МЕТОДИ ОЦІНЮВАНЯ ЕФЕКТИВНОСТІ РОЗПІЗНАВАННЯ ДІАЛОГУ ТА СКЛАДАННЯ ПІДСУМКІВ ДІАЛОГУ З ЗАЛУЧЕННЯМ ШІ

***Голян В.В.**, доцент, кафедра ПІ, ХНУРЕ*
***Моторін Р.С.**, магістр, кафедра ПІ, ХНУРЕ*

У сучасному світі інформаційних технологій важливість розуміння та обробки природної мови невпинно зростає. Штучний інтелект (ШІ) відіграє ключову роль у цьому процесі, особливо у сферах розпізнавання діалогів та автоматичного складання підсумків. Здатність машин до точного розуміння та реагування на людську мову може радикально змінити багато аспектів нашого життя, від покращення користувацького досвіду до ефективнішого аналізу великих обсягів даних.

Значення цих технологій особливо відчутне в сфері обслуговування клієнтів, де автоматизація взаємодій з користувачами через чат-боти та інтерактивні голосові асистенти відкриває нові можливості для бізнесу. Інтеграція ШІ не тільки сприяє збільшенню швидкості відповідей на запитання клієнтів, але й значно покращує якість обслуговування завдяки здатності системи адаптуватися та вчитися на основі минулих інтеракцій. Таке постійне вдосконалення сприяє створенню більш особистісного та ефективного досвіду для кожного користувача.

Оцінка ефективності систем на базі ШІ для розпізнавання та аналізу діалогів зазвичай включає кілька ключових аспектів [1].

Точність (Precision): це частка правильно ідентифікованих інстанцій відносно всіх інстанцій, які система визначила як позитивні.

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

де TP – істинно позитивні результати, тобто випадки, коли система правильно ідентифікувала наявність певної характеристики;

FP – хибно позитивні результати, випадки, коли система помилково ідентифікувала наявність певної характеристики.

Повнота (Recall): це частка правильно ідентифікованих інстанцій відносно всіх позитивних інстанцій у даних.

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

де T– істинно позитивні результати;

FN (False Negatives) – хибно негативні результати.

F-міра (F-measure): гармонічне середнє між точністю та повнотою, яке допомагає збалансувати обидва показники.

$$F1 - 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{3}$$

Для оцінки складання підсумків часто використовують такі метрики, як ROUGE (Recall-Oriented Understudy for Gisting Evaluation), що дозволяє оцінити кількість спільних одиниць (слів, n-грам) між автоматично сгенерованим підсумком і еталонним підсумком.

Автоматичне розпізнавання діалогів та складання підсумків знайшли застосування у багатьох сферах, включаючи системи віртуальних помічників, аналітику в соціальних медіа, корпоративні бази знань та багато інших. Основними викликами є:

Обробка неоднозначності мови: іронія, жарти та культурні відсилання можуть істотно ускладнити обробку тексту.

Розуміння контексту: важливість контексту в діалозі, де зміст повідомлення залежить від попередніх реплік.

Адаптація до нових даних: здатність системи адаптуватися до нових слів, жаргону або навіть мов без необхідності великомасштабного перенавчання.

Одним із обіцяючих напрямків є використання глибокого навчання для створення ембедінгів діалогу, що зможуть краще уловлювати семантичні нюанси [2]. Також важливою є розробка нових метрик для оцінки різноманітних аспектів розуміння мови, включаючи емоційний зміст і сарказм.

Важливим аспектом у вдосконаленні систем обробки природної мови є здатність до роботи з даними, які виходять за рамки звичайного розподілу. Дослідження Hendrycks et al. (2020) підкреслює, як переднавчані трансформерні моделі, такі як BERT та GPT, покращують стійкість систем NLP у сценаріях з використанням даних out-of-distribution. Ці моделі ефективно впораються з різноманітними та неочікуваними типами даних, що робить їх незамінними у сучасних застосуваннях штучного інтелекту, де гнучкість та адаптивність є ключовими для успішної роботи системи. Це дослідження не тільки підтверджує значення переднавчання в моделях, але й вказує на потенціал для подальших покращень у роботі з складними випадками використання NLP [3].

Висновки. Оцінка методів розпізнавання діалогів та складання підсумків із залученням ШІ є критично важливою для забезпечення надійності та точності цих технологій. Розвиток нових методів і технологій обіцяє значне покращення в цій області, що, у свою чергу, може принести користь різним сферам нашого життя.

Список літератури

1. Recent Advances in NLP via Large Pre-Trained Language Models: This source provides an in-depth survey of modern pre-trained language models, including autoregressive models like GPT, masked language models like BERT, and encoder-decoder models such as BART and T5. It discusses their training sources, dataset sizes, and model parameters, offering a comprehensive overview of the evolution and capabilities of these models. https://ar5iv.labs.arxiv.org/html/2111.01243.

2. Fabbri et al. (2021). Summeval: Re-evaluating Summarization Evaluation. This paper, published in the Transactions of the Association for Computational Linguistics, offers a fresh perspective on evaluating summarization techniques in NLP. It underscores the importance of rethinking current evaluation methodologies to better assess the performance of summarization algorithms.

3. Hendrycks et al. (2020). Pretrained Transformers Improve Out-of-Distribution Robustness. This research highlights how pretrained transformer models, such as BERT and GPT, enhance the robustness of NLP systems, particularly in scenarios involving out-of-distribution data. It provides insights into the effectiveness of these models in handling diverse and unexpected data types.