних, пройде навчання. Основна мета полягає в тому, щоб покращити розуміння моделлю складних нюансів української мови, зокрема ідіоматичних виразів, контекстуальних тонкощів та синтаксичних структур, притаманних лише українській мові. Наша гіпотеза полягає в тому, що в результаті ми отримаємо модель, здатну створювати не лише точні, але й контекстуально релевантні анотації українською мовою. Ми очікуємо, що завдяки збагаченню бази знань та лінгвістичних

можливостей моделі якість згенерованих анотацій відчутно покращиться. По суті, наша головна мета зрозуміла: підвищити продуктивність нашої системи в галузі анотування текстів українською мовою. Ітеративно вдосконалюючи нашу модель шляхом розширення набору даних і цілеспрямованого навчання, ми прагнемо досягти помітного покращення якості та глибини наших анотацій. Це цілком узгоджується з нашою місією розвивати сферу українського NLP і, таким чином, робити внесок у ширший спектр розуміння та обробки природної мови.

Результати експерименту

У нашому прагненні вирішити проблему скорочення тексту для українського тексту ми дослідили два різні підходи: використання моделі Pegasus, інтегрованої з перекладачем, і моделі mT5, розробленої спеціально для завдань скорочення тексту українською мовою. Оцінимо наші підходи згідно показника швидкості роботи та запишемо їх у таблицю 1.

Таблиця. 1

Швидкодія виконання однієї анотації

	Pegasus з перекладом	**mT5**
Мінімум	3.76 мсек	3.74 мсек
Середнє арифметичне	4.51 мсек	3.98 мсек
Медіана	4.59 мсек	4.01 мсек
Максимум	5.37 мсек	4.81 мсек

Очевидно, що модель mT5 продемонструвала перевагу в швидкості роботи над інтегрованою з перекладачем моделлю Pegasus. У більшості тестових кейсів mT5 випереджає свій аналог приблизно на 500 мілісекунд, що свідчить про його ефективність при узагальненні українського тексту. Хоча є випадки, коли обидва підходи працюють однаково, загальна тенденція свідчить про те, що модель mT5 стабільно виконує свої завдання швидше. Це спостереження підкреслює перевагу використання моделі, спеціально пристосованої для реферування українських текстів, оскільки вона усуває потребу в додаткових етапах перекладу, пов'язаних з підходом на основі Pegasus. Тобто поточні висновки відповідають нашій меті — підвищити ефективність і швидкість реагування системи.

Коли справа доходить до оцінки якості анотування тексту, важливо визнати, що не завжди просто кількісно виміряти результати за допомогою математичних формул. Проблема полягає в тому, що, хоча ми можемо оцінити рівень володіння мовою мовної моделі, визначення того, чи точно вона створила анотації найважливіших частин документа, є складнішим завданням. Автоматичні системи оцінювання стикаються з обмеженнями в оцінюванні семантичної релевантності та контекстуальної точності згенерованих анотацій. Хоча вони чудово справляються з оцінюванням на основі мови, їм часто не вдається вловити нюанси розуміння предмету. Тим не менш, існують метрики, такі як ROUGE (Recall-Oriented Understudy for Gisting Evaluation), які пропонують засоби для оцінки узагальнень у створених анотаціях. ROUGE порівнює автоматично створені анотації з анотаціями, створеними людиною. Однак варто зазначити, що навіть ROUGE не може повністю визначити, чи точно згенерований текст відповідає темі документа, тому суб'єктивна оцінка у цьому контексті відіграє важливу роль. Оцінюючи суб’єктивно результати, то результат задовільний. Бували випадки, коли узагальнення було некоректне, але майже всі анотаціх відображали основну думку оригінального тексту, тому наші системи функціонують коректно у контексті анотування текстів українською мовою.

Висновок

Ця робота присвячена розвитку галузі обробки природної мови шляхом розробки ефективної системи анотування текстів українською мовою. Нашою першочерговою метою було створення системи, здатної аналізувати вхідні українські тексти, виокремлювати основні поняття та генерувати змістовні анотації. Під час дослідження ми заглибилися в сферу анотування українських текстів, ретельно вивчивши дві різні методології: одну, що поєднує модель Pegasus із перекладачем, і другу, що використовує модель mT5, ретельно адаптовану до завдань анотування українських текстів.

Наша комплексна оцінка цих підходів, з особливою увагою до їхніх показників швидкості та продуктивності, дозволила зробити цінні висновки. Вона підкреслила ключову роль спеціалізованих моделей в ефективній обробці тонкощів української мови, пропонуючи багатообіцяючі результати в галузі анотування текстів. Крім того, ми підкреслили важливість суб'єктивності в оцінюванні ефективності системи в передачі основної суті вихідного тексту, визнаючи, що хоча математичні показники мають своє місце, вони не можуть повністю охопити семантичну точність.

По суті, ця наукова робота робить значний внесок у розвиток українських систем NLP, впроваджуючи інноваційні підходи до анотування текстів. Вона не лише проливає світло на виклики та перспективи в цій галузі, а й підкреслює першорядну важливість спеціалізованих моделей та суб'єктивних оцінок для досягнення точності та контекстуальної релевантності анотацій для українських текстів. Наша робота слугує

каталізатором подальшого поступу в галузі опрацювання української мови, обіцяючи глибше розуміння цього багатого лінгвістичного середовища та його практичних застосувань.

Література

1. Liu Y, Lapata M. Text summarization with pretrained encoders. В 2019. с. 3730–40.
2. Madhuri JN, Ganesh Kumar R. Extractive Text Summarization Using Sentence Ranking. В 2019.
3. Mohd M, Jan R, Shah M. Text document summarization using word embedding. Expert Syst Appl. 2020;143.
4. Vysotsk V, Holoshchuk S, Holoshchuk R. A comparative analysis for english and ukrainian texts processing based on semantics and syntax approach. В 2021. с. 311–56.
5. Telenyk S, Pogorilyy S, Kramov A. The Complex Method of Coreferent Pairs Detection in a Ukrainian-language Text Based on a BiLSTM Neural Network. В 2021. с. 205–10.
6. Wang Q, Liu P, Zhu Z, Yin H, Zhang Q, Zhang L. A text abstraction summary model based on BERT word embedding and reinforcement learning. Appl Sci Switz. 2019;9(21).

References

1. Liu Y, Lapata M. Text summarization with pretrained encoders. В 2019. с. 3730–40.
2. Madhuri JN, Ganesh Kumar R. Extractive Text Summarization Using Sentence Ranking. В 2019.
3. Mohd M, Jan R, Shah M. Text document summarization using word embedding. Expert Syst Appl. 2020;143.
4. Vysotsk V, Holoshchuk S, Holoshchuk R. A comparative analysis for english and ukrainian texts processing based on semantics and syntax approach. В 2021. с. 311–56.
5. Telenyk S, Pogorilyy S, Kramov A. The Complex Method of Coreferent Pairs Detection in a Ukrainian-language Text Based on a BiLSTM Neural Network. В 2021. с. 205–10.
6. Wang Q, Liu P, Zhu Z, Yin H, Zhang Q, Zhang L. A text abstraction summary model based on BERT word embedding and reinforcement learning. Appl Sci Switz. 2019;9(21).