

О. М. ШУШУРА

доктор технічних наук, професор,
професор кафедри цифрових технологій в енергетиці
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
ORCID: 0000-0003-3200-720X

К. В. НОВИЦЬКИЙ

магістр кафедри цифрових технологій в енергетиці
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
ORCID: 0009-0002-5000-1397

В. О. ІВАНОВ

магістр кафедри цифрових технологій в енергетиці
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
ORCID: 0009-0002-2498-5628

АВТОМАТИЧНЕ РЕЗЮМУВАННЯ НАУКОВИХ ДОКУМЕНТІВ НА ОСНОВІ МОДЕЛЕЙ-ТРАНСФОРМЕРІВ

Зростання обсягів наукової літератури актуалізує розробку ефективних методів автоматичної генерації стислих резюме. Традиційні підходи стикаються з обмеженням контекстного вікна, що робить неможливим безпосередню обробку документів довжиною понад кілька тисяч токенів.

Метою даної роботи є розробка гібридного методу автоматичного резюмування для узагальнення документів, які перевищують стандартні обмеження розміру контекстного вікна моделей-трансформерів.

Розроблений гібридний метод поєднує екстрактивні та абстрактивні методи резюмування для ефективної обробки документів довільної довжини. Для екстрактивної фази була використана модель Sentence-BERT, з метою отримати семантичні векторні представлення речень, що дозволило ідентифікувати найбільш важливі частини тексту. На відміну від статистичних методів, Sentence-BERT захоплює глибокий семантичний зміст незалежно від лексичного складу. Наступна фаза методу видаляє семантичні дублікати за допомогою косинусної подібності, що забезпечує компактність проміжного представлення. Метод ідентифікує як точні дублікати, так і перефразування, створюючи компактне резюме. Фаза абстрактивної генерації виконується з використанням моделі BART-large-CNN, що поєднує двонаправлене кодування та авторегресивну генерацію. Це забезпечує створення зв'язних резюме з власними формулюваннями моделі, здатність до перефразування та об'єднання інформації з різних частин документу.

Розроблено програмне забезпечення для реалізації методу згідно з SOLID принципами, забезпечуючи модульність та можливість розширення системи. Проведено порівняльне дослідження розробленого методу з чотирма категоріями базових підходів і спеціалізованою моделлю яка має розширене вікно контексту LongT5. Оцінка на вибірці з наукових статей з arXiv показала, що запропонований метод краще показує себе аніж традиційні методи та працює на рівні з LongT5, використовуючи при цьому стандартну модель BART-large-CNN. Метод був застосований без додаткового перед-навчання, що знижує обчислювальні вимоги.

Ключові слова: *резюмування тексту, обробка природньої мови, моделі трансформери, машинне навчання, інформаційні системи.*

О. М. SHUSHURA

Doctor of Technical Sciences, Professor,
Professor at the Department of Digital Technologies in Energy
National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
ORCID: 0000-0003-3200-720X



K. V. NOVYTSKYI

Master at the Department of Digital Technologies in Energy
National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
ORCID: 0009-0007-2508-4350

V. O. IVANOV

Master at the Department of Digital Technologies in Energy
National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
ORCID: 0009-0002-2498-5628

AUTOMATIC SUMMARISATION OF SCIENTIFIC DOCUMENTS BASED ON TRANSFORMER-MODELS

The rapid growth of scientific literature highlights the need for effective methods of automatic summary generation. Traditional approaches face limitations imposed by fixed context windows, which makes the direct processing of documents longer than several thousand tokens impractical.

The aim of this work is to develop a hybrid automatic summarization method for processing documents that exceed the standard context window limitations of transformer-based models.

The proposed hybrid method combines extractive and abstractive summarization techniques to efficiently handle documents of arbitrary length. In the extractive phase, the Sentence-BERT model is employed to obtain semantic vector representations of sentences, enabling the identification of the most informative parts of the text. Unlike statistical methods, Sentence-BERT captures deep semantic meaning regardless of lexical variation. The subsequent phase removes semantic duplicates using cosine similarity, ensuring the compactness of the intermediate representation. The method identifies both exact duplicates and paraphrases, producing a concise intermediate summary. The abstractive generation phase is performed using the BART-large-CNN model, which combines bidirectional encoding with autoregressive generation. This enables the production of coherent summaries with model-generated phrasing, paraphrasing capabilities, and the integration of information from different parts of the document.

Software implementing the proposed method was developed in accordance with the SOLID principles, ensuring modularity and extensibility of the system.

