УДК 004.891

АЛГОРИТМ ПОЛІЛІНГВАЛЬНОГО ПОШУКУ ВІДПОВІДЕЙ В ІНФОРМАЦІЙНІЙ СИСТЕМІ ПІДТРИМКИ ІНОЗЕМНИХ СТУДЕНТІВ

Заїка А.Р.[1]; Булгакова О.С., к.т.н., доц.[2]
[1,2] Національний університет кораблебудування імені адмірала Макарова
[1,2] Україна, Миколаїв
[1]antonina.zaika.04@gmail.com; [2]oleksandra.bulhakova@nuos.edu.ua

Анотація. *Розглянуто задачу полілінгвального інформаційного пошуку в системі підтримки іноземних студентів. Запропоновано алгоритм зіставлення природномовних запитів із записами багатомовної бази знань, який враховує особливості української, англійської та китайської мов.*
Ключові слова: *полілінгвальний інформаційний пошук; алгоритм пошуку; багатомовна база знань; оброблення природної мови; оцінювання відповідності; віртуальний асистент.*

Вступна частина. Іноземні студенти звертаються до університетських вебресурсів із повторюваними запитаннями щодо вступу, документів, розкладу, проживання та навчання. Доступ до відповіді ускладнюється, коли користувач не володіє мовою ресурсу або формулює запит інакше, ніж названо відповідний розділ. Для багатомовних систем запитань і відповідей дослідники розглядають нерівномірне представлення інформації різними мовами та способи доступу до неї [1]. Однією з основних проблем інформаційного пошуку є встановлення відповідності між запитом, сформульованим природною мовою, та інформаційними одиницями, представленими в базі знань. У традиційних системах пошуку за ключовими словами результат значною мірою залежить від буквального збігу текстових фрагментів, але однакові за змістом запитання можуть суттєво відрізнятися за структурою, порядком слів та використаними мовними конструкціями [2]. Буквальне входження ключової фрази не охоплює всіх формулювань: англомовний запит «How much does study cost?» відрізняється за порядком слів і складом службових одиниць від словосполучення «cost of study». Для китайських запитів додатковою умовою є відсутність пробілів між словами, що обмежує можливість застосування єдиного механізму токенізації та зіставлення ключових фраз для всіх мов [3].

Метою дослідження є розроблення алгоритму полілінгвального інформаційного пошуку для визначення найбільш релевантної відповіді на запит користувача шляхом формалізації процедур нормалізації тексту, зіставлення ключових фраз та обчислення інтегральної оцінки відповідності з урахуванням особливостей української, англійської та китайської мов.

Основна частина. Алгоритм призначений для пошуку відповідей у структурованій багатомовній базі знань, функціонування якого ґрунтується на зіставленні тексту запиту з множинами ключових фраз, пов’язаних із відповідними інформаційними записами. База знань містить 35 записів, розподілених між вісьмома тематичними категоріями: вступ, документи, навчальний процес, розклад, гуртожиток, бібліотека, академічна мобільність і контакти. Кожний запис бази знань представлено структурою: $b_i = (id_i, c_i, \{K_i^l, A_i^l\})$,

де id_i – унікальний ідентифікатор запису, c_i – тематична категорія, K_i^l – множина ключових фраз мовою l, A_i^l – відповідь мовою l. $L = \{uk, en, zh\}$ – множина підтримуваних мов. Під час

формування багатомовної бази знань використано принцип функціональної еквівалентності, відповідно до якого мовні версії одного запису передають однакову довідкову інформацію, але можуть містити різні набори ключових фраз.

Послідовність основних етапів алгоритму полілінгвального пошуку відповідей представлено на рис. 1. Для української та англійської мов використовується порівняння унікальних змістових слів $T_l(k)$ після вилучення службових одиниць. Локальна оцінка відповідності визначається функцією:

$$s_l(q,k) = \begin{cases} 10 + 0.1\min(n,10), C_1, \\ 8 + 0.1\min(m,10), C_2 \\ 4r, C_3 \\ 0, \textit{інакше} \end{cases} \quad (1)$$

де $m = |T_l(q) \cap T_l(k)|, n = |T_l(k)|, r = m/n$ C_1 – точне входження нормалізованої ключової фрази до запиту з урахуванням меж слів, C_2 – повний збіг змістових слів ($m = n > 0$), C_3 – частковий збіг ($m \geq 2, r \geq 0.67$).

