УДК 004.89:81`32
DOI https://doi.org/10.32782/2663-5941/2024.2/26

Фант М.О.
Державний університет «Житомирська політехніка»

ПОРІВНЯЛЬНИЙ АНАЛІЗ НЕКОНТРОЛЬОВАНИХ МЕТОДІВ ОЦІНКИ ЕКСТРАКТИВНИХ АНОТАЦІЙ

У статті представлено аналіз неконтрольованих методів оцінки екстрактивних анотацій. У дослідженні використано шість загальновідомих та поширених алгоритмів генерування екстрактивних анотацій: Луна, дивергенції Кульбака – Лейблера, редукції, TextRank, LexRank, LSA. За допомогою обраних алгоритмів в ході дослідження згенеровано анотації для 1000 англомовних текстів, відібраних з Вікіпедії. Для кожного тексту для порівняльного аналізу додатково також згенеровано по одній анотації випадковим чином. Для оцінок згенерованих анотацій було відібрано чотири методи, які належать до двох груп: SummaQA-prob, SummaQA-fscore, BLANC-help, BLANC-tune. Проаналізовано стандартне відхилення числового значення оцінок однотипних анотацій, отриманих одним методом. Доведено надійність та послідовність методів дослідження на основі гомогенності стандартного відхилення поміж усіх чотирьох методів оцінки анотацій. Встановлено, що обидва методи BLANC показують кращі результати стандартного відхилення в порівнянні з методами SummaQA. Помічено особливість, що всі чотири методи оцінки анотацій показують високий ступінь стандартного відхилення на анотаціях, згенерованих алгоритмом LSA. Випадкові анотації отримують низьке значення стандартного відхилення передусім через показники методів SummaGA. Розроблено спосіб визначення стабільності методів оцінки анотацій. Цей спосіб базується на припущенні, що оцінка невипадкових анотацій має бути вищою за оцінку випадкових анотацій і чим вища така відмінність тим кращим є метод оцінки. За показником стабільності перевагу отримали також методи BLANC. Встановлено, що методи BLANC переважають методи SummaQA майже за всіма параметрами. У висновку підкреслено важливість вживання методів BLANC у парі, оскільки кожен з методів отримує свої переваги. Подальші дослідження в цьому напрямку вбачаються у порівняльному аналізі методів оцінки абстрактивних анотацій.

Ключові слова: *машинне навчання, обробка природної мови, екстрактивна анотація, мовна модель.*

Постановка проблеми. Сучасний етап розвитку суспільства характеризується величезним напливом інформації на кожного незалежно від роду і виду діяльності. Впродовж дня кожен з нас обробляє велику кількість усних та письмових текстів не тільки на роботі, а й на дозвіллі. Це спричиняє неабияке навантаження на пізнавальні можливості людського мозку і вимагає зменшення обсягів оброблюваних текстів без зниження їх інформаційного наповнення. В таких умовах особливого значення набувають інструменти для автоматичного анотування, які надають змогу стискати обсяги текстів завдяки скороченого викладу основної думки тексту.

В світлі окресленої проблематики важливо дати визначення поняттю *анотування*, а також виокремити підвиди анотацій. В цій статті під *анотуванням* ми розуміємо процес створення нового тексту на основі вхідного тексту, при чому обсяг нового тексту повинен бути істотно меншим за обсяг вхідного, а інформативність максимально збережена. В залежності від техніки створення анотації виокремлюють *екстрактивні* та *абстрактивні* анотації.

Екстрактивні анотації створюються в результаті відбору найважливіших і найінформативніших речень з вхідного тексту. При цьому анотація може не відповідати традиційному поняттю текст, а є швидше множиною речень. Порядок слідування цих речень зазвичай не має значення. В прикладному аспекті такі анотації зручно використовувати для позначення відібраних речень прямо у вхідному тексті, наприклад, шляхом підсвічування найважливіших речень на веб-сайті новин.