A comparative study was conducted against four categories of baseline approaches as well as the specialized LongT5 model with an extended context window. Evaluation on a dataset of scientific articles from arXiv demonstrated that the proposed method outperforms traditional approaches and achieves performance comparable to LongT5, while relying on the standard BART-large-CNN model. The method was applied without additional pre-training, which significantly reduces computational requirements.

Key words: text summarization, natural language processing, transformer models, machine learning, information systems.

Постановка проблеми

Резюмування текстів є однією з ключових задач обробки природної мови, яке має широке застосування в багатьох сферах життя, таких як юриспруденція, наукова діяльність, тощо. Резюмування буває двох видів – екстрактивне і абстрактивне. Суть екстрактивного резюмування полягає в виборі найголовніших речень в тексті, без перефразування, в той час як абстрактивне резюмування створює короткий виклад, перефразовуючи і узагальнюючи зміст своїми словами. Через значні виклики при перефразуванні, домінуючим методом резюмування до винайдення моделей-трансформерів був екстрактивний [1]. Сучасні моделі-трансформери, такі як BART, PEGASUS, T5 демонструють дуже гарні результати для коротких і середніх за обсягом текстів, проте мають обмеження розміру контекстного вікна (зазвичай 512–1024 токенів), що значно ускладнює роботу з документами, довгими ніж контекстне вікно [2]. Усічення документів призводить до неповної обробки та, відповідно, неточного резюмування. Екстрактивне ж резюмування, як от з використанням графових методів, не завжди може забезпечити необхідну якість узагальнення. Це зумовило появу значної кількості досліджень, у яких пропонуються нові підходи для резюмування довгих текстів.

Аналіз останніх досліджень і публікацій

Абстрактивне узагальнення часто реалізують за допомогою таких передтренированих моделей як BART, PEGASUS, T5. Модель BART має гарні результати завдяки архітектурі енкодер-декодер і попередньому навчанні [3]. Модель PEGASUS оптимізована спеціально для задач резюмування завдяки використанню *gap-sentence generation* під час попереднього навчання [4]. Модель T5 (Text-to-text Transformer) розглядає резюмування як задачу перетворення тексту, досягаючи універсальності за рахунок єдиної архітектури.

Для обробки довгих документів було запропоновано декілька підходів. Модель Hierarchical Transformers [5] використовують багаторівневу архітектуру для обробки локальних та глобальних залежностей. Модель Longformer [6] застосовує механізм розрідженої уваги для масштабування до довгих послідовностей. Модель

LED (Longformer Encoder-Decoder) поєднує переваги обох підходів для задач резюмування. Однак ці моделі вимагають значних обчислювальних ресурсів та великих обсягів даних для навчання.

Гібридні підходи намагаються поєднати переваги екстрактивних та абстрактивних методів. Автори Zhang et al. [7] запропонували двоетапний підхід, де спочатку екстрактивна модель виділяє важливі фрагменти, а потім абстрактивна модель генерує фінальне узагальнення.

Нарешті, Sentence-BERT [8] запропонував ефективний спосіб отримання семантичних векторних представлень для речень, що широко використовується у задачах пошуку схожості та кластеризації. Застосування цього підходу для екстрактивного резюмування показало перспективні результати, але неясним залишається можливість його найкращого поєднання з абстрактивними моделями для обробки довгих документів.

Таким чином, незважаючи на значний прогрес у розробці спеціалізованих архітектур для обробки довгих документів, питання ефективного поєднання екстрактивних та абстрактивних методів на основі стандартних моделей-трансформерів залишається недостатньо дослідженим. Зокрема, потенціал інтеграції семантичних векторних представлень Sentence-BERT з абстрактивними моделями типу BART для створення масштабованих рішень без потреби у значних обчислювальних ресурсах та розширених контекстних вікнах потребує детального вивчення. Це визначає актуальність подальших досліджень гібридного застосування моделей-трансформерів для задачі резюмування документів.

Формулювання мети дослідження

Метою даної роботи є розробка гібридного методу резюмування для узагальнення документів, які перевищують стандартні обмеження розміру контекстного вікна моделей-трансформерів.

Для досягнення поставленої мети потрібно вирішити наступні задачі:

- розробити гібридний метод резюмування довгих документів, що поєднує екстрактивні та абстрактивні підходи в рамках чотирьох-етапної архітектури;
- здійснити програмну реалізацію розробленого методу з дотриманням принципів модульності та розширюваності;
- провести порівняльне дослідження запропонованого методу з існуючими підходами на репрезентативній вибірці наукових документів.

