

МОДЕЛЬ КАТЕГОРИЗАЦІЇ ЗАПИТІВ КОРИСТУВАЧІВ З ВИКОРИСТАННЯМ ТЕХНОЛОГІЙ NLP

Ільїн О.Ю., д.т.н., проф., oleg.ilin5454@gmail.com;

Горячев Т.В., магістрант, horiachev.tymur@gmail.com

*Державний університет інформаційно-комунікаційних технологій,
Київ, Україна*

У сучасному світі взаємодія людини та комп'ютера є невід'ємною частиною повсякденного життя, а бурхливий розвиток інформаційних технологій надає можливість постійної оптимізації цієї взаємодії. Сервіси пошуку інформації у мережі, автоматизованого перекладу, чат-боти та голосові помічники дозволяють звичайним користувачам взаємодіяти з ними використовуючи свою звичайну природню мову. Однак неформалізованість, залежність від контексту, часткова випадковість та інші особливості сутності природньої мови накладають певні обмеження, породжує складність її обробки та класифікації, що призводить до невиконання запитів кінцевих користувачів.

Традиційні методи аналізу тексту вже не є недостатніми, та великого значення набуває використання технологій обробки природньої мови (NLP). Машинне навчання та інші техніки аналізу тексту в поєднанні з NLP надають можливість покращення взаємодії користувачів з комп'ютерними системами та дозволяють реалізувати автоматизовані методи класифікації та обробки запитів [1].

У цьому контексті розробка моделі категоризації запитів користувачів за допомогою технологій NLP стає ключовим інструментом для покращення якості обробки інформації та швидкості взаємодії. Відтак, розробка ефективної моделі категоризації є актуальною задачею в умовах сучасного інформаційного суспільства.

Запити користувачів, відповідно до системи чи сервісу куди вони надходять, можуть містити від одного-двох слів (наприклад, для інформаційного пошуку у мережі Інтернет), до декількох речень (наприклад, електронний лист до установи з переліком питань). У останньому випадку класична мультикласова класифікація тексту, що з множини класів надає результатом лише один певний клас до якого текст віднесено, не є достатньою, оскільки запит одночасно може бути віднесено до кількох класів, хоч і з різною вірогідністю відповідності до кожного з класів. Більш відповідною до задачі, що необхідно вирішити, є багатокласова класифікація, що надає ймовірнісний розподіл за класами.

Для багато класової класифікації запиту природньою мовою доречним є використання методик машинного навчання, мовних моделей та інших технологій NLP. Формування певної кількості ключових фраз, розподіл їх по категоріям з певним ваговим коефіцієнтом як мірою важливості для даної категорії, обрахування відповідності запиту до ключових фраз заздалегідь

навченими моделями та підсилення ступені відповідності мірою важливості, надасть можливості більш якісної категоризації запитів користувачів, та, як слідство, формування більш релевантних відповідей на запит.

Пропонується використання наступної моделі категоризації запиту користувача за допомогою технологій NLP.

Множина категорій, де N – кількість категорій:

$$C = \{c_1, c_2 \dots c_N\}. \quad (1)$$

Множина ключових фраз для пошуку схожості, M – кількість фраз:

$$L = \{l_1, l_2 \dots l_M\}. \quad (2)$$

Вагові коефіцієнти фраз у категоріях, де w_{ij} – ваговий коефіцієнт та $w_{ij} \in [0;1]$:

$$W = \{(c_i, l_j, w_{ij}) | c_i \in C, l_j \in L\}. \quad (3)$$

Коефіцієнти відповідності запиту до фраз у категоріях, де sim_{ij} – коефіцієнт відповідності, що обраховується шляхом розрахунку метрики відповідності двох векторів: векторне уявлення ключової фрази та векторне уявлення запиту:

$$Sim = \{sim_{ij} | \forall (c_i, l_j) c_i \in C, l_j \in L\}. \quad (5)$$

Ступень належності запиту до фраз у категоріях:

$$R = \{w_{ij} \cdot sim_{ij} | w_{ij} \in W, sim_{ij} \in Sim\} \quad (6)$$

Вектор максимальних значень ступені належності запиту до i -категорії:

$$M = \{\max_i (r_{ij}) | i \in R\} \quad (7)$$

Вектор номерів категорій, до яких належить запит, де $limitValue$ – порогове значення:

$$SC = \{i | m_i > limitValue\} \quad (8)$$

Таким чином, запиту буде поставлено у відповідність множену номерів категорій, яким він відповідає.

