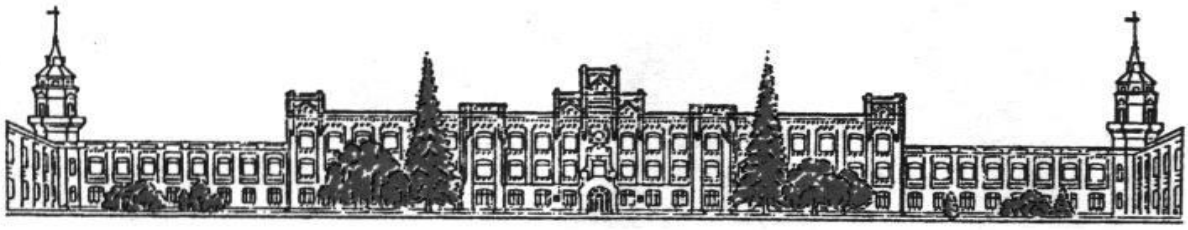


НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»

Інститут прикладного системного аналізу

Кафедра математичних методів системного аналізу

Есе

Тема: Методи розпізнавання текстів та пошуку ключових слів для автоматичного реферування текстів

Виконав:

Студент 2 курсу

групи КН-31Ф

Кузнєцов Олексій Андрійович

Перевірив:

Терентьєв О. М.

Київ – 2024

Вступ

В умовах сучасного інформаційного суспільства обсяг даних зростає експоненціально, що створює серйозні виклики для ефективного аналізу та обробки інформації. Однією з ключових задач стає швидке і точне виокремлення суттєвих даних з великих обсягів текстової інформації. Автоматичне реферування текстів, яке полягає у створенні стислого викладу основного змісту документа, стає важливим інструментом для розв'язання цієї проблеми.

Реферування тексту дозволяє зменшити обсяг інформації, зберігаючи при цьому ключові ідеї та факти, що значно полегшує процес прийняття рішень, навчання та наукових досліджень. Існує безліч методів автоматичного реферування, які використовуються для різних завдань, від створення анотацій наукових статей до автоматичного складання новинних зведень.

Ці методи можна розділити на дві основні категорії: добуваючі (extractive summarization) та абстрактні (abstractive summarization) підходи. Добуваючі методи полягають у виборі та поєднанні найбільш значущих речень з оригінального тексту, тоді як абстрактні методи передбачають створення нових речень, що узагальнюють основний зміст документа. Кожен з цих підходів має свої переваги та обмеження, які визначають вибір методу залежно від конкретного завдання.

Метою цього есе є аналіз основних методів автоматичного реферування текстів, їх ефективності та можливостей практичного застосування. Будуть розглянуті основні принципи добуваючих і абстрактних методів, їх реалізація, а також переваги і недоліки кожного підходу. Крім того, буде надано огляд сучасних досліджень у цій галузі, які спрямовані на вдосконалення існуючих методів та створення нових, більш ефективних алгоритмів автоматичного реферування текстів.

Основна частина

Основні підходи до автоматичного реферування тексту

Існує два головних підходи до автоматичного реферування тексту:

- Extractive summarization (добуваючий підхід);

- Abstractive summarization (абстрактний підхід).

Добуваючий підхід передбачає автоматичне анотування, яке ґрунтується на виділенні ключових фраз, слів або навіть цілих абзаців з первинних документів та додавання їх у вихідний документ без змін, у порядку їхньої появи в оригінальному тексті.

Абстрактний підхід передбачає автоматичне анотування, яке ґрунтується на виокремленні найважливішої інформації та навіть можливості створення нових текстів на основі узагальнених первинних документів. Цей метод, на відміну від добуваючого, дозволяє використовувати слова, яких не було у вихідному документі. Анотації, створені за допомогою цього підходу, схожі на ті, що пишуть люди. Проте реалізація цього методу є дуже складним завданням, оскільки модель має вирішувати складні проблеми, такі як семантичне представлення тексту та генерація природної мови. Тому, враховуючи складність цього підходу як у реалізації, так і в практичному застосуванні, а також значні обмеження, пов'язані з текстами, які можна реферувати за допомогою добуваючого методу, подальші дослідження методів реферування будуть проводитися в рамках добуваючого підходу.

