УДК 004.4:37

А. В. Снігур
В. В. Івасюк

АНАЛІЗ МЕТОДІВ АВТОМАТИЗОВАНОГО ОЦІНЮВАННЯ ТЕКСТОВИХ ВІДПОВІДЕЙ

Вінницький національний технічний університет

Анотація

У роботі проаналізовано сучасні методи автоматизованого оцінювання коротких відкритих текстових відповідей. Розглянуто лексичні та семантичні підходи зі зіставленням відповіді студента з еталоном. Визначено їхні обмеження та умови практичного застосування в освітніх системах.

Ключові слова: автоматизоване оцінювання, відкриті відповіді, семантична подібність, векторизація тексту, трансформери.

Abstract

The paper analyzes modern methods for automated grading of short open ended textual answers. Lexical matching and semantic comparison approaches between a student answer and a reference answer are considered. Their limitations and practical conditions of use in educational systems are identified.

Keywords: automated grading, open ended answers, semantic similarity, text vectorization, transformers.

Вступ

Дистанційне та змішане навчання перенесли значну частину перевірки в цифрові платформи, де відповідь, оцінка і коментар фіксуються в одному середовищі. Найбільшу складність оцінювання мають завдання з відкритою відповіддю, оскільки в них оцінюється зміст відповіді, а не факт вибору варіанта. При збільшенні кількості студентів ручна перевірка таких завдань потребує надзвичайно багато часу і результати оцінювання залежать від суб'єктивної думки викладача.

Автоматизоване оцінювання таких відповідей зазвичай зводиться до порівняння відповіді студента з еталоном. Підхід, що спирається лише на збіги слів, часто помиляється, коли правильна думка сформульована іншими словами. Тому на практиці застосовуються підходи, які оцінюють близькість саме за змістом через числовий опис тексту та розрахунок їх подібності. Для вибору придатного методу потрібен аналіз їх можливостей і обмежень у реальних сценаріях.

Результати дослідження

Коефіцієнт Жаккара є базовим методом лексичного порівняння, у якому еталон і відповідь подаються як множини слів, після чого обчислюється частка спільних елементів відносно їх об'єднання. Метод не потребує навчання, швидко обчислюється та дає прозорий результат. Водночас результат визначається тим, наскільки збігаються саме слова, а не смисл. Синоніми, граматичні варіанти та перестановка частин речення майже не відображаються в оцінці, тому змістовно правильні відповіді, подані іншими словами, часто отримують занижений показник схожості. Тому коефіцієнт Жаккара є доцільним як допоміжний механізм для контролю наявності ключових термінів і як швидкий попередній фільтр, але не як основний інструмент оцінювання відповіді [1].

Для підвищення змістової коректності оцінювання застосовуються семантичні підходи, у яких відповідь студента зіставляється з еталоном через близькість змісту. У таких підходах вирішальним є те, як саме текст переводиться в набір ознак. Практична реалізація спирається на векторні моделі, де кожен текст перетворюється на числовий вектор, а подібність визначається на основі близькості векторів еталону та відповіді. У межах аналізу ключовим є розмежування моделей, що формують ознаки з частот і статистики корпусу, та моделей, що формують контекстні представлення, чутливі до оточення слова у реченні [2].

Мішок слів є найпростішим варіантом векторного подання, у якому текст описується частотами термінів зі словника. Після виділення слів і приведення їх до єдиного вигляду формується ознаковий вектор фіксованої розмірності, де кожна координата відповідає конкретному терміну, а її значення

дорівнює кількості входжень цього терміна у відповіді. Вектор є розрідженим, тому обчислення подібності виконується швидко навіть для великої кількості відповідей. Оскільки порядок слів і синтаксичні зв'язки не фіксуються, різні слова з близьким змістом залишаються незалежними координатами. Це призводить до того, що правильне перефразування часто віддаляється від еталону без зміни реального змісту.