Викладення основного матеріалу дослідження

На основі комбінації Map-Reduce підходу з семантичними векторними представленнями та трансформерними моделями було розроблено гібридний метод резюмування документів, узагальнений алгоритм якого наведено на рисунку 1. Метод поєднує Map-Reduce підхід з семантичним відбором і абстрактивною генерацією для досягнення достатньої якості узагальнень, без потреби використання моделей-трансформерів з розширеним контекстним вікном.

Як видно на рисунку 1, на початку методу відбувається розбиття вхідного документу на фрагменти оптимального розміру, при цьому кожен наступний фрагмент включає останні 100 слів попереднього, що запобігає втраті семантичного зв'язку між реченнями на межі розбиття [9].

На другому етапі методу використовується модель Sentence-BERT для семантичного вибору найбільш важливих речень з кожного фрагменту.

На відміну від статистичних методів (TF-IDF) або позиційних методів (Lead-k), Sentence-BERT є оновленою архітектурою BERT, яка спеціально адаптована для ефективної генерації векторних представлень речень. Оригінальна модель BERT вимагає передачі речень парами через мережу, що призводить до складності $O(n^2)$. Модель Sentence-BERT вирішує цю проблему через використання сіамської архітектури, що дозволяє кодувати кожне речення незалежно за один прохід, знижуючи складність до лінійної $O(n)$. У даній роботі була використана модель all-MiniLM-L6-v2 з шістьма шарами, навчена методом контрастного навчання на понад мільярд пар речень. Алгоритм семантичної екстракції складається з декількох етапів. На першому етапі фрагмент документу розбивається на окремі речення за допомогою бібліотеки NLTK, що коректно обробляє абрєвіатури та інші специфічні конструкції наукових текстів. Далі кожне речення кодується моделлю all-MiniLM-L6-v2 у 384-вимірний вектор. Процес кодування включає токєнізацію за допомогою WordPiece tokenizer (словник 30 000 слів) та проходження через шість шарів трансформера з механізмом самоуваги, а також усереднення векторних подань для отримання підсумкового вектору речення. Для речення з m токєнами, представленими векторами $h_1, h_2, \dots, h_m \in R^{768}$, отримуємо фінальне векторне представлення:

$$e = \frac{1}{m} \cdot \sum_{i=1}^m h^i, \quad (1)$$

де $e \in R^{384}$ після проєкції з 768 на 384 виміри.

Потім обчислюється центроїд фрагменту – векторне представлення, яке задає загальну семантику тексту. Для фрагменту з n речень і векторами $e_1, e_2, \dots, e_n \in R^{384}$, центроїд $c \in R^{384}$ визначається як середнє арифметичне за формулою:

$$c = \frac{1}{n} \cdot \sum_{i=1}^n e^i. \quad (2)$$

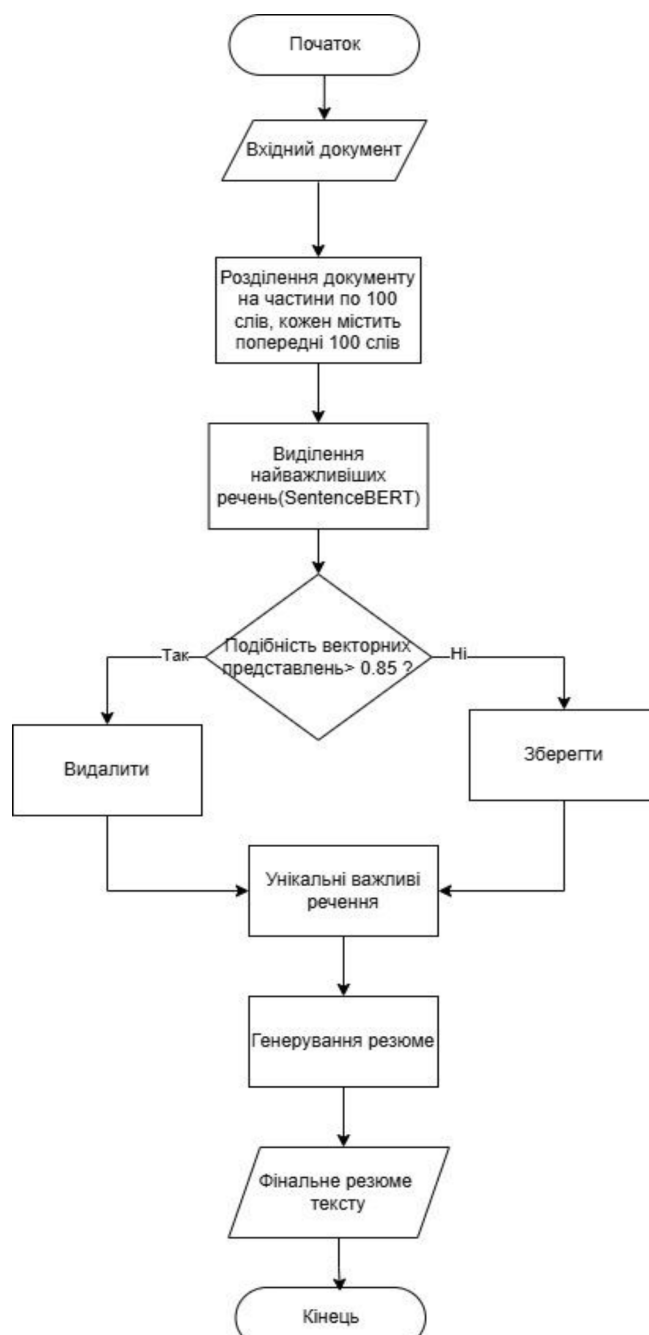


Рис. 1. Узагальнена схема методу гібридного резюмування документу

Геометрично центроїд є центром усіх векторних представлень у просторі, базуючись на припущенні, що найбільш репрезентативні речення про текст близькі до усередненого вектора усіх речень. Для кожного речення обчислюється косинусна подібність до центроїду за формулою:

$$\text{similarity}(e_i, c) = \frac{e_i \cdot c}{(\|e_i\| \cdot \|c\|)}, \quad (3)$$

де $e_i \cdot c$ – скалярний добуток, а $\|e_i\|, \|c\|$ – Евклідова норма.

Косинусна подібність нормалізована на інтервалі $[-1, 1]$ і незалежна від довжини векторів, що означає що вона є оптимальною для високимірних просторів. Відбираються декілька речень з найвищою подібністю, зберігаючи оригінальний порядок речень.

Третій етап методу видаляє семантичні дублікати на основі косинусної подібності векторних представлень. Оскільки перший етап створює фрагменти тексту, які містять частини інших для збереження цілісності тексту, а фаза виділення найважливіших частин обробляє кожен фрагмент незалежно, деякі речення можуть бути обрані

декілька разів. Крім того, різні речення можуть висловлювати ідентичні ідеї в різних формулюваннях, що негативно впливає на якість фінального резюме. Для визначення семантичної схожості речень з векторами $e_1, e_2, \dots, e_n \in R^{384}$ застосовується косинусна подібність векторних представлень:

$$\text{similarity}(e_i, e_j) = \frac{e_i \cdot e_j}{(\|e_i\| \cdot \|e_j\|)},$$

де $e_i \cdot e_j$ – скалярний добуток, а $\|e_i\|$ – Евклідова норма.

Косинусна подібність нормалізована в інтервалі $[-1, 1]$, де значення, наближені до 1, свідчать про високий рівень семантичної схожості між текстовими одиницями. Речення вважаються дублікатами за умови, що значення косинусної подібності перевищує поріг 0,85. Вибір даного порогового значення обумовлено результатами емпіричних спостережень: нижчі пороги призводять до надмірного видалення інформації, тоді як вищі не забезпечують фільтрацію дублікатів із незначними варіаціями формулювань.

Далі для набору з n речень $S = \{s_1, s_2, \dots, s_n\}$ з векторами $E = \{e_1, e_2, \dots, e_n\}$ ініціалізується порожня множина унікальних речень U . Для кожного речення s_i обчислюється попарна косинусна подібність з усіма реченнями U . Якщо максимальна подібність не перевищує задане значення, речення додається до множини U , інакше відкидається. Таким чином, алгоритм має складність $O(n \cdot m)$, де m – розмір фінальної множини U .

Наприкінці методу генерується фінальне резюме з використанням моделі BART-large-CNN (Bidirectional and Auto-Regressive Transformers), яка є архітектурою типу sequence-to-sequence на основі трансформерів, що поєднує двонаправлене кодування вхідного тексту і авторегресивну генерацію вихідного тексту. Довжина резюме регулюється параметрами, і зазвичай складає 80–120 слів. На основі розробленого методу було створене відповідне програмне забезпечення.