Абстрактивні анотації орієнтовані на створення читабельного текста, який перефразовує основний зміст вхідного тексту. Анотації такого типу найбільше розповсюджені в сфері науки, освіти та медіа. На відміну від екстрактивних анотацій важлива якісна характеристика абстрактивних анотацій – зв'язність мовлення у вихідному тексті.

Методи оцінки згенерованих анотацій можна умовно поділити на дві великі групи:

– ті що базуються на референтних анотаціях, створених людьми;

– повністю автоматизовані методи оцінки.

За аналогією до контрольованого і неконтрольованого машинного навчання, ми називаємо перший тип методів оцінки контрольованими, а другий – неконтрольованими. Варто детальніше розглянути алгоритми створення екстрактивних анотацій, а також методи їх неконтрольованої оцінки.

Аналіз останніх досліджень і публікацій. В теоретичних дослідженнях автоматичного анотування вчені в першу чергу звертають увагу на прагматичний аспект та виділяють такі корисні цілі анотування, як: зменшення часу прочитання; полегшення процесу відбору потрібної інформації з тексту; підвищення ефективності індексації; нівелювання суб'єктивності в порівнянні з анотаціями, створеними людьми вручну; корисні для створення систем запитання-відповідь [1-2]. Водночас, почасти поза фокусом уваги вчених залишається проблема оцінювання якості згенерованої анотації.

В дослідженні [3] проведено детальний та ґрунтовний огляд видів анотацій, створено розгалужену класифікацію анотацій: за принципом зіставлення, за типом вхідної колекції, за типом змісту, за розміром анотації, за потребою користувача. Водночас що стосується методів оцінки анотацій, то автори не вдавалися в детальний аналіз і використали лише метод оцінки якості ROGUE, за характеристиками Precision та Recall.

ROGUE – одна з найпопулярніших метрик для оцінки якості анотування, яка має три підтипи ROGUE-N, ROGUE-L, ROGUE-S. Спільна проблема сім'ї метрик ROGUE – це потреба у людській праці у вигляді створення опорних анотацій, з якими алгоритм метрики порівнює згенеровану анотацію. Так само на опорних анотаціях базуються: подібність косинусів, F-міра. Такі контрольовані методи оцінки анотацій не підходять при роботі з певними видами текстів, наприклад текстів малоресурсних мов з відсутніми або недоступними носіями, наприклад, історичних, мертвих або екзотичних мов.

В дослідженнях [4-8] відображено основні особливості і характеристики найпоширеніших і найуживаніших в сучасних дослідженнях і прикладних програмах алгоритмів анотування. Варто зазначити, що і в цих дослідженнях проблема оцінки якості анотації залишається мало освітленою. Зазвичай автори вдаються до загальних методів оцінки згенерованого тексту, а також до різних видів порівнянь анотацій з текстами референтних анотацій. Що стосується неконтрольованих методів оцінки, то вони маловживані.

В роботах [9-10] описано принципи роботи та використання неконтрольованих методів оцінки анотацій: SummaQA і BLANC. Суттєвою перевагою цих методів над контрольованими методами є те, що вони не потребують референтних текстів, для порівняння зі згенерованими анотаціями, що уможливлює їх використання, наприклад, для текстів мертвих або малоресурсних мов. Водночас, обидва методи описані безвідносно один до одного. Тому постає питання, з якими анотаціями і в яких випадках варто застосовувати кожен з них.

Постановка завдання. Метою роботи є порівняльний аналіз неконтрольованих методів оцінки екстрактивних анотацій.

Виклад основного матеріалу. Для проведення порівняльного аналізу неконтрольованих методів оцінки екстрактивних анотацій було обрано наступні алгоритми генерування анотацій:

– алгоритм Луна (LUHN) [4];

– алгоритм на основі дивергенції Кульбака – Лейблера (KL) [5];

– алгоритм редукції (RED);

– алгоритм TextRank (TR) [6];

– алгоритм LexRank (LR) [7];

– алгоритм LSA [8];

– випадковий підбір речень (RND).

Серед неконтрольованих методів оцінки екстрактивних анотацій ми обрали *SummaQA* і *BLANC*.

