УДК 004.89:004.912

ПОРІВНЯННЯ ФРАЗОВОГО ТА ПОСЛОВНОГО ЗАСТОСУВАННЯ ВІДСТАНІ ЛЕВЕНШТЕЙНА ДЛЯ РОЗПІЗНАВАННЯ НАМІРІВ В УКРАЇНОМОВНОМУ ЧАТ-БОТІ

Гунченко А. Ю.[1]; Мартинович Л. Я.[2]; Фуштор А. В.[3]
[1,2,3] Одеський національний університет імені І. І. Мечникова
[1,2,3] Україна, Одеса
[1] gunchenko.andrii@stud.onu.edu.ua; [2] larysa.yaroslavna@onu.edu.ua; [3] fushtor.anastasiia@stud.onu.edu.ua

Анотація. *Порівняно фразовий і пословний способи застосування відстані Левенштейна для визначення наміру в україномовному чат-боті. На 30 тематичних і 20 позатематичних запитах пословний спосіб правильно визначив намір у 93 % випадків проти 30 % для фразового і надійніше відсікав сторонні запити. Результати придатні для вузькотематичних чат-ботів з невеликою базою знань.*
Ключові слова: *чат-бот; відстань Левенштейна; розпізнавання намірів; обробка природної мови; українська мова.*

Вступ. Вузькотематичні чат-боти, що відповідають на запитання з невеликої бази знань, широко застосовуються в довідкових і навчальних системах [1]. Такі боти часто визначають намір користувача без навчання моделі: запит порівнюється з еталонними запитаннями, і обирається найближче за мірою схожості рядків. Найпоширенішою такою мірою є відстань Левенштейна [2], тобто мінімальна кількість вставок, видалень і замін символів, потрібних для перетворення одного рядка на інший. Її обчислюють методом динамічного програмування за алгоритмом Вагнера - Фішера [3], а огляд споріднених методів наближеного пошуку рядків наведено в роботі [4]. У невеликих проєктах відстань зазвичай обчислюють між цілими фразами. Проте для української мови з розвиненою системою закінчень і для коротких розмовних запитів ефективність такого підходу не очевидна.

Мета роботи - порівняти фразовий і пословний способи застосування відстані Левенштейна для визначення наміру в україномовному чат-боті та оцінити, як кожен із них розпізнає перефразовані, короткі та позатематичні запити.

Основна частина. Дослідження виконано на базі знань Telegram-бота про ушу з курсового проєкту. Вона містить 50 еталонних запитань, рівномірно розподілених за 5 напрямами: історія, види, відмінності від інших видів спорту, Таолу та Саньда. Перед порівнянням текст зводиться до нижнього регістру й очищається від розділових знаків.

Фразовий спосіб порівнює запит a з еталоном b як два суцільні рядки:

$$r(a, b) = d(a, b) / ((|a| + |b|) / 2) * 100\ \%, \quad (1)$$

де d(a, b) - відстань Левенштейна, |a| і |b| - довжини рядків. Намір визначається за еталоном із найменшим значенням r.

Пословний спосіб працює з окремими словами. Для кожного слова w запиту довжиною не менше чотирьох символів (так відкидаються короткі службові слова) шукається найближче слово x еталона, а підсумкова оцінка є середнім значенням таких відстаней:

$$s(a, b) = (1 / |W|) * \Sigma \min d(w, x) / \max(|w|, |x|) * 100\ \%, \quad (2)$$

де W - множина значущих слів запиту, сума береться по всіх словах $w \in W$, а мінімум - по всіх словах x еталона b. Завдяки цьому різні форми одного слова («зброя» - «зброї») дають невелику відстань.

Для перевірки сформовано тестову вибірку з 30 запитів, які не збігаються з еталонними (по шість на кожен намір): короткі фрази, перефразування і розмовні звороти. Окремо взято 20 поза тематичних запитів, на які бот повинен відповідати відмовою. Результати порівняння наведено в таблиці 1.

Таблиця 1 - Порівняння способів визначення наміру

Показник	Фразовий спосіб	Пословний спосіб
Правильний намір без порогу (30 запитів)	9 (30 %)	28 (93 %)
Середня відстань до найближчого еталона: тематичні / позатематичні запити	77,3 % / 86,1 %	22,9 % / 56,4 %
Поріг відстані	50 %	30 %
– правильно прийнято тематичних запитів	1 з 30	16 з 30
– хибно прийнято тематичних запитів	0	0
– прийнято позатематичних запитів	0 з 20	0 з 20
Поріг відстані	70 %	40 %
– правильно прийнято тематичних запитів	5 з 30	25 з 30
– хибно прийнято тематичних запитів	4	0
– прийнято позатематичних запитів	0 з 20	2 з 20

Як видно з таблиці 1, фразовий спосіб майже не придатний для запитів, які відрізняються від еталонів. Середня відстань для тематичних запитів (77,3 %) лише трохи менша, ніж для позатематичних (86,1 %), тому розділити їх порогом неможливо. Головна причина полягає в різниці довжин: короткий запит «що таке таолу» (13 символів) порівнюється з еталонами довжиною 30-60 символів, і більшу частину відстані становлять «зайві» символи довшого рядка. Через це відстань до правильного еталона «Що являє собою розділ таолу?» виявилася більшою, ніж до еталона «Як і де виникло ушу?» з іншого розділу .

