УДК 519.816
DOI: https://doi.org/10.17721/AIT.2024.1.04

Олексій КУЗНЄЦОВ, асп.
ORCID ID: 0000-0002-3537-9976
e-mail: oleksiy1908@gmail.com
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна

Геннадій КИСЕЛЬОВ, канд. техн. наук, ст. наук. співроб.
ORCID ID: 0000-0003-2682-3593
e-mail: g.kyselov@gmail.com
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна

ЗАСТОСУВАННЯ ТА АНАЛІЗ ФОРМАЛЬНИХ МЕТОДІВ ОЦІНЮВАННЯ РЕЛЕВАНТНОСТІ АВТОМАТИЧНО СТВОРЕНИХ РЕФЕРАТІВ ІНФОРМАЦІЙНИХ ТЕКСТІВ

Вступ. ***Розглянуто існуючі підходи до оцінювання якості автоматично створених рефератів інформаційних текстів. Дано огляд методів автоматичного реферування, включаючи класичні підходи та сучасні моделі на основі штучного інтелекту. Огляд містить екстрактивні методи реферування, такі як TF-IDF та PageRank, а також графові методи, зокрема TextRank. Особливу увагу приділено абстрактним підходам, що включають моделі Generative Pretrained Transformer (GPT) і Bidirectional and Auto-Regressive Transformers (BART). Оцінювання якості генерованих рефератів виконують за допомогою кількісних метрик оцінювання релевантності рефератів, зокрема і ROUGE та BLEU.***

Методи. ***Проаналізовано кілька підходів до автоматичного реферування текстів. Класичні екстрактивні методи, зокрема і TF-IDF, обчислюють важливість термів на основі частоти їхнього вживання в документі та в колекції документів. PageRank і TextRank використовують графові моделі для визначення значущості речень на основі зв'язків між ними. Абстрактні методи, такі як GPT і BART, генерують нові речення, що стисло передають зміст оригінального тексту. Оцінювання ефективності кожного підходу здійснюється метриками ROUGE і BLEU, які вимірюють збіг між автоматично згенерованими рефератами й еталонними текстами. Особливу увагу приділено аналізу їхньої точності, гнучкості, вимогам до ресурсів і простоті реалізації.***

Результати. ***Результати дослідження свідчать, що метрики ROUGE показують хорошу точність у вимірюванні збігів n-грам (послідовностей з n слів), тоді як BLEU ефективна у завданнях машинного перекладу, але може не враховувати деякі синтаксичні особливості тексту. Оцінювання методів автоматичного реферування за допомогою цих метрик показала, що екстрактивні методи реферування, такі як TF-IDF, є ефективними для оброблення простих текстів, але можуть втратити важливий контекст у складних текстах. PageRank і TextRank дозволяють враховувати зв'язки між реченнями, проте можуть давати менш релевантні результати для текстів із слабко вираженими структурними зв'язками. Абстрактні моделі GPT і BART забезпечують гнучкіший підхід до реферування, створюючи нові речення, що краще передають зміст, однак потребують значних обчислювальних ресурсів і складні у впровадженні.***

Висновки. ***Поєднання класичних і сучасних методів автоматичного реферування текстів дозволяє досягти вищої якості результатів. Важливо враховувати специфіку тексту та вимоги до кінцевого результату, адаптуючи обрані підходи та метрики відповідно до завдання.***

Ключові слова: ***автоматичне реферування, екстрактивні методи, абстрактні методи, GPT, BART, ROUGE, BLEU, TextRank, PageRank, TF-IDF.***

Вступ

У сучасному світі інформація є одним із найцінніших ресурсів. Щодня створюють величезні обсяги нових даних – від наукових статей до новинних повідомлень, технічної документації та багато іншого. У такому інформаційному потоці швидкий доступ до релевантної інформації стає надзвичайно важливим. Тут на допомогу приходять реферати, які дозволяють отримати суть документа без необхідності читати його повністю.

Реферати виконують важливу роль, особливо у наукових дослідженнях, де вони допомагають дослідникам швидко зрозуміти, чи є певний документ корисним для їхньої роботи. Однак створення рефератів вручну вимагає багато часу і зусиль, що не завжди можливо через великі обсяги даних. Ця проблема стає особливо гострою в умовах швидкого зростання інформаційних потоків, коли необхідність автоматизованого реферування стає очевидною. Для оцінювання результатів реферування використовують відповідні метрики, які дозволяють визначити, наскільки точно і повно реферат передає зміст оригінального документа.

Класичні методи автоматичного реферування – екстрактивні методи, де важливі речення просто витягуються з тексту, й абстрактні методи, що вимагають створення нових речень на основі розуміння змісту, мають свої переваги та недоліки. З появою штучного інтелекту з'явилися нові підходи до реферування, що базуються на глибинному навчанні й нейронних мережах, які обіцяють значне покращення якості автоматично створених рефератів. Однак, попри значні досягнення, задача автоматичного реферування текстів не може вважатись розв'язаною.

Постановка задачі. Основна мета цієї статті полягає в аналізі існуючих метрик оцінювання релевантності автоматично створених рефератів відносно оригінальних інформаційних текстів. На основі експериментів із використання цих метрик сформульовано рекомендації щодо застосування методів автоматичного реферування інформаційних текстів. Для досягнення цієї мети необхідно розв'язати такі задачі.

1. Проаналізувати наявні екстрактивні методи реферування текстів, такі як TF-IDF, PageRank і TextRank, й оцінити їхні переваги та недоліки в різних контекстах застосування.
2. Проаналізувати існуючі абстрактні методи автоматичного реферування текстів, такі як GPT та BART.
3. Дослідити можливості й обмеження використовуваних кількісних метрик оцінювання релевантності рефератів, зокрема ROUGE та BLEU.
4. Провести порівняльний аналіз різних методів реферування з використанням кількісних метрик оцінювання релевантності рефератів
5. Розробити рекомендації щодо вибору оптимальних методів реферування текстів різних типів, враховуючи особливості задачі та вимоги до точності, контекстуальності й обчислювальних витрат.

© Кузнєцов Олексій, Кисельов Геннадій, 2024

Ця робота націлена на пошук балансу між точністю, ресурсною ефективністю та якістю автоматично згенерованих рефератів, що дозволить покращити ефективність оброблення текстової інформації у різних сферах застосування.

Методи

В роботі використано екстрактивні й абстрактні методи автоматичного реферування текстів. Розглянемо їх по черзі.

Екстрактивні методи

Автоматичне реферування текстів є важливим інструментом оброблення великих обсягів інформації для забезпечення швидкого доступу до ключових даних. В основу класичних підходів до автоматичного реферування покладено методи, які дозволяють автоматично витягувати найважливіші речення з тексту, зберігаючи при цьому оригінальний стиль і формулювання автора. Ці методи відомі як екстрактивні та використовують різні техніки й алгоритми для визначення значущості та релевантності речень у контексті всього документа (Kuznietsov, & Kyselov, 2024).

Класичні екстрактивні методи включають TF-IDF (Term Frequency-Inverse Document Frequency), алгоритми ранжування, такі як PageRank, графові методи, наприклад, TextRank, та різноманітні статистичні підходи. Кожен із цих методів має свої унікальні характеристики, що дозволяють ефективно виокремлювати ключову інформацію з тексту.

Розуміння принципів роботи вказаних методів, їхніх переваг та обмежень є важливим для подальшого вдосконалення автоматичних систем реферування. Розглянемо основні класичні підходи до автоматичного реферування текстів, їхні особливості та вплив на якість створених рефератів. Це дозволить оцінити ефективність таких методів і визначити напрямки для подальших досліджень і розвитку технологій автоматичного реферування текстів.

- *Екстрактивний метод* ***TF-IDF***

TF-IDF (*частота терміна – обернена частота документа*) – це статистичний показник, що використовується для визначення значущості терміна в межах одного документа з певного набору документів.

TF (*частота терміна*) визначає відношення кількості появи слова до загальної кількості слів у документі (формула 1). Це дозволяє оцінити важливість терміна *t* в конкретному документі *d* (Кузнєцов, Кисельов, 2022).

$$tf(t,d) = n_t / \sum_k n_k, \quad (1)$$

де n_t – кількість входжень терму t в документ; $\sum_k n_k$ – загальна кількість слів у документі.

IDF (*обернена частота документа*) відображає частоту, з якою певне слово зустрічається у всіх документах колекції (формула 2). IDF знижує вагу поширених слів, надаючи перевагу унікальним термінам. Для кожного унікального слова в межах набору документів існує одне значення IDF. Терміни з високою частотою в конкретному документі та низькою частотою в інших документах отримують вищий бал у TF-IDF.

$$idf(\mathrm{t}, D) = log\left(|D| \Big/ \left|\left\{d_i \in |D| \mathrm{t} \in d_i\right\}\right|\right), \quad (2)$$

де $|D|$ – кількість документів у колекції; $\left|\left\{d_i \in |D| \mathrm{t} \in d_i\right\}\right|$ – кількість документів із колекції D, в яких зустрічається t (коли $n_t \neq 0$).

Відношення правдоподібності (log-likelihood ratio) є логарифмом відношення ймовірності спостереження слова з однаковою ймовірністю у корпусі вхідних документів і відповідних їм резюме до ймовірності появи слова з різними ймовірностями в цих корпусах (Lin, & Hovy, 2000).

Переваги методу TF-IDF

1. *Простота реалізації:* TF-IDF легко зрозуміти і впровадити. Його математична основа є інтуїтивно зрозумілою і базується на простих статистичних принципах.

Приклад. У будь-якій мові програмування, як-от Python, реалізація TF-IDF може бути здійснена за допомогою кількох рядків коду з використанням бібліотек, напр., scikit-learn.

2. *Ефективність для великих колекцій тексту:* TF-IDF добре масштабується для оброблення великих наборів даних і колекцій документів. Він дозволяє швидко й ефективно аналізувати і витягувати ключові слова з великої кількості текстів.

Приклад. Пошукові системи, такі як Google, використовують модифіковані версії TF-IDF для швидкого індексування і ранжування вебсторінок, дозволяючи користувачам отримувати релевантні результати пошуку за лічені секунди.

3. *Врахування значущості слів:* TF-IDF дозволяє оцінити важливість термінів у контексті документа, виділяючи ключові слова, які мають значення для конкретного документа.

Приклад. У статті про технології слова "штучний інтелект" і "машинне навчання" отримають високі значення TF-IDF, оскільки вони є ключовими термінами, що описують тему статті.

4. *Зменшення ваги поширених слів:* TF-IDF автоматично знижує вагу поширених слів, таких як сполучники і прийменники, які не несуть значної інформації.

Приклад. У новинній статті слово "повідомляють" матиме низьку вагу TF-IDF, оскільки воно часто зустрічається у багатьох новинах, тоді як специфічні терміни, такі як "економічна криза", отримають вищу вагу.

Недоліки методу TF-IDF

1. *Відсутність семантичного контексту:* TF-IDF не враховує значення слів у контексті, що може призводити до неправильної інтерпретації багатозначних слів.