Програмний продукт для застосування методу розроблено в середовищі PyCharm на мові Python версії 3.11. Програма складається з файлу розробленої моделі та файлів інших методів, які використовувались для порівняння. На рисунку 2 зображена діаграма класів розробленого методу.

Розроблена програма містить 4 інтерфейси, які визначають абстракції для кожної фази обробки документа. Інтерфейс IDocumentSplitter відповідає за розбиття вхідного тексту, ISentenceExtractor забезпечує виділення репрезентативних речень, IRedundancyReducer виконує видалення семантичних дублікатів, а IAbstractiveSummarizer

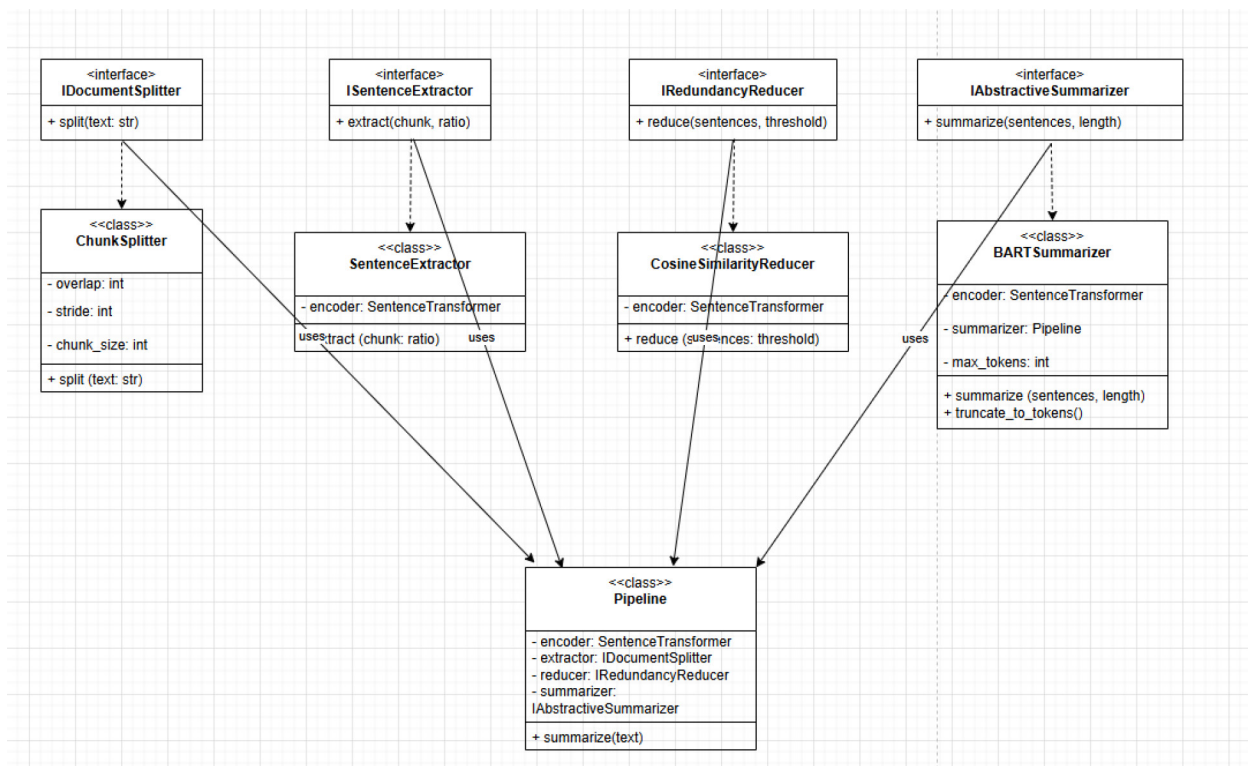


Рис. 2. Діаграма класів розробленого методу

генерує фінальне резюме. Кожен інтерфейс визначає один ключовий метод. Кожний інтерфейс має клас, який його реалізує. Клас ChunkSplitter розбиває документ з параметрами overlap і chunk_size. Наступний клас, SentenceExtractor використовує модель SentenceTransformer для створення векторних представлень і відбору

речень на основі центроїду. Клас CosineSimilarityReducer застосовує косинусну подібність для видалення дублікатів, також використовуючи SentenceTransformer. Головний клас BARTSummarizer використовує модель BART для генерації фінального резюме через Pipeline клас.

Клас Pipeline виступає координатором всієї системи, яка агрегує всі компоненти через їх інтерфейси. Використовуючи розроблене програмне забезпечення, було виконано порівняльне дослідження методу на тестових даних.