Різницю між способами добре видно на конкретних запитах (таблиця 2). Для коротких запитів пословний спосіб знаходить у еталоні всі значущі слова і дає нульову або близьку до нуля відстань, тоді як фразовий спосіб відносить їх до хибного розділу з відстанню понад 85%. Для позатематичних запитів пословна відстань залишається високою (понад 60 %), що дозволяє відсікти їх порогом. Прочерк у таблиці 2 означає, що запит відхилено порогом (50 % для фразового і 30 % для пословного способу).

Таблиця 2 - Приклади визначення наміру двома способами

Запит	Фразовий: r, %	Фразовий: намір	Пословний: s, %	Пословний: намір
що таке таолу	87,5	Історія (хибно)	0,0	Таолу
стилі ушу	92,9	Історія (хибно)	0,0	Види та напрями
правила саньда	103,0	Історія (хибно)	7,1	Саньда
чим ушу краще за карате	77,6	Відмінності	16,7	Відмінності
Яка погода в Одесі?	79,1	–	62,5	–
Як приготувати борщ?	78,9	–	61,4	–

Водночас фразовий спосіб добре виправляє описки в запитах, близьких до еталонних. У експерименті до кожного з 50 еталонів внесено випадкові спотворення (вставку, видалення чи заміну) у 10, 20 і 30 % літер, по чотири варіанти на еталон. В усіх 600 випадках фразовий

спосіб правильно знайшов вихідне запитання, тоді як відстань Левенштейна, обчислена між послідовностями цілих слів, при 30 % спотворень дала лише 88,5 % правильних відповідей.

Пословний спосіб розділяє тематичні та позатематичні запити краще: середні відстані відрізняються майже в 2,5 раза. При порозі 30 % бот розпізнав 16 із 30 запитів, не допустивши жодної хибної відповіді та не прийнявши стороннього запиту. Підвищення порогу до 40 % збільшує кількість розпізнаних запитів до 25, але два позатематичні запити проходять перевірку. Отже, поріг 30-35 % підходить для «обережного» бота, а 40 % - для бота, якому важливіше рідше відмовляти.

Пословний спосіб має і обмеження: він не враховує порядок слів і не розпізнає синонімів. Так, запит «хто створив ушу» не має спільних значущих слів з еталонами розділу «Історія та походження» і залишився нерозпізнаним. Такі випадки доцільно передавати класифікатору машинного навчання, наприклад логістичній регресії на символьних n-грамах, стійких до змін закінчень [5], як це зроблено в гібридній схемі розробленого бота.

Висновки. Для україномовного чат-бота з невеликою базою знань фразове застосування відстані Левенштейна доцільно використовувати лише як перший рівень перевірки, що виправляє описки, але не розпізнає перефразувань. Пословний спосіб без навчання правильно визначив намір у 93 % коротких і перефразованих запитів, а при порозі 30 % не дав жодної хибної відповіді. Поєднання пословної відстані Левенштейна з класифікатором машинного навчання та порогом упевненості є простим і ефективним рішенням для вузькотематичних ботів. Подальші дослідження доцільно спрямувати на врахування синонімів і зважування слів за їхньою інформативністю.

Використання інструментів ШІ. Під час підготовки роботи генеративні інструменти штучного інтелекту (Claude, Anthropic) використовувалися для редагування тексту, перекладу анотації англійською мовою та перевірки оформлення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial intelligence applications and innovations* (pp. 373–383). Cham: Springer. https://doi.org/10.1007/978-3-030-49186-4_31

[2] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

[3] Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM, 21*(1), 168–173.

[4] Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys, 33*(1), 31–88.

[5] Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval* (pp. 161–175). Las Vegas, NV.

Hunchenko A. Yu., Martynovych L. Ya., Fushtor A. V.

Comparison of phrase-level and word-level application of the Levenshtein distance for intent recognition in a Ukrainian-language chatbot

Abstract. *Phrase-level and word-level methods of applying the Levenshtein distance to intent recognition in a Ukrainian-language chatbot are compared. On 30 in-domain and 20 out-of-domain queries, the word-level method identified the correct intent in 93 % of cases versus 30 % for the phrase-level method and rejected out-of-domain queries more reliably. The results are applicable to narrow-domain chatbots with a small knowledge base.*

Keywords: *chatbot; Levenshtein distance; intent recognition; natural language processing; Ukrainian language.*