Приклад. У реченнях "банк річки" і "банк фінансової установи" TF-IDF не розрізняє різні значення слова "банк", оскільки метод базується лише на частоті слів без урахування їхнього контексту.

2. *Чутливість до частоти:* якщо слово часто зустрічається у багатьох документах, його вага може бути знижена навіть у контексті документа, де воно є ключовим.

Приклад. Якщо слово "технології" часто зустрічається в наборі документів, його вага може бути знижена навіть у документах, де воно є основною темою.

ISSN 2788-6603

3. *Ігнорування порядку слів:* TF-IDF не враховує порядок слів у тексті, що може призводити до втрати сенсу під час оброблення текстів.

Приклад. Фрази "машинне навчання" і "навчання машинне" матимуть однакове значення TF-IDF, хоча порядок слів має значення для розуміння сенсу.

4. *Неадаптивність до нових слів:* TF-IDF може неадекватно оцінювати нові або рідкісні слова, оскільки метод потребує достатньої кількості документів для оцінювання частоти.

Приклад. Якщо нове слово, як-от "криптовалюта", з'являється вперше в новому документі, TF-IDF може не надати йому високої ваги, оскільки метод потребує достатньої кількості документів для адекватного оцінювання частоти.

- *Екстрактивний метод у вигляді алгоритму ранжування PageRank*

PageRank – це алгоритм, розроблений Ларрі Пейджем і Сергієм Бріном 1996 р., який використовують для оцінювання важливості вебсторінок у пошукових системах. Основна ідея PageRank полягає в тому, що важливість сторінки визначається не лише кількістю посилань на неї, але й важливістю сторінок, які на неї посилаються (Mihalcea, & Tarau, 2004).

Формула для розрахунку PageRank сторінки P виглядає так:

$$PR(P) = \frac{1-d}{N} + d\sum_{i=1}^{k}\frac{PR(P_i)}{L(P_i)}, \quad (3)$$

де $PR(P)$ – значення PageRank сторінки P; d – коефіцієнт затухання (зазвичай встановлюється на рівні 0,85); N – загальна кількість сторінок; P_i – сторінка, що посилається на сторінку P; $L(P_i)$ – кількість посилань, які виходять зі сторінки P_i.

Коефіцієнт затухання d відображає ймовірність того, що користувач продовжить переходити за посиланнями, а не завершить перегляд. Зазвичай d встановлюється на рівні 0,85, що означає, що користувач з імовірністю 85 % перейде за посиланням і з імовірністю 15 % почне перегляд нової сторінки.

Основні принципи PageRank

1. Графова модель:
   - Інтернет розглядається як граф, де вузли – це вебсторінки, а ребра – це гіперпосилання між сторінками.
2. Вага посилань:
   - Кожне посилання з однієї сторінки на іншу розглядається як "голос" за цю сторінку.
   - Важливіші сторінки отримують більше голосів, і, відповідно, їх PageRank зростає.
3. Розподіл ваги:
   - Кожна сторінка передає свою вагу (PageRank) сторінкам, на які вона посилається.
   - Вага розподіляється рівномірно між усіма вихідними посиланнями.
4. Ітеративний розрахунок:
   - PageRank обчислюється ітеративно, поки значення не закінчать збігання (перестануть значно змінюватися між ітераціями).

Переваги PageRank

1. *Об'єктивність оцінювання важливості:* PageRank оцінює важливість сторінки на основі її зв'язків з іншими сторінками. Це дозволяє алгоритму визначати важливість сторінок об'єктивно, без урахування їхнього вмісту.

Приклад. Уявімо два блоги: один має багато посилань від авторитетних новинних сайтів, інший – від особистих блогів. PageRank надасть перевагу першому блогу через високий авторитет його посилань, навіть якщо їхня кількість менша.

2. *Зважування якості посилань:* алгоритм враховує не лише кількість посилань, але й їхню якість. Посилання з авторитетних сторінок мають більшу вагу, ніж посилання з менш важливих сторінок.

Приклад. Сторінка, яка отримала посилання від урядового сайту, отримає більший приріст PageRank, ніж сторінка, яка отримала посилання від маловідомого блога.

3. *Зниження впливу спаму:* PageRank стійкий до маніпуляцій через спам-сайти, оскільки посилання з низькоякісних сторінок мають менший вплив.

Приклад. Якщо хтось створює багато фальшивих сайтів, що посилаються на основний сайт, PageRank основного сайту не зросте значно, тому що ці нові сайти мають низький PageRank.

4. *Гнучкість застосування:* алгоритм PageRank можна застосовувати не лише до вебсторінок, але й до інших структур, таких як наукові статті, соціальні мережі тощо.

Приклад. PageRank може використовуватися для оцінювання впливу наукових робіт на основі цитувань від інших наукових статей.

Недоліки PageRank

1. *Часові витрати на обчислення:* обчислення PageRank вимагає багато ітерацій для досягнення збіжності, що може бути часовитратним і ресурсомістким.

Приклад. Пошукова система, що індексує мільярди сторінок, потребує значного часу й обчислювальних ресурсів для регулярного оновлення значень PageRank.

2. *Проблеми з новими сторінками:* нові сторінки спочатку мають низький PageRank, навіть якщо вони містять якісний контент, через відсутність вхідних посилань.

Приклад. Новий блог з високоякісними статтями може довго залишатися непоміченим, поки він не отримає достатню кількість посилань від авторитетних джерел.

3. *Вразливість до посилальних ферм:* хоча PageRank стійкий до багатьох видів маніпуляцій, він може бути вразливий до посилальних ферм, де сайти взаємно посилаються один на одного для штучного підвищення їхнього PageRank.

Приклад. Група вебмайстрів може створити мережу сайтів, що взаємно посилаються один на одного, тим самим штучно підвищуючи їхній PageRank.

4. *Недостатня увага до контенту:* алгоритм зосереджується на зв'язках між сторінками, але не враховує якість їхнього контенту.

Приклад. Сторінка з великою кількістю вхідних посилань може мати високий PageRank, навіть якщо її контент застарілий або неякісний.

- *Екстрактивний метод на основі графів відомий як TextRank*

Методи на основі графів є потужним інструментом для аналізу тексту, структурування інформації та автоматичного реферування. Вони базуються на ідеї представлення тексту у вигляді графа, де вузли можуть представляти слова, речення або документи, а ребра – відношення між ними, такі як схожість, зв'язність чи послідовність.

TextRank – це один із популярних алгоритмів для автоматичного реферування текстів, який заснований на ідеях алгоритму PageRank, але адаптований для обробки тексту. Цей метод, як і PageRank, базується на ідеї графового представлення даних, де вузли представляють об'єкти (наприклад, слова або речення), а ребра – зв'язки між ними (Mihalcea, & Tarau, 2004).

Основні концепції TextRank

1. Графове представлення тексту:
   - Вузли: у методі TextRank текст представляється у вигляді графа, де вузли можуть бути або словами, або реченнями, залежно від задачі.
   - Ребра: ребра між вузлами відображають певний тип зв'язку між цими елементами тексту. Наприклад, для графа слів це може бути суміжність у тексті, а для графа речень – схожість у змісті.
2. Ранжування вузлів:
   - Подібно до PageRank, TextRank використовує ітеративний процес для ранжування вузлів на основі кількості та ваги посилань, що вказують на них. Чим більше важливих вузлів посилаються на даний вузол, тим більший ранг він отримує.
3. Збіжність ітерацій:
   - Процес ранжування триває, поки значення рангу не стабілізується (збіжність). На практиці, це означає, що після кількох ітерацій алгоритм припиняє роботу, коли зміни в ранжуванні вузлів стають мінімальними.

Переваги TextRank

1. *Мовна незалежність:* TextRank можна застосовувати до текстів різними мовами без необхідності модифікації алгоритму. Це можливо тому, що TextRank працює зі структурою тексту, а не з його змістом на семантичному рівні.

Приклад. Ви можете використовувати TextRank для автоматичного реферування статей на англійській, українській, китайській чи будь-якій іншій мові з мінімальними змінами. Алгоритм просто будує граф слів або речень і аналізує їхні зв'язки, незалежно від мови.

2. *Автоматичне визначення значущості:* TextRank автоматично визначає важливі елементи тексту (речення або слова) на основі їхньої зв'язності з іншими елементами, що дозволяє отримати об'єктивні результати без необхідності навчання на великих наборах даних.

Приклад. Якщо у вас є наукова стаття, TextRank може виділити основні тези і висновки, які часто згадуються у тексті та мають сильний зв'язок з іншими важливими реченнями.

3. *Простота реалізації:* TextRank є відносно простим для реалізації, оскільки він не вимагає попереднього навчання моделей або наявності великих обсягів анотованих даних.

Приклад. Упровадити TextRank у додаток для оброблення тексту можна за допомогою готових бібліотек або, якщо написати кілька рядків коду. Це особливо корисно для невеликих проєктів або стартапів, які не мають ресурсів для складних моделей машинного навчання.

4. *Застосування до різних задач:* TextRank можна використовувати як для реферування текстів, так і для екстракції ключових слів. Це універсальний інструмент для аналізу тексту.

Приклад. Для блога можна автоматично згенерувати короткий підсумок статті або виділити ключові слова для SEO-оптимізації, використовуючи той самий алгоритм.

Недоліки TextRank

1. *Чутливість до вибору метрик і налаштувань:* якість роботи TextRank залежить від вибору метрик для вимірювання схожості між елементами тексту й інших налаштувань алгоритму. Неправильний вибір може призвести до неточних результатів.

Приклад. Якщо використовувати невідповідну метрику схожості між реченнями, TextRank може виявити не ті речення як найважливіші, що призведе до некоректного реферату. Наприклад, речення з однаковими стоп-словами можуть бути помилково визнані схожими.

2. *Ігнорування глибокого контексту:* TextRank не враховує глибокі семантичні зв'язки між словами або реченнями, що може обмежити його ефективність для складних текстів.

Приклад. У художньому тексті, де важлива контекстуальна інформація і метафори, TextRank може не виділити ключові речення правильно, оскільки він не розуміє семантичний зміст тексту, а лише його поверхневу структуру.

3. *Обмеження у довгих текстах:* TextRank може погано працювати з дуже довгими текстами, де кількість зв'язків між елементами стає занадто великою, що може призвести до зниження ефективності та збільшення обчислювальної складності.

Приклад. Якщо спробувати застосувати TextRank до великого роману, алгоритм може неефективно визначити важливі частини тексту через велику кількість взаємозв'язків між реченнями, що робить реферат непослідовним або неповним.

4. *Залежність від суміжності:* TextRank часто визначає важливість елементів на основі їхнього суміжного положення в тексті. Це може бути проблематично, коли важливі ідеї розподілені у тексті нерівномірно.

Приклад. У статті, де ключова ідея розкидана по різних частинах тексту, TextRank може не з'єднати ці фрагменти в один важливий вузол, ігноруючи в такий спосіб основну думку.