Оцінка анотації за методом SummaQA проводиться за двома значеннями: мірою ймовірності (SummaQA-prob) та F-мірою (SummaQA-fscore). Міра ймовірності виражає ступінь впевненості SummaQA в істинності виведеної відповіді до референтного питання. F-міра зазвичай використовується для оцінювання якості. Вона вимірює збіг між прогнозами та базовими відповідями.

В дослідженні використано дві метрики BLANC, а саме BLANC-help і BLANC-tune. Головна відмінність між цими двома метриками наступна: BLANC-help під час аналізу використовує текст анотації напряму поєднуючи його з кожним реченням вхідного тексту, в той час як BLANC-tune використовує анотацію для точного налаштування попередньо нетренованої моделі (в нашому випадку це модель BERT), а вже потім обробляє весь документ [10].

Для проведення порівняльного аналізу неконтрольованих методів оцінки екстрактив-

них анотацій було обрано 1000 текстів англійською мовою. Тексти представляють собою статті з англійської Вікіпедії. Було підібрано короткі статті розміром до 512 стем. Лімітацію по обсягу було введено задля економії часових і обчислювальних ресурсів, оскільки тексти більшого обсягу навряд чи могли б покращити якісні показники проведеного дослідження. Кожен з текстів перед анотуванням проходив попередню обробку, яка включала реченнєву та словесну токенізацію, видалення стоп-слів, стеммінг.

Розмір анотації не перевищував 40% оригінального тексту для всіх алгоритмів генерування анотацій. В якості мовної моделі для використання в методах оцінки анотацій було використано попередньо треновану модель BERT.

Важливим показником аналізу є показник *стандартного відхилення* методу оцінки для кожного алгоритму анотування. За допомогою стандартного відхилення можна визначити рівень розсіювання оцінок анотацій різних текстів, згенерованих одним алгоритмом. Іншими словами, стандартне відхилення дозволяє визначити рівень стабільності методу оцінки одного і того ж алгоритму.

Тобто для множини S всіх оцінок анотацій, згенерованих за допомогою однакового алгоритму стандартне відхилення σ вираховується за формулою некоригованого стандартного відхилення (1), де s_i – це i-та оцінка цієї множини, $\bar{s}$ – середня арифметична оцінка, а n – кількість всіх оцінок анотацій, згенерованих за допомогою однакового алгоритму. Ми використовуємо некориговане стандартне відхилення, оскільки розміри вибірки дозволяють виконати обрахунки на всій вибірці, в той час як кориговане стандартне відхилення рекомендують використовувати на частині вибірці задля уникнення девіацій.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left(s_i - \bar{s}\right)^2} \qquad (1)$$

Показник стабільності методу оцінки відповідає інтуїції, що більш надійний метод оцінки анотації повинен давати схожі результати при застосуванні до анотацій, згенерованих за допомогою однакового алгоритму.

На рис. 1 зображені значення стандартних відхилень для кожного з методів оцінки анотацій по відношенню до окремих алгоритмів генерування анотацій. На рисунку можна прослідкувати наступні важливі особливості:

1. Обидві групи методів оцінки анотацій мають схожу форму ліній, що свідчить про надійність та послідовність результатів дослідження,