Добуваючий підхід автоматичного реферування текстів

Наразі всі добуваючі методи автоматичного реферування складаються з трьох основних етапів[1]:

- Побудова проміжного представлення вхідного тексту: існують два основні підходи до створення такого представлення: тематичне представлення (topic representation) та індикаторне представлення (indicator representation). Тематичне представлення перетворює текст у проміжне представлення через інтерпретацію тем, що містяться в тексті. Індикаторне представлення анотує речення на основі переліку формальних ознак або індикаторів, які різняться залежно від алгоритму,

та використовує їх для ранжування тексту. До таких ознак належать довжина речення, його розташування у тексті, наявність ключових фраз тощо.

- Оцінка речень на основі проміжного представлення тексту: кожному реченню надається оцінка важливості для анотації. В різних підходах оцінка речення відображає, наскільки добре це речення передає найважливіші теми тексту.

- Формування підсумку: система обирає декілька найважливіших речень для створення кінцевої анотації. Для цього використовуються різні підходи: деякі методи застосовують жадібні алгоритми, тоді як інші перетворюють вибір речень на задачу оптимізації, метою якої є максимізація важливості та мінімізація надмірності слів.

Підходи з використанням тематичного представлення

Метод тематичних слів

Техніка тематичних слів є загальним підходом, основна ідея якого полягає у виділенні слів, що описують тему вхідного тексту. Тематичні слова можуть бути визначені різними способами. Наприклад, у роботі Луна[2] цей метод використовувався вперше для пошуку описових слів у документі за допомогою порогових значень частоти слів у тексті. У дослідженні Даннінга[3] застосовували логарифмічний алгоритм відношення правдоподібності для виявлення описових слів.

Оцінка важливості речення базується на розподілі інформативних слів у реченнях. Існує два способи обчислення важливості речення: як функція кількості тематичних слів, які воно містить, або як частка тематичних слів у одному реченні. Обидва методи оцінки речень стосуються одного й того ж подання теми, але можуть призначати різні оцінки різним реченням через особливості оцінювання. Перший метод може надавати вищі оцінки довшим реченням, оскільки вони містять більше слів, тоді як інший більше зосереджується на щільності тематичних слів. Найбільш поширеними методами в цій категорії є модель TF-IDF та лямбда-відношення правдоподібності і їх модифікації.

TF-IDF (term frequency — inverse document frequency, частота терму – зворотна частота документа) — це статистична міра, яка використовується для оцінки важливості термів у межах документа, що є частиною набору документів.

TF (term frequency – частота слова) — відношення кількості появи слова до загальної кількості слів у документі (формула 1). Таким чином, оцінюється важливість терма t в рамках окремого документа d.

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

де n_t – кількість входжень терма t в документ;

$\sum_k n_k$ – загальна кількість слів у документі.

IDF (inverse document frequency – зворотна частота документа) — інверсія частоти, з якою слово зустрічається в документах колекції (формула 2). IDF дозволяє зменшити оцінку часто вживаних слів, зосереджуючись на унікальних словах. Для кожного унікального слова в межах конкретного набору документів існує лише одне значення IDF. Більшу оцінку в TF-IDF отримують слова з високою частотою вживань у конкретному документі і низькою частотою вживань в інших документах.

$$idf(t, D) = \log (|D| / |\{d_i \in D \mid t \in d_i \}|) \quad (2)$$

де $|D|$ – кількість документів у колекції;

$|\{d_i \in D \mid t \in d_i \}|$ – кількість документів із колекції D, в яких зустрічається t (коли $n_t \neq 0$).

Лямбда-відношення правдоподібності (log-likelihood ratio) є логарифмом відношення ймовірності спостереження слова з однаковою ймовірністю в корпусі вхідних документів і корпусі відповідних їм резюме до ймовірності появи слова з різними ймовірностями в цих корпусах.

Метод заснований на центруванні речень

Головна ідея цього підходу полягає в припущенні, що найбільш цікава для анотації інформація міститься не лише в одному реченні. Він полягає у обчисленні «відстаней» між реченнями і у виборі тих з них, що в середньому знаходяться «ближче» до інших. Для визначення близькості речень використовуються алгоритми, засновані на наборах слів (Bag-of-words). Наприклад, найближчі в середньому речення можна визначити наступним чином:

- обчислити близькість між усіма парами речень за якимось параметром, наприклад, перекриттям змісту одного речення іншим;

- для кожного речення визначити середню близькість до інших речень;

- впорядкувати ці значення і вибрати з мінімальною близькістю (наприклад, використовуючи Латентний семантичний аналіз (LSA)).