Абстрактні методи

Абстрактні методи є складною та потужною технологією у сфері автоматичного реферування текстів. Вони відрізняються від екстрактивних методів тим, що не просто вибирають найважливіші фрази або речення з оригінального тексту, а створюють нові, узагальнювальні резюме, які можуть містити нові слова та структури речень (Moratanch, & Chitrakala, 2016). Ключові особливості абстрактних методів виглядають так.

1. *Генерація нового тексту:* абстрактні методи здатні створювати новий текст, який не просто повторює слова та фрази з вихідного документа, а передає його зміст у більш стислій і зручній для сприйняття формі.

ISSN 2788-6603

Це означає, що модель може перефразовувати, узагальнювати і навіть включати додатковий контекст для створення більш чіткого і зв'язного резюме. Наприклад, якщо в оригінальному тексті детально описується певна подія, абстрактний метод може згенерувати коротке резюме, яке передає сутність цієї події, без повторення всіх деталей.

Приклад. Якщо оригінальний текст містить кілька абзаців, що описують наукове дослідження і його результати, абстрактне резюме може виглядати так: "Дослідження показало, що новий метод лікування значно покращує виживаність пацієнтів із серцевою недостатністю".

2. *Глибоке розуміння контексту:* щоб створити змістовне резюме, абстрактні методи повинні розуміти зміст тексту на глибокому рівні, включаючи семантичні зв'язки між різними частинами тексту.

Ці методи часто використовують складні моделі машинного навчання, такі як трансформери або рекурентні нейронні мережі, які навчаються на величезних обсягах даних. Моделі здатні вивчати не тільки прямі зв'язки між словами, але й контекст, у якому вони вживаються, що дозволяє їм створювати більш узгоджені і зв'язні резюме.

Уявіть, що модель обробляє текст, де описують кілька взаємопов'язаних подій. Абстрактний підхід може зв'язати ці події в одне узагальнене резюме, що передає основну ідею, наприклад: "Ланцюг подій привів до значного зростання економіки країни".

3. *Використання нейронних мереж:* абстрактні методи часто базуються на нейронних мережах, зокрема і на моделях типу трансформерів, які є основою для багатьох сучасних систем оброблення природної мови (NLP).

Такі моделі, як GPT (Generative Pretrained Transformer) або BART (Bidirectional and Auto-Regressive Transformers), спочатку навчаються на великих обсягах тексту, щоб зрозуміти структуру і зміст природної мови. Після цього вони можуть генерувати новий текст, що відображає головні ідеї оригінального документа, але у стислій формі.

Використовуючи GPT, модель може проаналізувати статтю про нову технологію і згенерувати резюме: "Нова технологія штучного інтелекту покращує точність діагностики медичних зображень".

4. *Постпроцесинг для покращення якості:* після того, як текст генерується, він проходить етап постпроцесингу, щоб переконатися, що резюме відповідає граматичним, стилістичним і змістовим стандартам.

На цьому етапі можуть бути використані додаткові модулі або алгоритми, які перевіряють текст на узгодженість, логічність, коректність стилю, усувають можливі помилки або некоректності.

Якщо модель згенерувала резюме з неузгодженостями або орфографічними помилками, ці проблеми будуть виправлені на етапі постпроцесингу, щоб отримати кінцевий якісний генеративний (Generative) результат.

5. *Навчання на великих обсягах даних:* щоб досягти високої якості результатів, абстрактні методи потребують значних обсягів текстових даних для навчання.

Чим більше текстів використовують для навчання моделі, тим краще вона розуміє різні стилі письма, контексти та теми. Це дозволяє абстрактним методам створювати резюме, які є точнішими та змістовнішими.

Якщо модель навчалася на мільйонах статей і книг, вона буде ефективнішою у створенні резюме для нових текстів у різних галузях, таких як медицина, наука, література тощо.

Абстрактні методи автоматичного реферування тексту є потужним інструментом для створення стиснених резюме, які зберігають основний зміст оригіналу, але формулюються новими словами та фразами. Ці методи використовують складні нейронні мережі для аналізу і розуміння тексту, що дозволяє їм генерувати нові тексти, які є більш змістовними і компактними порівняно з вихідними. Однак такі методи потребують великих обчислювальних ресурсів і великих обсягів даних для навчання, що робить їхню реалізацію складнішою, ніж екстрактивні методи.

Трансформери для моделювання природної мови

Generative Pretrained Transformer

Generative Pretrained Transformer (генеративний трансформер із попереднім навчанням) – це потужна архітектура для моделювання природної мови, створена на основі трансформерів. Ця модель була розроблена компанією OpenAI і стала основою багатьох сучасних систем оброблення природної мови (NLP). Розглянемо детальніше ключові аспекти GPT (OpenAI, 2022):

1. *Архітектура трансформера:* модель GPT базується на архітектурі трансформера, яку запропоновано в статті "Attention is All You Need" (2017) (Vaswani et al., 2017). Трансформери використовують механізм уваги (attention) для оброблення послідовностей даних, що дозволяє їм ефективно враховувати контекст на всіх рівнях тексту.

Увага дозволяє моделі зосередитися на важливих частинах вхідних даних, що особливо важливо для завдань, які вимагають розуміння контексту. Наприклад, якщо модель аналізує речення, увага дозволяє їй зосередитися на тих словах, які є найбільш релевантними для поточного слова або фрази.

2. *Переднавчання (Pretraining):* модель GPT спочатку навчається на величезних обсягах текстових даних, використовуючи неконтрольоване навчання. Це означає, що модель просто намагається передбачити наступне слово в тексті, враховуючи попередні слова. Цей процес дозволяє моделі вивчити структуру мови, граматичні правила, семантику та загальні закономірності.

Переднавчання дозволяє моделі отримати базові знання про мову, що робить її здатною виконувати різні завдання оброблення природної мови після додаткового навчання.

3. *Генерація тексту:* після переднавчання GPT може використовуватися для генерування тексту. Коли модель отримує початковий текст (контекст), вона може передбачати та додавати нові слова, фрази або навіть абзаци, продовжуючи текст у стилі та темі вихідного матеріалу.

GPT здатна генерувати довгі та зв'язні тексти на основі лише кількох початкових слів або речень. Це робить її дуже корисною для завдань, які вимагають креативного підходу або автоматизації написання тексту, наприклад, для створення новин, оповідань, резюме або відповідей на запитання.

4. *Навчання з перенесенням (Transfer Learning):* після переднавчання модель може бути додатково налаштована (fine-tuned) на конкретні завдання, використовуючи менший набір даних. Наприклад, якщо потрібно створити модель для генерації юридичних текстів, модель GPT може бути навчена на корпусі юридичних документів.

Навчання з перенесенням дозволяє ефективно адаптувати модель до нових завдань, зберігаючи знання, отримані під час переднавчання. Це значно підвищує продуктивність моделі та робить її гнучкою для використання в різних галузях (Hosna et al., 2022).

5. *Шкала та потужність:* з кожною новою версією GPT (GPT-2, GPT-3, GPT-4) модель стає більшою, з більшою кількістю параметрів. Наприклад, GPT-3 має 175 мільярдів параметрів, що робить її однією з найбільших і найпотужніших моделей у сфері оброблення природної мови.

Велика кількість параметрів дозволяє моделі краще узагальнювати знання та точно виконувати завдання, навіть якщо вони суттєво відрізняються від того, на чому модель навчалася.

6. *Застосування:*

- *Чат-боти та віртуальні асистенти:* GPT використовують для створення чат-ботів, які можуть вести природні діалоги з користувачами, відповідаючи на запитання та допомагаючи з різними завданнями.
- *Автоматичне написання текстів:* модель може генерувати статті, резюме, електронні листи, технічну документацію та багато іншого.
- *Творче письмо:* GPT здатна створювати поеми, сценарії та інші творчі тексти на основі певної тематики або стилю.

Переваги GPT

1. *Здатність генерувати високоякісний текст:* GPT може генерувати зв'язний, змістовний і граматично правильний текст.

Приклад. GPT може створювати статті, що виглядають як написані людиною. Наприклад, на основі запиту "Опишіть значення штучного інтелекту в медицині" GPT може згенерувати детальну статтю з описом впливу ШІ на діагностику та лікування пацієнтів. Це робить GPT корисним для автоматизації написання текстів, таких як новини, блоги або резюме.

2. *Універсальність і гнучкість:* GPT може бути використаний у багатьох завданнях, як-от переклад, створення резюме, чат-боти, відповіді на запитання і багато іншого.

Приклад. У сценарії використання в чат-ботах GPT може підтримувати природну розмову з користувачами, відповідати на їхні запитання та надавати інформацію на основі введених даних. Одна модель може бути налаштована на різні завдання, що робить її надзвичайно гнучкою.

3. *Широке переднавчання:* GPT навчається на величезних обсягах текстових даних, що дозволяє їй отримати знання з багатьох галузей.

Приклад. GPT може генерувати текст на різноманітні теми – від науки до мистецтва. Наприклад, GPT може написати есе на тему філософії або технічний звіт про новітні досягнення у сфері квантових обчислень. Це робить GPT здатною адаптуватися до різних запитів і надавати релевантну інформацію в різних контекстах.

4. *Можливість налаштування на специфічні завдання:* GPT можна додатково налаштовувати на конкретні завдання або теми за допомогою навчання на менших, спеціалізованих наборах даних.

Приклад. Якщо потрібно, щоб GPT створювала юридичні документи, її можна налаштувати на наборі юридичних текстів, щоб модель стала експертом у цій галузі. Це дозволяє отримати більш точні та спеціалізовані результати в конкретних сферах.

Недоліки GPT

1. *Ризик генерації некоректної інформації:* GPT може генерувати текст, який виглядає правдоподібно, але містить фактичні помилки або некоректну інформацію.

Приклад. На запит "Скільки планет у Сонячній системі?" GPT може помилково згенерувати відповідь, що включає планети, яких не існує. Це робить GPT ненадійним у випадках, коли потрібна точна і перевірена інформація, особливо в наукових або технічних контекстах.

2. *Відсутність справжнього розуміння контексту:* GPT не має справжнього розуміння контексту або намірів, що може призводити до випадкових або недоречних відповідей.

Приклад. Якщо GPT запитати про етичні аспекти технологій, то система може згенерувати поверхневий або навіть безглуздий текст, оскільки модель не розуміє етичні концепції на глибинному рівні. Це може обмежувати здатність GPT генерувати текст, який вимагає глибокого розуміння складних концепцій або контекстів.

3. *Проблеми з упередженнями:* GPT може відображати упередження, які містяться в навчальних даних, оскільки вона навчалася на текстах, написаних людьми з різними поглядами.

Приклад. Якщо GPT навчена на текстах з інтернету, вона може відтворювати упереджені думки або стереотипи. Наприклад, GPT може генерувати тексти, що містять гендерні або расові стереотипи. Це створює ризик поширення дезінформації або закріплення упереджень, що є небажаним в етичному контексті використання технологій.