Зважування за схемою TF-IDF враховує не лише частоту терміна, а і його інформативність у межах колекції документів. Компонента TF відображає частоту терміна в конкретному документі, компонента IDF зменшує вагу термінів, що трапляються в багатьох документах, і підсилює терміни, що є рідкіснішими для корпусу. У результаті формується розріджений вектор ваг, який краще виділяє предметні поняття та зменшує вплив загальної лексики. Однак TF-IDF не формує семантичної близькості між різними словами, тому синонімія та варіативність формулювань продовжують знижувати подібність, а для відкритих відповідей це призводить до недооцінювання коректних перефразувань.

Латентно-семантичний аналіз (LSA) виконує узагальнення частотної інформації через побудову матриці термін документ і подальший сингулярний розклад, після чого зберігаються лише найбільш значущі компоненти. Завдяки переходу до скороченого семантичного простору тексти представляються не окремими словами, а комбінаціями латентних тем, тому різна лексика може давати близькі вектори, якщо вона використовувалась у подібних контекстах корпусу. Це дає підвищену стійкість до перефразування в межах однієї предметної області. Порівняно з TF-IDF це зазвичай краще працює на перефразуваннях у межах однієї предметної області. Обмеженням є залежність від корпусу та параметрів зменшення розмірності, а також слабке розрізнення багатозначності, оскільки різні значення одного терміна можуть перетворюватися в одні й ті самі латентні компоненти [3].

Word2Vec є нейромережевою моделлю розподілених представлень слів, яка навчається на великому текстовому корпусі та ставить у відповідність кожному слову щільний вектор фіксованої розмірності. Навчання виконується так, щоб слова, які з'являються в подібних контекстах, отримували близькі вектори, тому семантична спорідненість кодується через статистику співвживання. Зазвичай використовуються дві основні схеми. У моделі CBOW (Continuous Bag-of-Words) вхідними даними є контекстне вікно навколо слова, а цільовим значенням є саме слово, яке має бути відновлене. У моделі Skip-gram навпаки, на вході подається поточне слово, а мережею прогнозуються слова з його контекстного вікна, що дає змогу точніше налаштовувати вектори для рідкісних термінів, важливих у технічних дисциплінах. Для порівняння еталону з відповіддю потрібно отримати вектор не лише слова, а всього тексту, тому виконується агрегація векторів слів відповіді, найчастіше через усереднення або зважене усереднення. Унаслідок цього синонімічні заміни й близькі за змістом терміни рідше призводять до різкого падіння подібності, оскільки відповідні вектори розташовуються близько у просторі ознак. Кожне слово має один сталий вектор, тому різні значення слова в різних реченнях не розрізняються належним чином. Водночас слово має один сталий вектор, тому різні значення в різних реченнях відділяються слабко. Агрегація послаблює вплив порядку слів і синтаксичних залежностей, через що заперечення та уточнення можуть враховуватися недостатньо при оцінюванні коротких відповідей [4].

BERT (Bidirectional Encoder Representations from Transformers) є контекстною мовною моделлю на основі трансформерів, у якій представлення слова формується з урахуванням усіх слів у реченні. Завдяки механізму самоуваги модель враховує зв'язки між словами на всій довжині речення, а не лише в локальному контекстному вікні. Через це один і той самий термін у різних реченнях отримує різні векторні представлення, оскільки змінюється контекст, що істотно зменшує проблему багатозначності. На відміну від підходів зі сталим вектором слова, BERT здатний відображати вплив заперечення, уточнень і логічних зв'язків між словами, оскільки зміст кодується на рівні всього речення. Це підвищує стійкість до перефразувань і зменшує кількість ситуацій, коли схожість визначається лише повторенням термінів. Практичним обмеженням є висока обчислювальна складність, оскільки обробка кожного речення потребує повного проходу моделі, що ускладнює масштабування на великі потоки коротких відповідей без оптимізації [5].

Для систематизації та наочного порівняння розглянутих методів векторизації, їхні характеристики зведено в таблиці 1.