Тестування проводилось на наборі наукових статей з arXiv. Були відібрані 5 документів різної довжини (від 3000 до 6000 слів) з галузей фізики і математики. Кожен документ містив короткий опис на початку документу який був використаний в якості порівняння для оцінки якості згенерованого резюме.

Оцінка якості проводилась з використанням двох груп метрик. Перша група – метрики ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), які вимірюють лексичне перекриття і найдовшої спільної послідовності. Друга група метрик – BERTscore, які вимірюють семантичну подібність між згенерованим і еталонним резюме на основі концептуальних векторних представлень.

В таблиці 1 наведені результати порівнянь за метриками ROUGE. Були розглянуті як традиційні методи (TF-IDF), так і більш новітні, на основі машинного навчання. Метод на основі позиції речення демонструє високі результати завдяки специфічній структурі наукових статей, де найважливіша інформація традиційно концентрується на початку документа. Однак цей підхід має обмежену універсальність і не забезпечує ефективної роботи з документами альтернативної структури.

Таблиця 1

Порівняння ROUGE-1 метрики для різних методів

Метод	Складність	ROUGE-1
Lead-k(позиція речення)	$O(1)$	0.33
TF-IDF + косинусна подібність	$O(n^2)$	0.27
Sentence-BERT + центроїд(розроблений)	$O(n)$	0.35
BERT cls token	$O(n^2)$	0.38

Метод TF-IDF показав найнижчі результати серед досліджуваних підходів. Це пояснюється тим, що він враховує лише частотність слів, ігноруючи їх семантичну значущість, що призводить до неефективної обробки текстів з великою кількістю спеціалізованої термінології.

Метод на основі BERTcls токенів продемонстрував найкращі показники якості, проте поступається використаному в дослідженні методу Sentence-BERT з вибором за центроїдом за швидкістю обробки. Ця різниця стає критичною при роботі з довгими документами, де час обробки є важливим фактором.

Застосування семантичних векторних представлень Sentence-BERT забезпечує оптимальний баланс між якістю відбору найважливіших речень та ефективністю обробки документів. На відміну від методів, що базуються на позиції речення чи частотності слів, цей підхід здатний вилучати інформативні речення з будь-якої частини документа, включаючи середину та кінець, а не лише з початку, як це характерно для традиційних методів. Така властивість робить його більш універсальним для різних типів документів та структур викладу інформації.

На рисунку 3 зображені результати порівняння за ROUGE-метриками. Були розглянуті наступні методи:

- розроблений метод: Map-Reduce з вектором семантичних представлень;
- LongT5: спеціалізована модель-трансформер, здатна оброблювати довгі документи(до 16000 токенів);
- Map-Reduce+BART: базовий Map-Reduce без семантичної екстракції;
- TextRank+BART: виділення найважливіших речень на основі графового алгоритму і генерація резюме через BART;

- T5-base: стандартна модель трансформер, яка приймає до 1024 токени.

Як видно з діаграми на рисунку 3, за метрикою ROUGE-1 F1 наявна перевага запропонованого методу – він показав найвищий результат (0.5852 ± 0.027). Це на 4.4 % краще ніж спеціалізована модель-трансформер LongT5(0.5603 ± 0.017) і на 16.1 % краще за базовий Map-Reduce+BART.

За метрикою ROUGE-2, що вимірює збіг біграм, запропонований метод показав кращі результати також: 0.1420 ± 0.032 проти 0.0980 ± 0.028 у LongT5 (+44.9 %). Це вказує на те, що згенеровані резюме не тільки містять правильні слова, але й зберігають типові словосполучення з оригінальних окументів. За метрикою ROUGE-L F1 (найдовша спільна підпослідовність) результат запропонованого методу становить 0.2385 ± 0.045 проти 0.1840 ± 0.038 у LongT5, що свідчить про краще збереження структури та порядку інформації з оригінального тексту.

Результати порівняння за метрикою BERTscore представлені на рисунку 4.

Як видно з діаграми, запропонований метод показав найкращий результат з усіх методів – 0.5852. Це на 4.4 % краще за спеціалізовану модель-трансформер LongT5 і на 16 % краще за схожий Map-Reduce метод.