4. *Велика потреба в обчислювальних ресурсах:* GPT потребує значних обчислювальних ресурсів для навчання і використання, особливо коли йдеться про великі моделі, такі як GPT-3.

Приклад. Щоб розгорнути GPT-3 для генерації тексту в реальному часі, може знадобитися потужний сервер або хмарні обчислювальні ресурси. Це може бути дорогим і недоступним для деяких користувачів або організацій, що обмежує можливості широкого впровадження цієї технології.

Generative Pretrained Transformer має значні переваги, такі як здатність генерувати високоякісний текст, універсальність у різних завданнях, широке переднавчання і можливість налаштування. Водночас, модель має і недоліки, серед яких ризик генерації некоректної інформації, відсутність справжнього розуміння контексту, можливість упереджень і висока потреба в обчислювальних ресурсах. Ці плюси та мінуси роблять GPT потужним інструментом, але вимагають обережного й усвідомленого підходу до її використання.

Bidirectional and Auto-Regressive Transformers

Bidirectional and Auto-Regressive Transformers (BART) або Gemini – це нейромережна модель, розроблена компанією Facebook AI (тепер Meta AI), яка поєднує два підходи до трансформерів: двонаправлений (bidirectional) та авторегресивний (auto-regressive). BART створено як універсальну модель для оброблення природної мови (NLP), що може виконувати різні завдання, такі як генерація тексту, перефразування, автоматичне реферування тощо (Lewis et al., 2020).

Основні концепції BART

1. *Двонаправленість (Bidirectional):* двонаправлені трансформери, такі як BERT (Bidirectional Encoder Representations from Transformers), читають текст повністю, аналізуючи контекст як зліва направо, так і справа наліво. Це дозволяє моделі краще розуміти значення слів у контексті (Devlin et al., 2019).

Приклад. У реченні "Кішка сидить на підвіконні і дивиться у вікно", двонаправлена модель аналізує всі слова одночасно, щоб зрозуміти, що "кішка" – це суб'єкт дії, а слова "сидить" та "дивиться" описують її дії.

ISSN 2788-6603

2. *Авторегресивність (Auto-Regressive):* авторегресивні моделі, такі як GPT, генерують текст послідовно, прогнозуючи наступне слово на основі попередніх. Це означає, що на кожному кроці модель бере до уваги тільки ті слова, які вже були згенеровані.

Приклад. Якщо GPT генерує речення "Погода сьогодні чудова", то після слова "погода" модель передбачає, яке слово може бути наступним, на основі попередніх слів.

3. *Комбінація двох підходів у BART:* BART поєднує двонаправлений підхід, як у BERT, для кодування тексту, і авторегресивний підхід, як у GPT, для декодування тексту. Це робить модель універсальною та здатною до виконання широкого спектра завдань оброблення природної мови.

Приклад. Для задачі автоматичного реферування BART може використовувати двонаправлений підхід, щоб повністю зрозуміти вихідний текст, і авторегресивний підхід, щоб згенерувати короткий зміст тексту, використовуючи цю інформацію.

BART має енкодер-декодерну архітектуру, де:

- енкодер працює в двонаправленому режимі, аналізуючи контекст і зліва направо, і справа наліво.
- декодер працює в авторегресивному режимі, генеруючи текст послідовно, слово за словом.

Це дозволяє BART ефективно працювати як для завдань розуміння тексту (напр., класифікація тексту), так і для завдань генерації тексту (напр., машинний переклад або автоматичне реферування).

Переваги BART

1. *Універсальність:* BART є дуже гнучкою моделлю, здатною виконувати широкий спектр задач оброблення природної мови – автоматичне реферування, текстове перефразування, машинний переклад і заповнення пропусків у тексті.

Приклад. Якщо вам потрібно перефразувати текст, ви можете використовувати BART, і він дасть альтернативний, але збережений за змістом варіант. Наприклад, для фрази "Штучний інтелект змінює медицину" BART може запропонувати варіант "Інтелектуальні системи революціонізують охорону здоров'я".

2. *Висока точність:* завдяки двонаправленому енкодеру, який аналізує текст у повному контексті, і авторегресивному декодеру, який генерує текст послідовно, BART досягає високої точності в багатьох задачах.

Приклад. У задачі автоматичного реферування BART може виділити найбільш значущі фрагменти тексту для створення зведеного викладу. Наприклад, із довгої статті про зміну клімату модель може створити стислий, але інформативний реферат.

3. *Ефективність у розв'язанні різних задач:* оскільки BART поєднує двонаправлені й авторегресивні підходи, він ефективний у різних задачах генерації і розуміння тексту.

Приклад. BART може бути використаний як для задачі класифікації тексту (де важливо розуміти контекст), так і для генерації тексту (напр., написання короткого опису до заданого контенту).

4. *Можливість fine-tuning:* BART можна адаптувати до конкретних завдань або доменів додатковим навчанням на специфічних наборах даних, що робить його універсальнішим.

Приклад. Якщо вам потрібна модель для автоматичного створення підсумків юридичних документів, ви можете налаштувати BART на спеціалізованому наборі юридичних текстів, щоб досягти високої точності в цьому завданні.

Недоліки BART

1. *Великі обчислювальні ресурси:* як і інші великі трансформерні моделі, BART вимагає значних обчислювальних ресурсів для навчання і використання, що може бути проблематично для невеликих компаній або дослідницьких груп.

Приклад. Навчання BART на великому корпусі текстів може зайняти багато часу та потребує потужного обладнання, такого як графічні процесори (GPU) або тензорні процесори (TPU).

2. *Можливість генерації некоректної інформації:* як і інші моделі генерації тексту, BART може створювати змістовні, але некоректні або непослідовні фрагменти тексту, особливо якщо дані навчання були неякісними або незбалансованими.

Приклад. Якщо модель неправильно інтерпретує контекст або отримує суперечливу інформацію, вона може згенерувати текст, який не відповідає реальності. Наприклад, у задачі створення новинного заголовка BART може створити сенсаційний, але неправдивий заголовок.

3. *Ризик перенавчання на специфічних даних:* якщо BART надмірно налаштовувати на специфічний набір даних, модель може втратити здатність генералізувати, тобто погано працювати на даних, які відрізняються від тих, на яких вона навчалася.

Приклад. Якщо BART налаштований тільки на тексти наукової літератури, він може погано справлятися з генерацією текстів для популярних медіа або реклами.

4. *Складність налаштування й інтерпретації:* висока складність моделі ускладнює її налаштування та розуміння, що може створювати труднощі для дослідників та інженерів за адаптації BART для специфічних задач.

Приклад. Для правильного налаштування BART під задачу автоматичного перекладу необхідні спеціалізовані знання в галузі оброблення природної мови, що може бути достатньо проблематичним для команд без відповідного досвіду.

Метрики для оцінювання релевантності автоматично генерованих рефератів

Кількісні метрики оцінювання релевантності рефератів є важливим інструментом для об'єктивного вимірювання якості автоматично генерованих рефератів порівняно з еталонними, створеними вручну. Такі метрики дозволяють оцінювати, наскільки точно та повно автоматичний реферат передає зміст оригінального тексту. Розглянемо основні кількісні метрики детальніше.

- *Метрики типу Recall-Oriented Understudy for Gisting Evaluation*

Recall-Oriented Understudy for Gisting Evaluation **(ROUGE)** – це набір метрик, розроблених для оцінювання якості автоматичних рефератів та інших текстів, що генеруються алгоритмами. Цю метрику широко використовують в обробленні природної мови (NLP), оскільки вона дозволяє порівнювати автоматично створені тексти з еталонними рефератами, написаними людьми.

Основна ідея ROUGE полягає в тому, щоб виміряти кількість спільних елементів (зазвичай це слова або фрази) між автоматичним і еталонним рефератом. Метрика визначає, наскільки добре автоматично створений текст відображає зміст еталонного тексту (Lin, 2004).

Види ROUGE

Існує кілька різновидів ROUGE, кожен з яких призначений для оцінювання різних аспектів збігу між текстами:

ROUGE-N

- ROUGE-N вимірює кількість збігів n-грам між автоматичним і еталонним рефератами. N-грам – це послідовність з N елементів (напр., слів), які розглядають як одне ціле.
- ROUGE-1: вимірює збіг уніграм, тобто окремих слів. Це найпростіший і найчастіше використовуваний варіант ROUGE.
- ROUGE-2: вимірює збіг біграм, тобто послідовностей із двох слів.
- ROUGE-3 і більше: можливі також триграми та інші n-грами, але вони використовуються рідше.

Формула:

$$ROUGE\text{-}N = \frac{\sum_{gram_n \in Reference} Count\_match(gram_N)}{\sum_{gram_n \in Reference} Count(gram_N)}, \quad (4)$$

де *Count_match* – кількість збігів *n*-грам в автоматичному рефераті, *Count* – загальна кількість *n*-грам в еталонному рефераті.

Приклад. Якщо еталонний реферат містить фразу "Зміна клімату викликає глобальні проблеми", а автоматичний реферат містить "Кліматичні зміни викликають глобальні проблеми", то для ROUGE-1 збіги включатимуть слова "клімат", "глобальні", "проблеми".

ROUGE-L

ROUGE-L вимірює довжину найдовшої загальної підпослідовності (LCS, Longest Common Subsequence) між автоматичним і еталонним текстами. Цей підхід враховує і слова, і порядок їхнього розташування в тексті.

Формула:

$$ROUGE\text{-}L = \frac{LCS}{Довжина\ еталонного\ тексту}, \quad (5)$$

де *LCS* – довжина найдовшої підпослідовності, яка зустрічається і в автоматичному, і в еталонному рефераті.

Приклад. Якщо еталонний реферат містить фразу "Танення льодовиків призводить до підвищення рівня моря", а автоматичний реферат містить "Підвищення рівня моря викликано таненням льодовиків", то LCS включатиме "Танення льодовиків... рівня моря", що показує збіг послідовності слів.

ROUGE-S (ROUGE-Skip)

ROUGE-S вимірює збіг так званих "сплітних" біграм. Цей підхід дозволяє враховувати збіги, коли слова в біграмі можуть бути розділені іншими словами.

Формула:

$$ROUGE\text{-}N = \frac{\sum_{skip_2 \in Reference} Count\_match(skip_2)}{\sum_{skip_2 \in Reference} Count(skip_2)}, \quad (6)$$

де *skip_2* – це біграма зі словом або словами між ними, які дозволено пропускати.

Приклад. Якщо еталонний реферат містить фразу "Глобальне потепління швидко прогресує", а автоматичний реферат містить "Швидке прогресування глобального потепління", ROUGE-S врахує збіги фраз типу "глобальне... потепління".

ROUGE-W (Weighted ROUGE)

ROUGE-W є розширенням ROUGE-L, яке враховує вагу збігу залежно від довжини підпослідовності. Цей підхід надає більшу вагу довшим спільним підпослідовностям, ніж коротким збігам.

Формула: залежить від обраної ваги для різних підпослідовностей, що дозволяє точніше налаштовувати метрику під конкретні задачі.