Було проведено експеримент з використанням моделі на спрощених похідних даних: наборах запитів українською мовою та ключових фразах з апіорі відомою відповідністю до запитів.

Векторні уявлення для похідних наборів обраховано за попередньою токенизацією з використанням моделей BertModel, XLNetModel, XLMRobertaModel, які побудовані на архітектурі трансформер, призначені для попередньо навчених мовних уявлень та є мультимовними [2], коефіцієнти відповідності обраховано за метрикою косинусної подоби векторів, що надає значення відповідності у діапазоні $[0;1]$. Відповідно до апіорі відомої подоби ключових фраз до запитів для них було встановлено вагові коефіцієнти: $w_{11}C_1L_1 = 0.95$, $w_{22}C_2L_2 = 0.5$, $w_{33}C_3L_3 = 0.1$ та виконано категоризацію для порогових значень $limitValue = 0.45$ та $limitValue = 0.50$. Результати моделювання наведено у таблиці 1, де сірим кольором вказано значення, що перевищують порогове, а відповідна категорія відноситься до запиту.

Таблиця 1

Кількість запитів	Мовні моделі	Косинусна подоба			Ступінь належності					
					limitValue = 0,45			limitValue = 0,5		
		C_1L_1	C_2L_2	C_3L_3	C_1L_1	C_2L_2	C_3L_3	C_1L_1	C_2L_2	C_3L_3
N = 10	BertModel	0.900	0.710	0.450	0,86	0,36	0,05	0,86	0,36	0,05
	XLNetModel	0.990	0.980	0.870	0,94	0,49	0,09	0,94	0,49	0,09
	XLNetModel	0.997	0.990	0.952	0,95	0,50	0,10	0,95	0,50	0,10
N = 30	BertModel	0.908	0.690	0.447	0,86	0,35	0,04	0,86	0,35	0,04
	XLNetModel	0.980	0.977	0.865	0,93	0,49	0,09	0,93	0,49	0,09
	XLNetModel	0.995	0.989	0.950	0,95	0,49	0,10	0,95	0,49	0,10
N = 100	BertModel	0.920	0.675	0.437	0,87	0,34	0,04	0,87	0,34	0,04
	XLNetModel	0.979	0.976	0.860	0,93	0,49	0,09	0,93	0,49	0,09
	XLNetModel	0.995	0.985	0.947	0,95	0,49	0,09	0,95	0,49	0,09

За результатами моделювання можна зробити висновки, що модель BertModel найбільш адекватно відображає результат категоризації, а значення косинус них подоб може бути скореговано розробленою моделлю за рахунок налаштування ваги ключових фраз та пороговим значенням limitValue.

Література

1. Manning Christopher D Foundations of statistical natural language processing / D. Manning Christopher, Schutze Hinrich. [Електронний ресурс] – Режим доступу: https://doc.lagout.org/science/0_Computer%20Science/2_Algorithms/Statistical%20Natural%20Language%20Processing.pdf (дата звернення 20.11.2023). – Назва с екрану.
2. Офіційний сайт платформи Hugging Face [Електронний ресурс] – Режим доступу: <https://huggingface.co/> (дата звернення 20.11.2023). – Назва с екрану.

Анотація

Точна багатокласова класифікація запитів користувачів природньою мовою забезпечує надання швидкої та релевантної відповіді. Пропонується модель категоризації запитів, що використовує сучасні моделі NLP та дозволяє використовувати параметри для більш якісної категоризації.

Ключові слова: модель, запит, категорія, NLP.

Abstract

Accurate multi-class natural language classification of user queries ensures fast and relevant response. A query categorization model is proposed that uses modern NLP models and allows using parameters for better categorization.

Keywords: model, query, category, NLP.