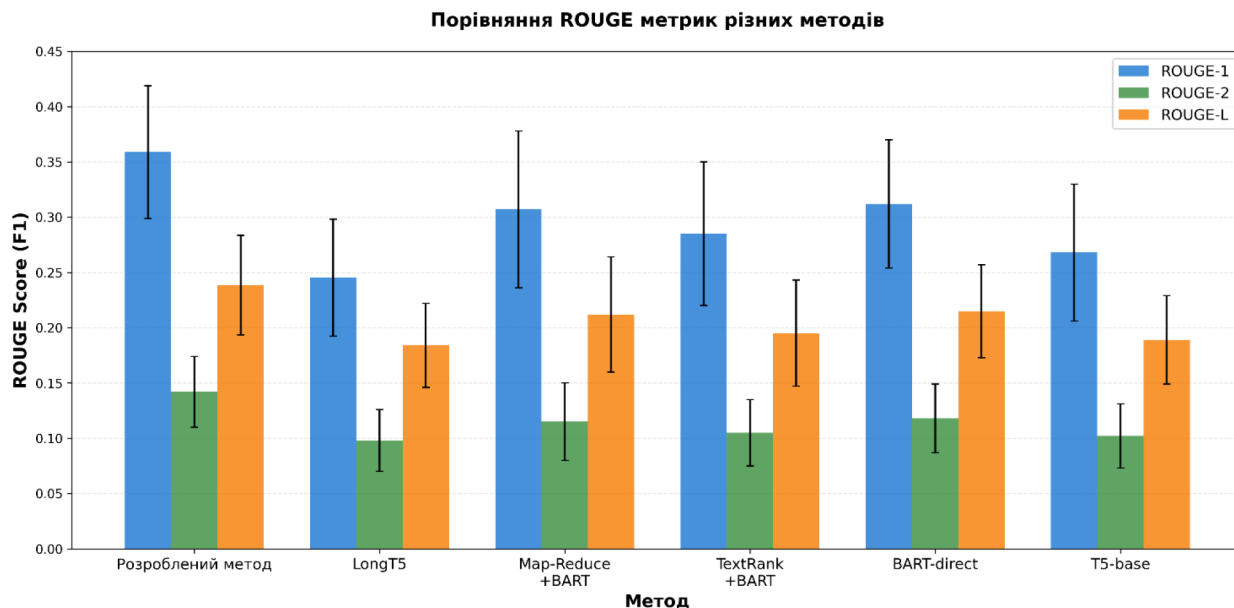


Рис. 3. Порівняння результатів ROUGE-метрик для різних методів

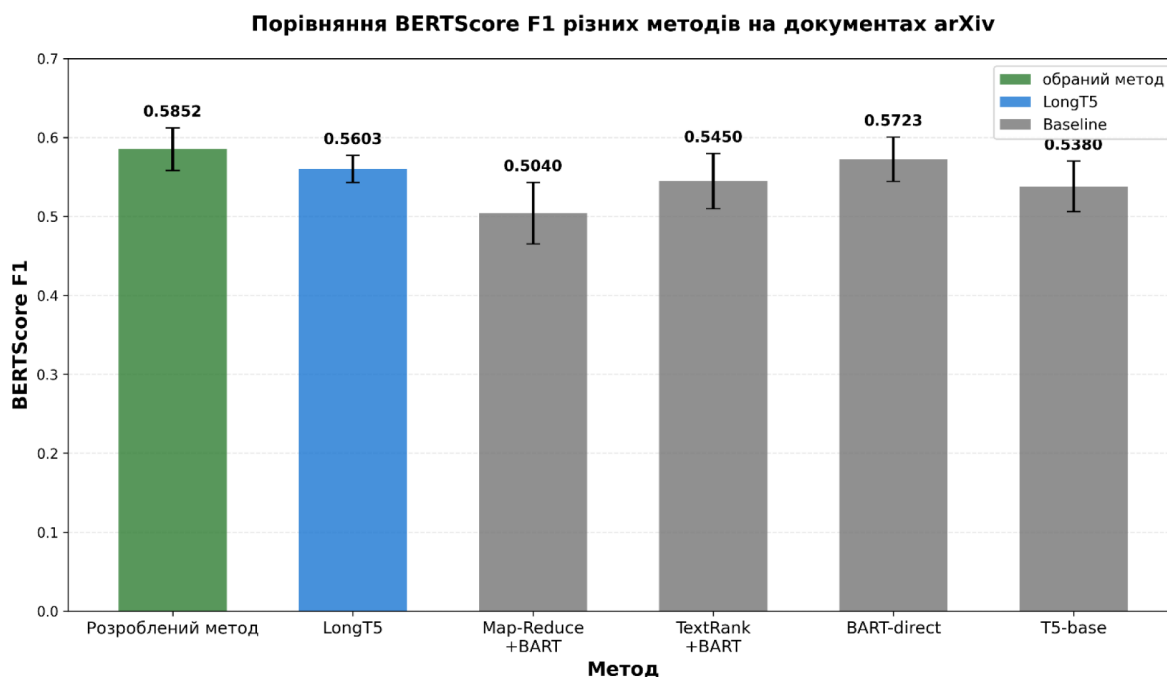


Рис. 4. Порівняння результатів BERTscore-метрик для різних методів

Висновки

Розроблено гібридний метод резюмування довгих наукових документів, що поєднує розбиття документу на перекриваючі фрагменти, семантичну екстракцію з використанням Sentence-BERT, видалення семантичних дублікатів через косинусну подібність векторних представлень та абстрактивну генерацію за допомогою BART-large-CNN. Метод дозволяє обробляти документи довільної довжини без використання спеціалізованих моделей з розширеним контекстним вікном.

Здійснено програмну реалізацію методу з дотриманням SOLID принципів, що забезпечує модульність та можливість заміни окремих компонентів системи. З використанням розробленого програмного забезпечення проведено порівняльне дослідження методу на вибірці наукових статей arXiv. За метрикою ROUGE-1 F1 запропонований метод показав результат 0.5852 ± 0.027 , що на 4.4 % краще за спеціалізовану модель LongT5 та на 16.1 % краще за базовий Map-Reduce + BART. Результати підтверджують ефективність гібридного підходу для резюмування документів з використанням стандартних моделей-трансформерів без додаткового дотренування.

Отримані результати підтверджують доцільність подальших досліджень з поєднання семантичної екстракції на основі векторних представлень з абстрактивною генерацією для задачі резюмування наукових документів.

Список використаної літератури

1. See A., Liu P. J., Manning C. D. Get To The Point: Summarization with Pointer-Generator Networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. Vol. 1. P. 1073–1083. <https://doi.org/10.18653/v1/P17-1099>
2. Huang L., Wu L., Wang L. An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics. ACM Computing Surveys. 2022. Vol. 55, № 8. Article 157. <https://doi.org/10.1145/3545176>
3. Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 7871–7880. <https://doi.org/10.18653/v1/2020.acl-main.703>
4. Zhang J., Zhao Y., Saleh M., Liu P. J. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. Proceedings of the 37th International Conference on Machine Learning. 2020. P. 11328–11339. <https://doi.org/10.48550/arXiv.1912.08777>