Переваги ROUGE

1. *Простота використання*

ROUGE легко розрахувати й інтерпретувати. Він оцінює, наскільки добре автоматичний реферат збігається з еталонним, на основі простих підрахунків збігів слів або фраз.

Приклад. Якщо ми маємо автоматичний реферат "Сонце сходить на сході", а еталонний – "Схід сонця на сході", то ROUGE-1 покаже високий рівень збігу, оскільки більшість слів повторюються, навіть якщо порядок дещо інший.

2. *Широке використання та підтримка*

ROUGE став стандартом де-факто для оцінювання якості автоматичних рефератів та інших текстових задач. Він підтримується багатьма інструментами та програмними бібліотеками, що робить його доступним для широкого кола дослідників і розробників.

Приклад. Використання ROUGE дозволяє легко порівнювати результати різних досліджень і алгоритмів, оскільки метрика широко прийнята в науковому співтоваристві. Наприклад, у порівнянні двох моделей реферування можна використати ROUGE для демонстрації того, яка модель краще генерує тексти.

3. *Гнучкість*

ROUGE може бути налаштований для різних типів тексту і завдань. Наприклад, можна використовувати ROUGE-1 для загального оцінювання збігу слів, ROUGE-2 для оцінювання послідовностей слів і ROUGE-L для оцінювання збігу підпослідовностей.

Приклад. Якщо потрібно оцінити реферати, які складаються з коротких ключових фраз, краще використати ROUGE-1 і ROUGE-2. А для довших текстів, де важливо врахувати послідовність і зв'язність, можна застосувати ROUGE-L.

4. *Незалежність від мови*

ROUGE можна застосовувати для текстів будь-якою мовою, що робить його універсальним інструментом для оцінювання якості текстів.

Приклад. Незалежно від того, чи працюєте ви з англійським, українським або китайським текстом, ROUGE можна використовувати без змін, що робить його зручним для міжнародних досліджень.

ISSN 2788-6603

Недоліки ROUGE

1. *Чутливість до лексичних відмінностей*

ROUGE орієнтований на точні збіги слів або фраз. Він не враховує синоніми або перефразування, що може призводити до заниження оцінювання для текстів, які передають однаковий зміст, але використовують різні слова.

Приклад. Автоматичний реферат "Сонце піднімається на сході", а еталонний реферат "Схід сонця на сході". Хоча зміст однаковий, ROUGE може дати низьку оцінку через відмінності у використаних словах, таких як "піднімається" і "схід".

2. *Неврахування семантики*

ROUGE оцінює лише поверхневі збіги, не враховуючи глибший зміст тексту. Це означає, що навіть якщо два тексти мають однакові ключові слова, але передають різні ідеї, ROUGE може показати високу оцінку.

Приклад. Два реферати можуть містити ті самі слова, але в різному контексті. Наприклад, у першому рефераті йдеться про "економічне зростання в Китаї", а в другому – про "економічне зростання у світі", хоча збіги слів можуть бути високими, зміст текстів різний, але ROUGE цього не врахує.

3. *Ігнорування синтаксису та контексту*

ROUGE не враховує синтаксичні зв'язки або контекстуальні взаємодії між словами, що може бути критичним для оцінкювання якості тексту.

Приклад. У фразі "Кіт побачив собаку" і "Собака побачив кота" порядок слів змінює значення, але ROUGE оцінить їх однаково, оскільки збіги слів однакові, ігноруючи зміну контексту.

4. *Чутливість до довжини тексту*

ROUGE може бути упередженим щодо довших текстів, оскільки більші тексти мають більше шансів на збіг з еталонними текстами просто через велику кількість слів.

Приклад. Довший реферат, навіть якщо він містить багато зайвих слів, може отримати вищу оцінку ROUGE, ніж короткий і точний реферат, оскільки в довшому тексті більше можливостей для збігу з еталонним текстом.

- *Метрики типу Bilingual Evaluation Understudy*

Bilingual Evaluation Understudy (BLEU) – це одна з найпопулярніших і найвідоміших метрик для автоматичного оцінювання якості машинного перекладу. Вона використовується для кількісного вимірювання того, наскільки близький згенерований переклад (кандидат) до одного або кількох еталонних перекладів, створених людиною (Papineni et al., 2002). Наведемо детальніший розгляд основних принципів і компонентів BLEU.

Основні принципи BLEU

1. *Оцінювання на основі n-грам:* BLEU оцінює схожість між кандидатським перекладом і еталонним за допомогою порівняння *n*-грам (послідовностей з *n* слів). *N*-грамна точність означає, що для певного *n* ми підраховуємо, скільки *n*-грам із кандидатського перекладу наявні в еталонному перекладі. Для BLEU зазвичай використовують *n*-грами від 1 до 4.

2. *Використання модифікованої точності:* BLEU використовує модифіковану *n*-грамну точність (modified precision), що означає, що для кожної з *n*-грам підраховується максимальна кількість її повторень в еталонному перекладі, щоб уникнути переоцінювання повторень у кандидатському перекладі. Це допомагає запобігти маніпуляціям із результатами шляхом багаторазового повторення певних слів або фраз у кандидатському перекладі.

3. *Штраф за кратність (Brevity Penalty):* BLEU включає штраф за кратність (brevity penalty), щоб враховувати різницю в довжині між кандидатським і еталонним перекладами. Штраф застосовують, коли кандидатський переклад коротший за еталонний, що запобігає генерації занадто коротких перекладів, які можуть мати високий збіг у *n*-грамі, але втрачати важливі аспекти змісту.

Основні компоненти BLEU

n-грамна точність (*n*-gram precision):

- 1-грамна точність (unigram precision) оцінює збіги окремих слів між кандидатським і еталонним перекладом;
- 2-грамна точність (bigram precision) оцінює збіги пар слів;
- 3-грамна точність (trigram precision) оцінює збіги трійок слів;
- 4-грамна точність (4-gram precision) оцінює збіги четвірок слів.

Загальний BLEU-бал розраховують як геометричне середнє від точності для всіх *n*-грам, що дозволяє оцінити відповідність перекладу як на рівні окремих слів, так і на рівні коротких фраз.

1. *Модифікована точність (Modified Precision):* щоб уникнути надмірної оцінки кандидатського перекладу за рахунок повторення тих самих слів, BLEU використовує модифіковану точність. Для кожного *n*-грам розраховують, скільки разів слово зустрічається в еталонному перекладі, і цю кількість збігів використовують для обчислення точності. Наприклад, якщо в кандидатському перекладі слово "кіт" зустрічається тричі, а в еталонному лише один раз, то для розрахунку модифікованої точності враховують лише один збіг.

2. *Штраф за кратність (Brevity Penalty):* Brevity Penalty використовують для запобігання генерації надто коротких перекладів, які можуть досягати високих показників точності за рахунок скорочення довжини (Callison-Burch, Osborne, Koehn, & 2006).

Формула для розрахунку штрафу така:

$$BP = \begin{cases} 1, \text{ якщо } c > r, \\ e^{\left(1-\frac{r}{c}\right)}, \text{ якщо } c \le r, \end{cases} \tag{7}$$

де c – довжина кандидатського перекладу, а r – довжина еталонного перекладу. Якщо кандидат довший або дорівнює еталону, штраф не застосовують.

3. *Геометричне середнє точності (Geometric Mean of Precision):* остаточний BLEU-бал розраховують як геометричне середнє від точності для кожної з *n*-грам, помножене на штраф за кратність. Це середнє забезпечує пропорційність між точністю для різних *n*-грам, дозволяючи оцінювати як збіг окремих слів, так і збіг коротких фраз.

Переваги BLEU

1. *Автоматизованість:* BLEU дозволяє автоматично оцінювати якість перекладу без необхідності залучати людських експертів.

Приклад. Припустимо, ви маєте тисячу перекладених речень і бажаєте швидко оцінити їхню якість. Використовуючи BLEU, можна швидко порівняти кандидатські переклади з еталонними, отримавши оцінку точності без витрат на ручну перевірку.

2. *Масштабованість:* BLEU добре працює з великими обсягами текстів, дозволяючи легко обробляти й оцінювати масштабні корпуси перекладів.

Приклад. В контексті онлайн-сервісу машинного перекладу, де щодня обробляють мільйони текстів, BLEU дозволяє постійно моніторити якість перекладу без значних затрат на ресурси.

3. *Універсальність:* BLEU можна використовувати для оцінювання результатів перекладів різними мовами та для різних типів текстів, від технічної документації до художньої літератури.

Приклад. Незалежно від того, чи перекладає система новини з англійської на французьку, чи наукові статті з китайської на іспанську, BLEU можна застосовувати для оцінювання якості перекладу.

4. *Урахування збігів n-грам:* BLEU враховує не тільки збіги окремих слів (1-грам), але й коротких фраз (2-грам, 3-грам тощо), що дозволяє оцінити не лише правильність окремих слів, а й збереження контексту.

Приклад. Якщо кандидатський переклад "The cat on mat" збігається з еталонним "The cat is on the mat" у 3-грамі ("The cat on"), BLEU покаже, що кандидат майже правильно передав основну ідею, хоч і пропустив деякі деталі.

Недоліки BLEU

1. *Ігнорування контексту:* BLEU оцінює тільки поверхневі збіги *n*-грам, ігноруючи глибокий зміст, семантику і контекст перекладу.

Приклад. Якщо переклад "He gave her cat food" замінити на "He gave her food for the cat", BLEU може знизити оцінку через відсутність точних збігів, хоча обидва варіанти мають однаковий зміст.

2. *Чутливість до варіативності:* BLEU погано враховує різні можливі варіанти правильного перекладу. Якщо еталонний переклад використовує одні слова, а кандидат – інші синонімічні, BLEU може дати низьку оцінку.

Приклад. Еталонний переклад: "The quick brown fox jumps over the lazy dog". Кандидат: "The fast brown fox leaps over the lazy dog". Незважаючи на те, що обидва переклади мають однаковий зміст, BLEU може знизити оцінку через відсутність збігів для слів "quick" і "jumps".

3. *Штраф за кратність (Brevity Penalty):* штраф за кратність може несправедливо знижувати оцінку, якщо кандидатський переклад коротший за еталонний, навіть якщо коротший переклад є точним і достатнім.

Приклад. Еталонний переклад: "The cat is on the mat". Кандидат: "The cat on mat". Незважаючи на те, що цей переклад є зрозумілим, BLEU може знизити оцінку через відсутність деяких слів і штраф за коротку довжину.

4. *Не враховує граматичні та синтаксичні помилки:* BLEU не враховує якість граматики та синтаксису в перекладі, тому текст із правильною послідовністю *n*-грам, але з поганою граматикою може отримати високу оцінку.

Приклад. Кандидат: "The cat on the mat is sitting". Якщо це порівнювати з еталоном "The cat is sitting on the mat", BLEU може показати високу оцінку, попри зміни порядку слів, що можуть порушувати граматику.

5. *Проблеми з довжиною n-грам:* використання великих *n*-грам може знижувати оцінки через те, що дуже рідко зустрічаються довгі послідовності, навіть якщо загальний зміст передано правильно.