Тематична модель, представлена у роботі Дірвестера та інших[4], є неконтрольованим методом вилучення прихованої семантики тексту на основі того, що слова з подібним значенням частіше зустрічаються разом, ніж окремо. Простіше кажучи, LSA бере текстові документи та відтворює їх у декількох різних частинах, де кожна частина виражається у різний спосіб погляду на значення тексту. Якщо уявити текстові дані як ідею, існувало б декілька різних способів розгляду цієї ідеї або кілька різних способів адаптації усього тексту.

Метод, представлений Гонгом та Лю[5], полягає у виборі одного речення для кожної з тем (рисунок 1), таким чином, щоб зберегти початкову кількість тем. Ця стратегія має недолік, оскільки обирається лише одне речення для кожної теми. Тому були запропоновані кілька альтернативних рішень для покращення ефективності методів узагальнення на основі LSA. Одним із покращень було використання ваги кожної теми, що надало гнучкість у варіативності кількості речень. Інше покращення полягає у тому, що автори зрозуміли[6], що речення, які обговорюють важливі теми, є хорошими кандидатами для узагальнення теми. Для

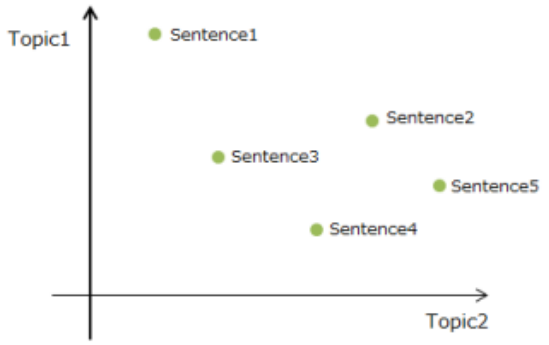
того, щоб знайти ці речення, вони визначили вагу речення за допомогою формули (формула 3):

$$(si) = \sqrt{\sum}^{m}{}_{j=1}\, d^2{}_{ij} \quad (3)$$

де g() – функція «зважування»;

m – кількість речень;

d_{ij} – вага теми і в реченні j.



	Topic1	Topic2
Sent1	0.9	0.25
Sent2	0.7	0.7
Sent3	0.45	0.4
Sent4	0.3	0.6
Sent5	0.4	0.9

Рис. 1 – Простий вибір основних речень, що представляють теми документу (LSA)[7]

Підходи з використанням індикаторного представлення

Підходи до подання індикаторів спрямовані на моделювання тексту як набору ознак, що використовуються для класифікації. До них належать методи на основі графів та техніки машинного навчання, які застосовуються для визначення речень, що мають бути включені у фінальне резюме. Основні ознаки, що використовуються в цих методах, включають:

- Речення на початку або в кінці тексту є більш інформативними;

- Занадто короткі або занадто довгі речення є менш інформативними;

- Наявність визначених сигнальних фраз або ключових слів;

- Слова з заголовків можуть свідчити про відношення речення до основної теми;

- Наявність емоційно забарвлених розділових знаків (знак питання, знак оклику, трикрапка тощо).

Методи на основі графів, впроваджені під впливом алгоритму PageRank[8], представляють документ як зв'язаний граф. Кожне речення утворює вершину графа, а вага ребер між реченнями вказує на зв'язок подібності між двома реченнями. Приклад подання тексту у вигляді графа можна побачити на рисунку 2. Подібність речень може вимірюватися як змістовним перекриттям між реченнями, так і за допомогою методу TF-IDF.

Графічна інтерпретація тексту має два основні результати. По-перше, кожен підграф є окремим розділом, пов'язаним однією темою. По-друге, ідентифікація важливих речень у документі базується на припущенні, що речення, пов'язані з багатьма іншими реченнями, швидше за все, є важливими та відображають тематику вхідного тексту. Тому їх потрібно включити у вихідний документ.