5. Liu Y., Lapata M. Hierarchical Transformers for Multi-Document Summarization. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 5070–5081. <https://doi.org/10.18653/v1/P19-1500>
6. Beltagy I., Peters M. E., Cohan A. Longformer: The Long-Document Transformer. arXiv preprint arXiv:2004.05150. 2020. <https://doi.org/10.48550/arXiv.2004.05150>
7. Zhang X., Wei F., Zhou M. HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 5059–5069. <https://doi.org/10.18653/v1/P19-1499>
8. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019. P. 3982–3992. <https://doi.org/10.48550/arXiv.1908.10084>
9. Automatic text summarization of scientific articles using transformers: A brief review. Journal of Artificial Intelligence. 2024. Vol. 7, № 5. <https://doi.org/10.32629/jai.v7i5.1331>

References

1. See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 1073–1083). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-1099>
2. Huang, L., Wu, L., & Wang, L. (2022). An empirical survey on long document summarization: Datasets, models, and metrics. ACM Computing Surveys, 55(8), Article 157. <https://doi.org/10.1145/3545176>
3. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871–7880). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.703>
4. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the 37th International Conference on Machine Learning (pp. 11328–11339). PMLR. <https://doi.org/10.48550/arXiv.1912.08777>
5. Liu, Y., & Lapata, M. (2019). Hierarchical transformers for multi-document summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 5070–5081). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1500>
6. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150. <https://doi.org/10.48550/arXiv.2004.05150>
7. Zhang, X., Wei, F., & Zhou, M. (2019). HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 5059–5069). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1499>
8. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.48550/arXiv.1908.10084>
9. Automatic text summarization of scientific articles using transformers: A brief review. (2024). Journal of Artificial Intelligence, 7(5). <https://doi.org/10.32629/jai.v7i5.1331>

Дата першого надходження статті до видання: 14.01.2026

Дата прийняття статті до друку після рецензування: 19.02.2026

Дата публікації (оприлюднення) статті: 30.04.2026