Після отримання векторних представлень оцінювання визначається вибором метрики, яка перетворює пару векторів на числовий показник подібності. Косинусна подібність вимірює кут між векторами, тому вона слабше залежить від довжини відповіді й краще підходить як для розріджених

TF-IDF подань, так і для щільних семантичних ембедингів. Евклідова відстань вимірює пряму відстань у просторі ознак, тому її значення суттєво залежить від масштабу координат і нормування векторів, інакше результат може відображати технічні властивості подання, а не зміст. Манхеттенська відстань є природною для розріджених векторів, де відмінності концентруються в окремих координатах, однак для щільних семантичних просторів частіше використовується косинусна подібність через стабільнішу поведінку при високій розмірності.

Таблиця 1 — Порівняльний аналіз моделей векторизації тексту

Критерій	Мішок слів / TF-IDF	LSA	Word2Vec	BERT / Трансформери
Основний принцип	Підрахунок частоти слів	Матриця термінів і тем	Навчання нейромережі	Механізм уваги
Врахування контексту	Ні	На рівні документа	Локальне	Глибоке та динамічне
Вирішення синонімії	Ні	Так	Так	Так
Вирішення полісемії	Ні	Ні	Частково	Так
Обчислювальна складність	Низька	Середня	Висока	Дуже висока

Окрему групу становлять підходи, у яких функція подібності не задається фіксованою метрикою, а навчається на прикладах. Сіамські нейронні мережі застосовують два однакові енкодери зі спільними вагами для еталону та відповіді, після чого оптимізується відстань між їх представленнями так, щоб коректні пари зближувалися, а некоректні віддалялися. Це дозволяє сформувати простір подібності, адаптований до конкретної дисципліни та типових формулювань, проте потребує розмічених пар для навчання і контролю узагальнення, інакше модель відтворює локальні шаблони замість змістових ознак правильності [6].

Висновки

У роботі виконано порівняльний аналіз підходів до автоматизованого оцінювання відкритих текстових відповідей. Проаналізовано методи на основі лексичного зіставлення та методи семантичного порівняння що орієнтовані на близькість змісту. Розглянуто методи векторного перетворення тексту та способи порівняння векторів. Проаналізовано, як перефразування, синонімія та багатозначність впливають на результати автоматизованого оцінювання. Визначено переваги й обмеження розглянутих рішень за точністю, інтерпретованістю та обчислювальною складністю що дає підстави обирати метод оцінювання відповідно до вимог навчального сценарію.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Jaccard Similarity [Електронний ресурс] — Режим доступу: https://www.geeksforgeeks.org/python/jaccard-similarity
2. Mohler M., Mihalcea R. Text-to-text Semantic Similarity for Automatic Short Answer Grading. 2009. 567–575.
3. What is latent semantic analysis? [Електронний ресурс] — Режим доступу: https://www.ibm.com/think/topics/latent-semantic-analysis
4. Understanding Word2Vec: CBOW & Skip-Gram [Електронний ресурс] — Режим доступу: https://www.kaggle.com/code/ftaham/understanding-word2vec-cbow-skip-gram
5. BERT Model — NLP [Електронний ресурс] — Режим доступу: https://www.geeksforgeeks.org/nlp/explanation-of-bert-model-nlp
6. Similarity Learning with Siamese Networks [Електронний ресурс] — Режим доступу: https://www.mygreatlearning.com/blog/siamese-networks/

Снігур Анатолій Васильович – к.т.н., доцент кафедри обчислювальної техніки, Вінницький національний технічний університет, Вінниця.

Івасюк Вадим Віталійович – студент групи 1КІ-24м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: vadim200339@gmail.com

Snigur Anatoliy Vasyliovych – Ph.D., Associate Professor of the Computer Engineering Department, Vinnytsia National Technical University, Vinnytsia.

Ivasiuk Vadym Vitaliyovych – student of group 1KI-24m, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: vadim200339@gmail.com