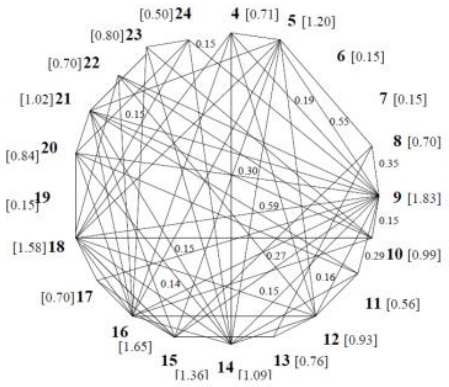


Рис. 2 – Графічне подання тексту у вигляді зваженого графу (речення представляють вершини з встановленою результуючою оцінкою)[9]

Підходи машинного навчання розглядають реферування як проблему класифікації. У дослідженні Купієка та інших[10] представлена рання спроба застосування технік машинного навчання для реферування. Автори розробили

класифікаційну функцію, засновану на наївному баєсовському класифікаторі. Ймовірності для класифікації визначалися за допомогою тренувальних даних з використанням правила Байеса (формула 4):

$$\mathrm{P}(s \in S|F_1, F_2, \dots, F_k) = \mathrm{P}(F_1, F_2, \dots, F_k|s \in S)P(s \in S) / P(F_1, F_2, \dots, F_k) \quad (4)$$

де s – це речення із колекції документів;

$F_1, F_2, \dots, F_k$ – ознаки, що використовуються для класифікації;

S – резюме, що має бути створено.

Отже, незважаючи на різноманітність методів і підтверджену ефективність кожного з них, вони мають суттєві недоліки. Наприклад, однією з основних проблем, крім складності деяких підходів, є те, що для цих алгоритмів потрібна спеціальна вибірка, і не кожна випадкова вибірка документів підходить для класифікації. Створення такої вибірки є складнішим завданням, ніж створення власноруч об’єднаних речень, переформульованих фраз чи нових речень. Навчальні дані також мають бути створені, і ефективність алгоритму повністю залежить від якості цих даних. Якби вдалося усунути ці недоліки та обмеження, можна було б використовувати ці алгоритми у ширшому спектрі областей, але наразі вони накладають значні обмеження на тексти.

Висновок

Автоматичне реферування текстів є важливою та актуальною технологією, яка дозволяє значно полегшити обробку великих обсягів інформації. У цьому есе було розглянуто два основних підходи до автоматичного реферування тексту: добуваючий та абстрактний.

Добуваючий підхід базується на виділенні ключових фраз, слів або абзаців з оригінального тексту та їх перенесенні до резюме без змін. Основні методи цього підходу включають техніку тематичних слів, метод центрування речень та індикаторне представлення. Вони забезпечують високу ефективність, проте мають обмеження щодо смислової цілісності тексту.

Абстрактний підхід передбачає створення нових текстів на основі узагальнення первинної інформації. Цей метод дозволяє створювати анотації, які наближаються до тих, що пишуть люди, але його реалізація є складнішою через необхідність вирішення проблем семантичного представлення тексту та генерації природної мови.

Методи на основі графів та машинного навчання дозволяють моделювати текст як набір ознак для класифікації речень, що підлягають включенню до підсумку. Ці методи використовують різноманітні алгоритми та підходи, такі як PageRank та наївний баєсівський класифікатор, проте потребують якісних навчальних даних для ефективної роботи.

Незважаючи на численні переваги та можливості застосування автоматичного реферування текстів, існують суттєві недоліки та обмеження, пов'язані з кожним із підходів. Подальші дослідження спрямовані на вдосконалення існуючих методів та розробку нових, більш ефективних алгоритмів для розширення спектру застосування автоматичного реферування текстів.

Список використаних джерел

1. A. Nenkova, K. McKeown. A survey of text summarization techniques.
2. Hans Peter Luhn. 1958. The automatic creation of literature abstracts. IBM Journal of research and development 2, 2 (1958), 159–165.
3. Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. Computational linguistics 19, 1 (1993), 61–74.
4. Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. JASIS 41, 6 (1990), 391–407
5. Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 19–25
6. Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Ježek. 2007. Two uses of anaphora resolution in summarization. Information Processing & Management 43, 6 (2007), 1663–1680.
7. Lee, R.S.T. Natural Language Processing: A Textbook with Python Implementations; Springer: Berlin/Heidelberg, Germany, 2023
8. Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. Association for Computational Linguistics.
9. Thakkar, K.S.; Dharaskar, R.V.; Chandak, M.B. Graph-Based Algorithms for Text Summarization. In Proceedings of the 2010 3rd International Conference on Emerging Trends in Engineering and Technology, Goa, India, 19–21 November 2010; pp. 516–519.
10. Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 68–73