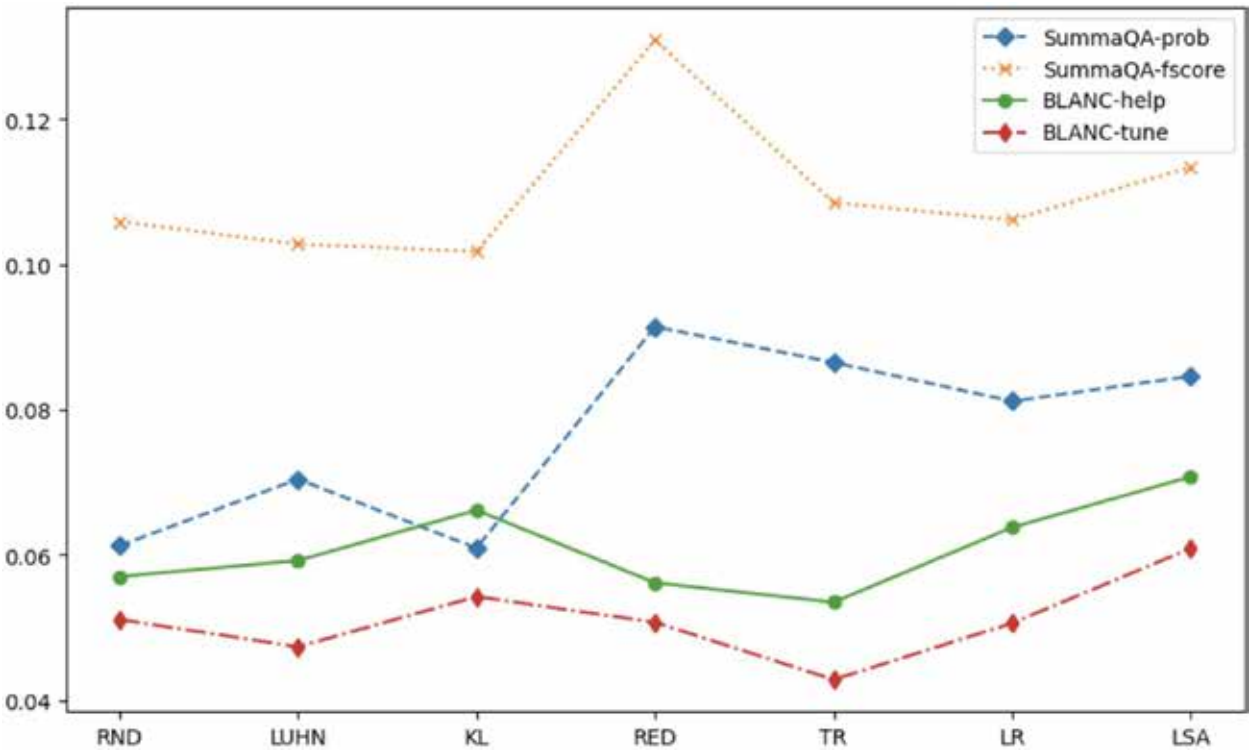

Рис. 1. Стандартне відхилення методів оцінки анотацій

адже зрозуміло, що споріднені методи оцінки мають подібним чином оцінювати анотації, згенеровані однаковим алгоритмом. Наприклад, ми бачимо, що група методів SummaQA показує найбільше розсіювання на алгоритмі редукції RED, а найменше – на алгоритмі дивергенції Кульбака – Лейблера KL. Водночас група методів BLANC найкраще впоралася з алгоритмом TR і найгірше з LSA.

2. Група методів BLANC послідовно показує менший рівень розсіювання на всіх алгоритмах. Виняток становить лише алгоритм KL, де метод оцінки BLANC-tune проявив вищий ступінь розсіювання аніж SummaQA-prob. Середні значення величини стандартного відхилення по методам оцінки анотацій можна побачити в Таблиці 1. За нею найкращим методом оцінки анотацій є BLANC-tune, на другому місці BLANC-help, а далі ідуть SummaQA-prob і SummaQA-fscore. Варто відмітити, що SummaQA-fscore суттєво відстає від інших методів.

3. Всі чотири методи оцінки анотацій показують підвищений ступінь розсіювання для алгоритму LSA, який вважається одним з найкращих. Це може бути спричинено вищою технічною складністю алгоритму LSA.

4. Анотації, згенеровані випадковим чином (RND), отримують зазвичай невисокий ступінь розсіювання, що також може свідчити про недосконаліть методу оцінки алгоритму. Тут варто виокремити метод SummaQA-prob, який показав в цілому невисокий ступінь розсіювання, але в якого випадкові анотації отримали дуже низький рівень стандартного відхилення в порівнянні з іншими методами. Водночас найкраще співвідношення між випадковими і невипадковими анотаціями має метод BLANC-tune, який крім того помітно виграє в загальному розподіленні стандартного відхилення.