Приклад. Якщо кандидатський переклад "The quick fox" збігається з еталонним "The quick brown fox", BLEU знизить оцінку через відсутність 3-грамного збігу "quick brown fox", навіть якщо втрачене слово не є критичним.

Результати використання метрик оцінювання релевантності текстів для розглянутих моделей

- *Приклад застосування моделі TF-IDF*

Розглянемо набір з трьох документів:

- Документ 1: "Штучний інтелект змінює світ".
- Документ 2: "Штучний інтелект у медицині і технологіях".
- Документ 3: "Технології і медицина розвиваються завдяки штучному інтелекту".

Розрахунок TF (частота терміна):

- В документі 1 слово "штучний" з'являється 1 раз, "інтелект" 1 раз, "змінює" 1 раз, "світ" 1 раз. Загальна кількість слів у документі 4.
- В документі 2 слово "штучний" з'являється 1 раз, "інтелект" 1 раз, "медицина" 1 раз, "технології" 1 раз. Загальна кількість слів у документі 5.
- В документі 3 слово "технології" з'являється 1 раз, "медицина" 1 раз, "розвиваються" 1 раз, "штучному" 1 раз, "інтелекту" 1 раз. Загальна кількість слів у документі 7.

Розрахунок IDF (зворотна частота документа):

- Для слова "штучний": з'являється у 2 документах з 3, IDF = log(3/2) = 0,176.
- Для слова "інтелект": з'являється у 2 документах з 3, IDF = log(3/2) = 0,176.
- Для слова "медицина": з'являється у 2 документах з 3, IDF = log(3/2) = 0,176.
- Для слова "технології": з'являється у 2 документах з 3, IDF = log(3/2) = 0,176.
- Для слів "змінює", "світ", "розвиваються", "штучному", "інтелекту": з'являються по одному разу, IDF = log(3/1) = 0.477.

Розрахунок TF-IDF:

- Для слова "штучний" у документі 1: TF = 1/4 = 0,25; TF-IDF = 0,25 × 0,176 = 0,044.
- Для слова "інтелект" у документі 1: TF = 1/4 = 0,25; TF-IDF = 0,25 × 0,176 = 0,044.
- Для слова "змінює" в документі 1: TF = 1/4 = 0,25; TF-IDF = 0,25 × 0,477 = 0,119.
- Для слова "світ" у документі 1: TF = 1/4 = 0,25; TF-IDF = 0,25×0,477 = 0,119.

Цей приклад показує, як TF-IDF допомагає визначити важливість термінів у контексті документа, одночасно знижуючи вагу загальних слів і підвищуючи вагу рідкісних і значущих слів.

ISSN 2788-6603

- *Приклад застосування моделі PageRank*

Розглянемо простий приклад, щоб продемонструвати, як працює алгоритм PageRank. Припустимо, у нас є невелика мережа вебсторінок: *A*, *B*, *C*, і *D*, з такими посиланнями між ними:

- Сторінка A посилається на сторінку B.
- Сторінка B посилається на сторінки A і C.
- Сторінка C посилається на сторінку B.
- Сторінка D посилається на сторінки C і A.

Початкова ініціалізація

Припустимо, що ми маємо 4 сторінки. Спочатку всім сторінкам призначають однакове значення PageRank:

$$PR(A) = PR(B) = PR(C) = PR(D) = \frac{1}{4} = 0{,}25.$$

Для простоти, використовуємо коефіцієнт затухання d = 0,85.

Обчислення PageRank для сторінки A

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(D)}{L(D)}\right),$$

$$PR(A) = \frac{1-0{,}85}{4} + 0{,}85\left(\frac{0{,}25}{2} + \frac{0{,}25}{2}\right) = 0{,}25\ .$$

Обчислення PageRank для сторінки B

$$PR(B) = \frac{1-d}{N} + d\left(\frac{PR(A)}{L(A)} + \frac{PR(C)}{L(C)}\right),$$

$$PR(B) = \frac{1-0{,}85}{4} + 0{,}85\left(\frac{0{,}25}{1} + \frac{0{,}25}{1}\right) = 0{,}4625.$$

Обчислення PageRank для сторінки C

$$PR(C) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)}\right),$$

$$PR(C) = \frac{1-0{,}85}{4} + 0{,}85\left(\frac{0{,}25}{2}\right) = 0{,}14375.$$

Обчислення PageRank для сторінки D

$$PR(D) = \frac{1-d}{N} + d\left(\frac{PR(C)}{L(C)}\right),$$

$$PR(D) = \frac{1-0{,}85}{4} + 0{,}85\left(\frac{0{,}25}{1}\right) = 0{,}25.$$

Алгоритм PageRank вимагає багаторазових ітерацій для досягнення стабільності (збіжності). На практиці цей процес триває, поки зміни у значеннях PageRank між ітераціями не стають дуже малими.

Після декількох ітерацій (зазвичай десятків або сотень) алгоритм збігається, і ми отримуємо стабільні значення PageRank для всіх сторінок. У реальних сценаріях для великих мереж вебсторінок обчислення PageRank можуть бути складнішими, але основні принципи залишаються такими самими.

- *Приклад застосування TextRank*

Щоб продемонструвати роботу TextRank, розглянемо приклад автоматичного реферування тексту. Припустимо, у нас є такий текст:

"Сонце яскраво світило над морем, і хвилі спокійно накочувалися на берег. Птахи співали в гілках дерев, а вітер ніжно шелестів у листі. Мандрівник стояв на вершині скелі і дивився на простори, які розкинулися перед ним. Він відчував себе вільним і щасливим, адже дорога привела його до місця, де він завжди мріяв побувати. Спокій і краса цього місця заполонили його душу".

Кроки роботи TextRank

1. Побудова графа:
   - Кожне речення у тексті представляється як вузол у графі.
   - Зв'язки між реченнями визначаються на основі схожості (напр., на основі кількості спільних слів або схожих фраз).
2. Обчислення схожості між реченнями:
   - TextRank обчислює схожість між кожною парою речень (напр., якщо два речення мають спільні слова або схожі фрази, вони отримують вищий бал схожості).
3. Ранжування речень:
   - Алгоритм ітеративно обчислює "вагу" кожного речення, тобто визначає його важливість, з огляду на те, скільки інших речень із ним пов'язано та наскільки вони важливі.
4. Вибір найважливіших речень:
   - Після кількох ітерацій TextRank отримує ранжування речень. Найважливіші з них вибираються для формування реферату.

Результат та його інтерпретація. Після застосування TextRank до нашого тексту, наприклад, можуть бути відібрані такі речення:

"Мандрівник стояв на вершині скелі і дивився на простори, які розкинулися перед ним. Він відчував себе вільним і щасливим, адже дорога привела його до місця, де він завжди мріяв побувати."

TextRank вибрав саме ці речення, оскільки вони є центральними у тексті та найбільш зв'язані з іншими реченнями через ключові слова та ідеї, такі як "мандрівник", "щасливим", "мріяв", "місце". Інші речення, такі як ті, що описують природу ("Сонце яскраво світило над морем..."), були менш зв'язані із центральною ідеєю тексту, тому їх не включили до реферату.

Цей приклад показує, як TextRank автоматично відбирає найважливіші частини тексту для створення короткого реферату, зберігаючи основний зміст.

- *Приклад застосування GPT*

Уявімо, що вам потрібно створити коротку статтю про вплив штучного інтелекту на сучасну медицину. Ви можете використовувати GPT для автоматичного створення тексту на цю тему.

Запит і вихідний текст, згенерований GPT. "Напиши коротку статтю про те, як штучний інтелект змінює медицину".

Штучний інтелект: Революція в сучасній медицині. Штучний інтелект (ШІ) стає ключовим елементом у розвитку сучасної медицини, відкриваючи нові горизонти для діагностики, лікування та досліджень. Здатність ШІ аналізувати величезні обсяги даних дозволяє лікарям швидше і точніше виявляти захворювання на ранніх стадіях, що значно підвищує шанси на успішне лікування.

Одним із найяскравіших прикладів використання ШІ є розроблення алгоритмів для аналізу медичних зображень. Такі системи вже нині здатні діагностувати ракові утворення з точністю, що перевершує людські можливості. Крім того, ШІ допомагає персоналізувати лікування, створюючи індивідуальні терапевтичні плани на основі генетичних даних та історії хвороби пацієнта.

Не менш важливим є застосування ШІ у сфері фармакології. Завдяки моделям машинного навчання, час розроблення нових ліків значно скорочується, а їхня ефективність підвищується. Наприклад, алгоритми можуть прогнозувати, як різні сполуки взаємодіятимуть із людським організмом, що дозволяє уникнути багатьох побічних ефектів.

Однак, незважаючи на всі переваги, використання ШІ в медицині стикається з певними викликами. Наприклад, питання конфіденційності даних пацієнтів та етичні аспекти впровадження нових технологій потребують ретельного вивчення й регулювання.

Загалом, штучний інтелект вже нині змінює медицину, робить її точнішою, ефективнішою та доступнішою. У майбутньому ці технології стануть ще більш інтегрованими в медичну практику, забезпечуючи нові можливості для здоров'я та добробуту людей.

Аналіз прикладу

- Структура: текст має чітку структуру зі вступом, основним змістом і висновком.
- Контент: модель згадує про конкретні сфери медицини, де ШІ вже активно застосовується, такі як діагностика та фармакологія.
- Зв'язність: текст добре зв'язаний, кожен абзац логічно випливає з попереднього.
- Загальна якість: модель генерує текст, який виглядає правдоподібно і може бути корисним у реальних сценаріях.

Цей приклад демонструє, як GPT може використовуватися для автоматичного створення текстів на задану тему, зберігаючи при цьому зв'язність, логічність і структурованість тексту.

- *Приклад застосування моделі BART*

BART можна застосовувати для різних задач оброблення тексту. Розглянемо приклад використання BART для **автоматичного реферування тексту** – це процес створення короткого викладу (реферату) довгого документа.

Сценарій. Уявімо, що ви працюєте з великим документом, який містить детальний огляд досліджень у галузі зміни клімату. Потрібно створити короткий реферат цього документа, щоб швидко передати основні ідеї читачам.

"Зміна клімату стала однією з найважливіших глобальних проблем нашого часу. Вчені прогнозують, що підвищення середньої температури на планеті призведе до значних екологічних і соціальних наслідків. Танення льодовиків, підвищення рівня моря, зміни в екосистемах і погодних умовах – це лише деякі з наслідків, які вплинуть на людство. Дослідження показують, що вразливі групи населення, зокрема жителі прибережних регіонів і країн, що розвиваються, будуть найбільш вразливі до цих змін. Крім того, зміна клімату може негативно вплинути на світову економіку, сільське господарство і здоров'я людей".

Постановка завдання і результат, який був згенерований BART

Згенерувати короткий реферат цього тексту.

"Зміна клімату спричиняє серйозні екологічні та соціальні наслідки, такі як танення льодовиків, підвищення рівня моря і вплив на вразливі групи населення. Ці зміни також можуть мати негативний вплив на світову економіку і здоров'я людей".

