Фант М. О., к.філол.н.,
Державний університет «Житомирська політехніка»

ОГЛЯД МЕТОДІВ ОЦІНКИ ТЕКСТОВИХ АНОТАЦІЙ

Огляд методів оцінки текстових анотацій – важливий інструмент для розвитку ефективних моделей автоматичного створення анотацій, а також для покращення якості інформаційних систем в цілому. Дослідження в цій області включає в себе різноманітні підходи та методики, орієнтовані на різні аспекти анотацій, такі як конкретність, повнота, послідовність та релевантність.

Найчастіше вживаними метриками для оцінки абстрактивних анотацій є метрики ROGUE. ROGUE використовується для порівняння згенерованих анотацій з референтними анотаціями, які створені людиною. Варто розрізняти між трьома підтипами цієї метрики: ROGUE-N, ROGUE-L, ROGUE-S. ROGUE-N оцінює схожість між n-грамами (послідовностями з n слів) в згенерованій анотації та референтною анотацією. ROGUE-N вимірює точність збігу n-грам між двома анотаціями. ROGUE-L вимірює схожість шляхом знаходження найдовшої спільної підпослідовності між згенерованою та референтною анотаціями. ROGUE-S оцінює схожість між словами в згенерованій та референтній анотаціях, використовуючи синоніми та семантичну схожість. ROGUE-S розглядає слова як ідентичні, якщо вони мають подібні синоніми або близький семантичний зміст.

Для певних задач, наприклад, у випадку з генерацією анотацій для текстів малоресурних мов, що характеризуються відсутністю мовних експертів, наприклад текстів історичних періодів розвитку конкретних природних мов важливо застосовувати методи оцінки, які не вимагають використання референтних анотацій. Серед таких методів виокремлюють Summa QA і BLANC.

Summa QA базується на прагматико-семантичному аспекті вхідного тексту, а саме на тому факторі, що кожен текст містить певну інформацію, на основі якої можна побудувати питання і відповіді до них. Такий підхід – близький до принципу створення анотації, адже анотація має містити відповіді на найістотніші питання, на які містить відповіді вхідний текст [1].

Постає питання, до яких саме частин тексту варто генерувати питання, щоб вони відображали найістотніші інформаційні характеристики. Дослідження показали, що найефективнішим підходом є генерування питань до іменованих сутностей (namedentity, NE) вхідного тексту. Таким чином, для кожної іменованої сутності NE_k вхідного тексту генерується трійка (I_k, Q_k, A_k), де I_k – це речення вхідного тексту, яке містить приховану NE_k, Q_k – питання до I_k, а A_k – відповідь на A_k, яка розкриває приховану NE_k.

Оцінка анотації за методом Summa QA проводиться за двома значеннями: мірою ймовірності (Summa QA-prob) та F-мірою (Summa QA-fscore).Міра ймовірності виражає ступінь впевненості Summa QA в істинності виведеної відповіді до референтного питання. Це відповідає, для кожної трійки, ймовірності істинної відповіді відповідно до моделі Summa QA. Оцінки ймовірності Summa QA-prob усереднюються для кожної анотації.

F-міра зазвичай використовується для оцінювання якості. Вона вимірює збіг між прогнозами та базовими відповідями. Для кожної анотації, яку потрібно оцінити, вираховується середнє значення оцінки F-міри, обчислене для кожної трійки.

BLANC-help і BLANC-tune – це дві різні версії BLANC, метрики для автоматичної оцінки якості машинного перекладу. BLANC-help: використовується для автоматичного оцінювання перекладу без додаткового налаштування або підгонки. Вона зазвичай базується на попередньо навчених моделях або правилах, які використовуються для порівняння перекладу з джерелом. BLANC-tune передбачає можливість налаштування метрики під конкретний корпус текстів або задачу перекладу. Вона може використовувати додаткові дані для підгонки параметрів метрики або для побудови моделей, що враховують специфіку конкретного набору даних.

Отже, серед методів оцінки анотацій можна виокремити три сім'ї: ROGUE, Summa QA і BLANC. Метрики ROGUE є найбільш вживаними та найпопулярнішими, проте їх використання вимагає наявність референтних анотацій. Summa QA і BLANC, в свою чергу, не вимагають використання референтних анотацій, проте внаслідок використання мовних моделей в межах цих методів потрібно зважати на підвищену ресурсоємність таких методів.

Список використаних джерел

1. Scialom Th., Lamprier S., Piwowarski B., Staiano J. Answers Unite! Un supervised Metrics for Reinforced Summarization Models. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conferenceon Natural Language Processing. – 2019. – PP. 3246–3256.
2. Vasilyev O., Dharnidharka V., Bohannon J. Fillinthe BLANC: Human-free quality estimation of document summaries. Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems – 2020. – PP. 11-20.