5. Досить несподіваний результат можна отримати, якщо створити вибірку по середнім значенням стандартного відхилення оцінки анотацій за алгоритмами анотування, що показано в табл. 2. За цими показниками лідирують випадкові анотації з найменшим показником розсіювання. Найбільший показник розсіювання отримав алгоритм LSA. Проте якщо співставити ці результати з результатами на рис. 1, то стає зрозуміло, що девіація спричинена в першу чергу методами SummaQA, які оцінили випадкові анотації з меншим розсіюванням відносно до невипадкових анотацій.

Ще одним важливим показником є *стабільність* здатності відмежовувати анотації, згенеровані випадковим чином, від анотацій, згенерованих невипадковим методом. Під стабільністю методу оцінки анотування ми розуміємо середнє арифметичне значення відстаней від оцінки випадкової анотації (RND) певного тексту до середнього арифметичного значення оцінок всіх інших анотацій цього тексту.

Тобто для кожного тексту t_i множини всіх текстів T ми знаходимо середнє значення оцінок всіх анотацій крім випадкової анотації s_i, і віднімаємо від нього оцінку анотації, згенерованої випадковим чином $s_{i,\,RND}$. Загальний показник стабільності методу оцінки анотування – середнє арифметичне значення всіх різниць $s_i - e_{i,\,RND}$, як показано у формулі (2).

$$R = \frac{1}{n}\sum_{i=0}^{n}\left(\overline{s_i} - s_{i,\,RND}\right) \qquad (2)$$

Показник стабільності базується на припущенні, що всі алгоритми анотування мають створювати кращі анотації ніж випадкова. Таким чином оцінка всіх анотацій крім випадкової має бути вищою за неї. Показники стабільності методів оцінки анотацій зображено в табл. 3.

За результатами аналізу стабільності методів оцінки анотацій можна виокремити наступні співзвучні з попередніми закономірності:

1. Група методів BLANC показує вищий рівень стабільності порівняно з групою методів SummaQA. Проте варто зазначити, що на відміну від показників стандартного відхилення де, найкращим методом виявився BLANC-tune,

Таблиця 1

Середнє значення стандартного відхилення методів оцінки анотацій

SummaQA-prob	*SummaQA-fscore*	*BLANC-help*	*BLANC-tune*
0.076622	*0.109944*	0.060947	**0.05111**

Таблиця 2

Середнє значення стандартного відхилення оцінки анотацій за алгоритмами анотування

RND	*LUHN*	*KL*	*RED*	*TR*	*LR*	*LSA*
0.068823	0.069952	0.070793	0.082314	0.072834	0.075432	*0.082438*

Таблиця 3

Стабільність методів оцінки анотацій

SummaQA-prob	*SummaQA-fscore*	*BLANC-help*	*BLANC-tune*
0.016605	0.028529	**0.033215**	0.026254

а найгіршим SummaQA-fscore, за показником стабільності перше місце отримує BLANC-help, а останнє SummaQA-prob.

2. Різниця в оцінках між випадковими і невипадковими анотаціями є загалом невеликою, що частково пояснюється високим відсотком розміру згенерованих анотацій (40% від оригінального тексту), а також самим типом анотування – екстрактивним. Зрозуміло, що при абстрактивному анотуванні відмінність між випадковим і невипадковим текстом буде вищою.

Висновки. Проведено порівняльний аналіз неконтрольованих методів оцінки екстрактивних анотацій. Анотації згенеровано за допомогою семи різних алгоритмів: випадкового, Луна, дивергенції Кульбака – Лейблера, редукції, TextRank, LexRank, LSA. Оцінювання анотацій виконано за допомогою чотирьох методів, які належать до двох груп: SummaQA-prob, SummaQA-fscore, BLANC-help, BLANC-tune.

Для аналізу результатів дослідження розроблено власний розрахунок показника стабільності (формула 4), який базується на відмінності оцінки випадкових і невипадкових анотацій.