Аналіз результату

- Стислий виклад: BART зміг виокремити ключові аспекти тексту, зокрема основні загрози, пов'язані зі зміною клімату.
- Збереження смислу: згенерований реферат зберігає основні ідеї, викладені в оригінальному тексті.
- Зв'язність: текст залишився логічним і зв'язним, що важливо для автоматичного реферування.

Такий підхід можна використовувати в новинних агенціях для створення коротких підсумків новинних статей, у наукових дослідженнях для автоматичного створення анотацій статей або в бізнесі для підготовки звітів на основі великих документів.

Переваги та недоліки цього прикладу

- Переваги: економія часу та зусиль під час створення коротких підсумків великих текстів.
- Недоліки: можлива втрата важливих деталей або генерація узагальнень, які не повністю відображають усі аспекти оригінального тексту.

Цей приклад демонструє, як BART може використовуватися для автоматичного створення коротких рефератів із довгих текстів, що є корисним у багатьох професійних і дослідницьких контекстах.

- *Приклад застосування моделі ROUGE*

Давайте додамо докладні розрахунки для кожної метрики ROUGE в наведеному прикладі.

ISSN 2788-6603

Вхідні дані

- Оригінальний текст статті: "Сьогодні вранці у центрі Києва стався потужний вибух. За попередніми даними, постраждало кілька людей. Вибух стався біля головного офісу відомої компанії. На місце події прибули рятувальники та поліція. Причини вибуху наразі з'ясовуються".
- Еталонний реферат (створений людиною): "Вибух у центрі Києва: постраждало кілька людей, причини з'ясовуються".
- Автоматичний реферат (згенерований системою): "Потужний вибух у Києві: постраждало кілька людей, причини невідомі".

Застосування ROUGE

ROUGE-1 (уніграми)

Ціль: оцінити кількість збігів окремих слів між автоматичним рефератом і еталонним.

- Слова в еталонному рефераті: "вибух", "у", "центрі", "Києва", "постраждало", "кілька", "людей", "причини", "з'ясовуються".
- Слова в автоматичному рефераті: "потужний", "вибух", "у", "Києві", "постраждало", "кілька", "людей", "причини", "невідомі".

Спільні слова між еталонним і автоматичним рефератами:

- "Вибух", "у", "постраждало", "кілька", "людей", "причини".

Розрахунок Precision (точність):

- Спільні уніграми: 6.
- Загальна кількість уніграмів в автоматичному рефераті: 9.
- ROUGE-1 Precision = 6/9 ≈ 0,67.

Розрахунок Recall (повнота):

- Спільні уніграми: 6.
- Загальна кількість уніграмів в еталонному рефераті: 9.
- ROUGE-1 Recall = 6/9 ≈ 0.67.

Розрахунок F1-міри (Christen, Hand, & Kirielle, 2024):

- ROUGE-1 F1 = 2 × (Precision × Recall) / (Precision + Recall).
- ROUGE-1 F1 ≈ 2 × (0,67 * 0,67) / (0,67 + 0,67) ≈ 0,67.

ROUGE-2 (біграми)

Ціль: оцінити кількість збігів пар слів (біграм) між автоматичним рефератом і еталонним.

- Біграми в еталонному рефераті: "вибух у", "у центрі", "центр Києва", "Києва постраждало", "постраждало кілька", "кілька людей", "людей причини", "причини з'ясовуються".
- Біграми в автоматичному рефераті: "потужний вибух", "вибух у", "у Києві", "Києві постраждало", "постраждало кілька", "кілька людей", "людей причини", "причини невідомі".

Спільні біграми між еталонним і автоматичним рефератами:

- "Вибух у", "постраждало кілька", "кілька людей", "людей причини".

Розрахунок Precision (точність):

- Спільні біграми: 4.
- Загальна кількість біграмів в автоматичному рефераті: 8.
- ROUGE-2 Precision = 4/8 = 0,50.

Розрахунок Recall (повнота):

- Спільні біграми: 4.
- Загальна кількість біграм в еталонному рефераті: 8.
- ROUGE-2 Recall = 4/8 = 0,50.

Розрахунок F1-міри:

- ROUGE-2 F1 = 2 * (Precision × Recall) / (Precision + Recall).
- ROUGE-2 F1 = 2 * (0,50 × 0,50) / (0,50 + 0,50) = 0,50.

ROUGE-L (довгі спільні підпослідовності)

Ціль: Оцінити довжину найдовшої спільної підпослідовності слів, що зберігає порядок слів.

- Еталонний реферат: "Вибух у центрі Києва: постраждало кілька людей, причини з'ясовуються".
- Автоматичний реферат: "Потужний вибух у Києві: постраждало кілька людей, причини невідомі".

Найдовша спільна підпослідовність:

- "Вибух у Києві постраждало кілька людей причини".

Розрахунок Precision (точність):

- Довжина спільної підпослідовності: 7 слів.
- Загальна кількість слів в автоматичному рефераті: 9.
- ROUGE-L Precision = 7/9 ≈ 0,78.

Розрахунок Recall (повнота):

- Довжина спільної підпослідовності: 7 слів.
- Загальна кількість слів в еталонному рефераті: 9.
- ROUGE-L Recall = 7/9 ≈ 0,78.

Розрахунок F1-міри:

- ROUGE-L F1 = 2 × (Precision × Recall) / (Precision + Recall).
- ROUGE-L F1 ≈ 0,78.

- *Приклад застосування моделі BLEU*

Наведемо приклад застосування метрики BLEU для оцінювання якості машинного перекладу з розрахунками.

Уявімо, що маємо еталонний переклад і кандидатський переклад від системи машинного перекладу. Ми хочемо оцінити якість кандидатського перекладу за допомогою метрики BLEU.

Еталонний переклад: "The quick brown fox jumps over the lazy dog".

Кандидатський переклад: "The fast brown fox jumps over the lazy dog".

Розрахунок n-грам

Метрика BLEU порівнює *n*-грам кандидатського перекладу з *n*-грамами еталонного перекладу. Давайте почнемо з розрахунку 1-грам (одиночних слів), а потім перейдемо до 2-грам (пари слів).

1-грам:

- Еталонний переклад: ["The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog"].
- Кандидатський переклад: ["The", "fast", "brown", "fox", "jumps", "over", "the", "lazy", "dog"].

2-грам:

- Еталонний переклад: ["The quick", "quick brown", "brown fox", "fox jumps", "jumps over", "over the", "the lazy", "lazy dog"].
- Кандидатський переклад: ["The fast", "fast brown", "brown fox", "fox jumps", "jumps over", "over the", "the lazy", "lazy dog"].

Підрахунок збігів n-грам

1-грам:

- Кількість збігів: ["The", "brown", "fox", "jumps", "over", "the", "lazy", "dog"] = 8 збігів із 9 (тут "fast" не збігається з "quick").
- Загальна кількість 1-грам у кандидатському перекладі: 9.

2-грам:

- Кількість збігів: ["brown fox", "fox jumps", "jumps over", "over the", "the lazy", "lazy dog"] = 6 збігів із 8.
- Загальна кількість 2-грам у кандидатському перекладі: 8.

Обчислення точності для n-грам

1-грамна точність (p_1):

$$p_1 = \frac{\textit{Кількість збігів 1-грам}}{\textit{Загальна кількість 1-грам у кандидаті}} = \frac{8}{9} \approx 0{,}889.$$

2-грамна точність (p_2):

$$p_2 = \frac{\textit{Кількість збігів 2-грам}}{\textit{Загальна кількість 2-грам у кандидаті}} = \frac{6}{8} = 0{,}75.$$

Обчислення середньої геометричної точності

Метрика BLEU зазвичай використовує середню геометричну точність для різних *n*-грам. Для простоти візьмемо тільки 1-грамну і 2-грамну точності:

$$\textit{Середня геометрична точність} = \sqrt{p_1 \times p_2} = \sqrt{0{,}889 \times 0{,}75} \approx 0{,}816.$$

Штраф за кратність (Brevity Penalty)

BLEU включає штраф за кратність для запобігання надмірно коротким перекладам. Якщо довжина кандидатського перекладу менша за довжину еталонного, застосовується штраф.

В нашому випадку, довжина кандидатського перекладу (9 слів) така сама, як і довжина еталонного перекладу (9 слів), тому штраф за кратність дорівнює 1.

Обчислення BLEU

Остаточна формула BLEU включає обчислену середню геометричну точність і штраф за кратність:

BLEU = Штраф за кратність * Середня геометрична точність.

Оскільки штраф за кратність дорівнює 1:

BLEU = 1 × 0,816 ≈ 0,816.

Отже, у цьому прикладі BLEU оцінка для кандидатського перекладу становить приблизно 0,816 або 81,6 %, що свідчить про високу схожість кандидатського перекладу з еталонним. Цей результат показує, що переклад є якісним, хоча є деякі незначні відмінності, такі як заміна "quick" на "fast".

Результати

Результати дослідження свідчать, що метрики ROUGE показують хорошу точність у вимірюванні збігів *n*-грам (послідовностей з *n* слів), тоді як BLEU ефективна у завданнях машинного перекладу, але може не враховувати деякі синтаксичні особливості тексту. Оцінювання методів автоматичного реферування за допомогою цих метрик показало, що екстрактивні методи реферування, такі як TF-IDF, є ефективними для оброблення простих текстів, але можуть втратити важливий контекст у складних текстах. PageRank і TextRank дозволяють враховувати зв'язки між реченнями, проте можуть давати менш релевантні результати для текстів із слабко вираженими структурними зв'язками. Абстрактні моделі GPT і BART забезпечують гнучкіший підхід до реферування, створюючи нові речення, що краще передають зміст, однак потребують значних обчислювальних ресурсів і складні у впровадженні.

Дискусія і висновки

Порівняльний аналіз результатів за допомогою метрик ROUGE і BLEU показав, що кожен підхід має свої сильні та слабкі сторони. Метрика ROUGE добре підходить для оцінювання екстрактивних методів, де важливо враховувати збіг *n*-грам з еталонним текстом. BLEU, зі свого боку, ефективна у задачах, де важливо зберігати структурні особливості перекладу або переказу, проте її застосування в абстрактних методах може бути обмеженим через недостатнє врахування контексту та синонімії.

Класичні підходи, такі як екстрактивне реферування, використовують різні статистичні та лінгвістичні методи для вибору найбільш значущих речень із тексту. Хоча саме класичні добре підходять для формальних і структурованих текстів, вони мають обмеження в передачі сенсу, особливо для текстів із складними взаємозв'язками між реченнями.

ISSN 2788-6603

TF-IDF є ефективним інструментом для виділення ключових термінів у тексті на основі їхньої частоти й унікальності. Однак цей метод не враховує контекст і не працює добре із синонімами, що може призводити до втрати сенсу у виділених реченнях.

PageRank є потужним інструментом для ранжування елементів на основі зв'язків між ними. У текстовому аналізі він допомагає визначити важливість речень через взаємозв'язки, але може мати проблеми з контекстуальним розумінням інформації, що іноді призводить до неправильних результатів.

Методи на основі графів, такі як TextRank, ефективно використовують структурні зв'язки в тексті для виявлення важливих елементів. Вони є гнучкими та підходять для багатьох типів текстів, проте вразливі до слабких зв'язків і можуть пропускати важливу інформацію в текстах зі складною структурою.

Абстрактні методи надають можливість створювати нові речення, що відображають загальний зміст тексту. Це робить їх гнучкішими порівняно з екстрактивними методами, але вони значно складніші у впровадженні та потребують великих обсягів навчальних даних для досягнення високої якості результату.

GPT є однією із провідних моделей для абстрактного реферування завдяки своїй здатності генерувати текст, що відображає глибокий зміст і контекст. Переваги включають високу якість і креативність генерованих текстів, але модель може іноді генерувати неправдиву або несуттєву інформацію, що є її основним недоліком.

BART поєднує переваги двох підходів: двонаправленого й авторегресивного, що дозволяє досягати високої точності в завданнях текстового реферування. Він особливо добре працює зі складними текстами, але вимагає значних обчислювальних ресурсів і обсягу навчальних даних.

Для оцінювання якості рефератів використовують кількісні метрики – ROUGE і BLEU, які забезпечують об'єктивні критерії для порівняння кандидатських рефератів з еталонними. Однак ці метрики мають свої обмеження, пов'язані з нездатністю повністю врахувати семантику і контекст тексту.

ROUGE є основною метрикою для оцінювання рефератів, яка враховує збіги *n*-грам між кандидатським і еталонним текстом. Це робить її корисною для оцінювання якості рефератів, але метрика погано враховує синоніми і варіативність у формулюваннях, що може знижувати її точність.

BLEU є потужною метрикою для оцінювання якості машинного перекладу, але її застосування обмежене нездатністю враховувати синоніми, контекст і граматику. Її результати слід інтерпретувати з урахуванням цих обмежень і, за можливості, доповнювати іншими методами оцінювання.

На основі проведеного аналізу можна надати такі рекомендації.

1. **Використання екстрактивних методів** доцільно для задач, де важлива швидкість оброблення текстів і коли тексти мають чітку структуру. Ці методи особливо корисні для оброблення великих обсягів даних, де необхідна простота реалізації та мінімальні вимоги до ресурсів.
2. **Абстрактні методи** є придатнішими для задач, де необхідна висока точність передачі змісту та контексту тексту. Вони рекомендовані для застосування в умовах, де є доступ до значних обчислювальних потужностей і де якість рефератів має пріоритетне значення.
3. **Поєднання обох підходів** може бути оптимальним рішенням для створення гібридних систем реферування, які можуть використовувати переваги екстрактивних і абстрактних методів залежно від специфіки текстів і вимог до кінцевого результату.

Отже, вибір методу для автоматичного реферування текстів повинен базуватися на конкретних вимогах до точності, швидкості та ресурсів, а також на особливостях самих текстів, що обробляються. Поєднання різних підходів і адаптація моделей під конкретні задачі дозволить отримати найкращі результати у створенні якісних і релевантних рефератів.

Внесок авторів: Олексій Кузнєцов – дослідження й аналіз методів та метрик, використаних у статті, написання частини статті; Геннадій Кисельов – огляд літературних джерел, розроблення висновків і рекомендацій, координація роботи, написання частини статті.

Список використаних джерел

Кузнєцов, О., & Кисельов, Г. (2022). Методи розпізнавання текстів та пошуку ключових слів для автоматичного реферування текстів. *Системні науки та інформатика: збірник доповідей І науково-практичної конференції "Системні науки та інформатика", 22–29 листопада 2022 року* (с. 331–335). Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського". https://ai.kpi.ua/ua/document2022.pdf

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. *In 11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 249–256). Trento, Italy. Association for Computational Linguistics. https://aclanthology.org/E06-1032/

Christen, P., J. Hand, D., & Kirielle, N. (2024). A review of the F-measure: Its History, Properties, Criti-cism, and Alternatives. *ACM Computing Surveys*, 56(3), 1–24. https://doi.org/10.1145/3606367

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1 (Long and Short Papers), (pp. 4171–4186). Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Hosna, A., Merry, E., Gyalmo, J., Alom, Z, Aung, Z., & Azim, M. A. (2022). Transfer learning: a friendly introduction. *Journal of Big Data,* 9(102). https://doi.org/10.1186/s40537-022-00652-w

Kuznietsov, O., & Kyselov, G. (2024). An overview of current issues in automatic text summarization of natural language using artificial intelligence methods. *Technology Audit and Production Reserves*, 4(78), 12–19. https://journals.uran.ua/tarp/article/view/309472

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,* online (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out (pp. 74–81). Barcelona, Spain. Association for Computational Linguistics. https://aclanthology.org/W04-1013.pdf

Lin, C.-Y., & Hovy, E. H. (2000). The Automated acquisition of topic signatures for text summarization. In *Proceedings of COLING-00*. Saarbrücken, Germany (pp. 495–501). International Committee on Computational Linguistics. https://doi.org/10.1007/978-3-540-74851-9_25

Mihalcea, R., & Tarau P. (2004). *TextRank: Bringing order into texts.* Association for Computational Linguistics. https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf

Moratanch, N., & Chitrakala, S. (2016). A survey on abstractive text summarization, 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT) (pp. 1–7). Nagercoil, India. Institute of Electrical and Electronics Engineers. https://ieeexplore.ieee.org/document/7530193

OpenAI. (2022). ChatGPT. https://openai.com/chatgpt/

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, (pp. 311–318). Association for Computational Linguistics. http://aclweb.org/anthology/P/P02/P02-1040.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, Aidan, N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems,* 30. Curran Associates, Inc. https://arxiv.org/abs/1706.03762

References

Callison-Burch, C., Osborne, M., Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. In 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy. Association for Computational Linguistics (pp. 249–256). https://aclanthology.org/E06-1032/

Christen, P., J. Hand, D., & Kirielle, N. (2024). A review of the *F*-measure: Its History, Properties, Criticism, and Alternatives. *ACM Computing Surveys*, 56(3), 1–24. https://doi.org/10.1145/3606367

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota. Association for Computational Linguistics, (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Hosna, A., Merry, E., Gyalmo, J., Alom, Z, Aung, Z., & Azim, M. A. (2022). Transfer learning: a friendly introduction. *Journal of Big Data,* 9(102). https://journals.uran.ua/tarp/article/view/309472

Kuznetsov, O., & Kyselyov, G. (2022). Methods of Text Recognition and Keyword Search for Automatic Text Summarization. System Sciences and Informatics: *Proceedings of the 1st Scientific and Practical Conference "System Sciences and Informatics", November 22–29, 2022* (pp. 331–335). National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" [in Ukrainian]. https://ai.kpi.ua/ua/document2022.pdf

Kuznietsov, O., & Kyselov, G. (2024). An overview of current issues in automatic text summarization of natural language using artificial intelligence methods. Technology Audit and Production Reserves, 4(78), 12–19. https://journals.uran.ua/tarp/article/view/309472

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, Online. (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out (pp. 74–81). Barcelona, Spain. Association for Computational Linguistics. https://aclanthology.org/W04-1013.pdf

Lin C.-Y., & Hovy E.H. (2000) The Automated acquisition of topic signatures for text summarization. In Proceedings of COLING-00. Saarbrücken, Germany. (pp. 495-501). International Committee on Computational Linguistics. https://doi.org/10.1007/978-3-540-74851-9_25

Mihalcea, R., & Tarau P. (2004). TextRank: Bringing order into texts. Association for Computational Linguistics. https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf

Moratanch, N, & Chitrakala, S. (2016). A survey on abstractive text summarization, 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, India, 2016, (pp. 1–7). Institute of Electrical and Electronics Engineers. https://ieeexplore.ieee.org/document/7530193/

OpenAI. (2022). ChatGPT. https://openai.com/chatgpt/

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, (pp. 311–318). Association for Computational Linguistics. http://aclweb.org/anthology/P/P02/P02-1040.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, Aidan, N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems,* 30. Curran Associates, Inc. https://arxiv.org/abs/1706.03762

Отримано редакцією журналу / Received: 15.08.24
Прорецензовано / Revised: 10.09.24
Схвалено до друку / Accepted: 22.09.24

Oleksii KUZNIETSOV, PhD Student
ORCID ID: 0000-0002-3537-9976
e-mail: oleksiy1908@gmail.com
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

Gennadiy KYSELOV, PhD (Engin.), Assoc. Prof.
ORCID ID: 0000-0003-2682-3593
e-mail: g.kyselov@gmail.com
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

USING AND ANALYSIS OF FORMAL METHODS FOR EVALUATING THE RELEVANCE OF AUTOMATICALLY GENERATED SUMMARIES OF INFORMATIONAL TEXTS

Background. ***The article reviews existing approaches to evaluating the quality of automatically generated summaries of informational texts. It provides an overview of automatic summarization methods, including classical approaches and modern models based on artificial intelligence. The review covers extractive summarization methods such as TF-IDF and PageRank, as well as graph-based methods, specifically TextRank. Special attention is given to abstractive approaches, including Generative Pretrained Transformer (GPT) and Bidirectional and Auto-Regressive Transformers (BART) models. The quality of generated summaries is evaluated using quantitative metrics of summary relevance, particularly ROUGE and BLEU.***

Methods. ***The article analyzes several approaches to automatic text summarization. Classical extractive methods, such as TF-IDF, calculate the importance of terms based on their frequency within a document and across a collection of documents. PageRank and TextRank utilize graph models to determine the significance of sentences based on the connections between them. Abstractive methods, such as GPT and BART, generate new sentences that succinctly convey the content of the original text. The effectiveness of each approach is assessed using ROUGE and BLEU metrics, which measure the overlap between automatically generated summaries and reference texts. Particular attention is given to analyzing their accuracy, flexibility, resource requirements, and ease of implementation.***

Results. ***The results of the study show that ROUGE metrics demonstrate good accuracy in measuring n-gram overlaps (sequences of n words), while BLEU is effective in machine translation tasks but may not account for certain syntactic features of the text. The evaluation of automatic summarization methods using these metrics revealed that extractive summarization methods, such as TF-IDF, are effective for processing simple texts but may lose important context in complex texts. PageRank and TextRank consider the connections between sentences but may produce less relevant results for texts with weak structural connections. Abstractive models like GPT and BART provide a more flexible approach to summarization, creating new sentences that better convey the meaning, though they require significant computational resources and are complex to implement.***

Conclusions. ***Combining classical and modern methods of automatic text summarization allows for achieving higher quality results. It is important to consider the specificity of the text and the requirements for the final outcome, adapting the selected approaches and metrics according to the task.***

Keywords: ***automatic summarization, extractive methods, abstractive methods, GPT, BART, ROUGE, BLEU, TextRank, PageRank, TF-IDF.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

ISSN 2788-6603