Виявлено, що група методів BLANC має перевагу за показником стандартного відхилення (найкращим виявився BLANC-tune) і за показником стабільності (найкращим виявився BLANC-help).

Що стосується екстрактивних алгоритмів, то варто зазначити, що алгоритм LSA, який вважається одним з найкращих послідовно отримує високий показник стандартного відхилення, що характеризує його з негативної точки зору. Водночас алгоритми TextRank і LexRank виділяються низьким рівнем стандартного відхилення в межах кожного методу оцінки.

Отже, для оптимальної оцінки екстрактивних анотацій варто використовувати методі оцінки BLANC, при чому бажано використовувати їх в парі, оскільки кожен з них має свої переваги (стандартне відхилення або стабільність). Вважаємо доцільним провести подальші дослідження в цьому напрямку з оцінкою абстрактивних анотацій.

Список літератури:

1. Archana A., Sunitha C. An overview on document summarization techniques, *International Journal on Advanced Computer Theory and Engineering*. 2013. Vol. 1. No. 2. Pp. 113–118.

2. Torres-Moreno J. Automatic Text Summarization. ISTE Ltd and John Wiley & Sons, Inc. 2014. 348 p.

3. Водолазький В., Холєв В., Росінський Д., Барковська О. Рішення задачі прискореного анотування текстових документів як елемент ЕБС. *Міжнародний науковий журнал «Грааль науки»*. № 6. 2021. С. 182-190.

4. Luhn H. P. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*. 1958. Vol. 2. Issue 2. Pp. 159–165.

5. Haghighi A., Vanderwende L. Exploring Content Models for Multi-Document Summarization. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2009. Pp. 362–370.

6. Mihalcea R., Tarau P. TextRank: Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. 2004. Pp. 404–411.

7. Erkan G., Radev D. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal Of Artificial Intelligence Research*. 2004. Vol. 22. Pp. 457–479.

8. Steinberger J., Ježek K. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. *Proceedings of ISIM*. 2004. Pp. 93–100.

9. Scialom Th., Lamprier S., Piwowarski B., Staiano J. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2019. Pp. 3246–3256.

10. Vasilyev O., Dharnidharka V., Bohannon J. Fill in the BLANC: Human-free quality estimation of document summaries. *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. 2020. Pp. 11–20.

Fant M.O. COMPARATIVE ANALYSIS OF UNSUPERVISED EVALUATION METHODS OF EXTRACTIVE SUMMARIZATIONS

The paper presents a comprehensive analysis of unsupervised evaluation methods of extractive summarizations. Six well-known and common algorithms for generating extractive summarizations were used in the study: Luhn, Kullback-Leibler divergence, reduction, TextRank, LexRank, LSA. With the help of selected algorithms, summarizations were generated for 1000 texts selected from English Wikipedia. For each text one summarization was additionally generated randomly. Four methods were selected for evaluation of the generated summarizations: SummaQA-prob, SummaQA-fscore, BLANC-help, BLANC-tune. The standard deviation of the numerical value of the estimates of the same type of summarizations obtained by one method was analysed. The reliability and consistency of research methods are proven based on the homogeneity of the standard deviation between all four summarization evaluation methods. Both BLANC methods are found to show better standard deviation results compared to SummaQA methods. A peculiarity was noticed that all four summarization evaluation methods show a high degree of standard deviation on the summarizations generated by the LSA algorithm. A method of determining the stability of summarization assessment methods has been developed. This method is based on the assumption that the evaluation of non-random summarizations should be higher than the evaluation of random summarizations, and the higher this difference, the better the evaluation method. The BLANC methods also gained an advantage in terms of stability. BLANC methods are found to outperform SummaQA methods in almost all parameters. The conclusion emphasises the importance of using BLANC methods in pairs, as each of the methods receives its advantages. Further research in this direction can be seen in the comparative analysis of abstract summarization evaluation methods.

Key words: *machine learning, natural language processing, extractive summarization, language model.*