Для китайської мови застосовується аналогічний принцип пріоритетності збігів, замість слів використовуються символьні послідовності. Точне входження ключової фрази оцінюється з урахуванням її довжини, а часткова відповідність визначається за часткою збіжних двосимвольних фрагментів за умови наявності щонайменше двох збігів та перекриття не менше 0,67. Інтегральна оцінка відповідності запису враховує найкращий збіг ключової фрази та додаткові збіги:

$$S_i(q,l) = \max s_l((q,k) + 0.01\min\left(\sum s_l(q,k),10\right) \quad (2)$$

Додаткова складова обмежена величиною 0.1, що забезпечує перевагу найкращого збігу над накопиченням часткових відповідностей. Вибір найбільш релевантного запису здійснюється за правилом:

$$i^* = \arg\max_{1 \leq i \leq N} S_i(q,l). \quad (3)$$

За однакових максимальних оцінок обирається перший запис у базі знань. Якщо всі оцінки дорівнюють нулю, формується повідомлення про відсутність відповіді.

Розглянемо приклад: для англомовного запиту «How much does study cost?» та ключової фрази «cost of study» після вилучення службових слів отримуємо два спільні змістові слова: study і cost, m=n=2, відповідно до (1) локальна оцінка становить $s_{en}(q,k)$=8.2. Таким чином, алгоритм встановлює відповідність незалежно від порядку змістових слів, на відміну від пошуку за буквальним входженням ключової фрази.

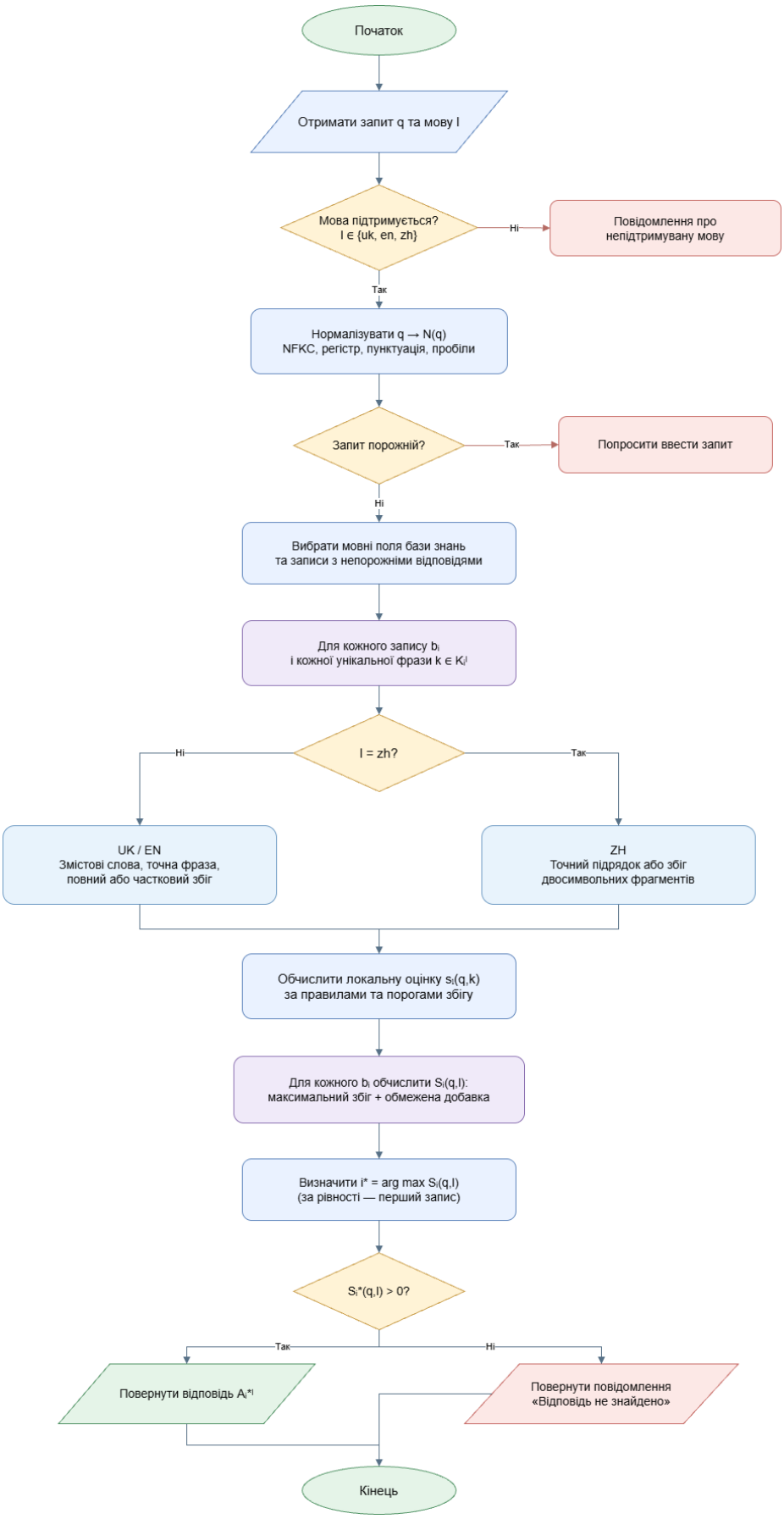


Рис.1. Блок-схема алгоритму полілінгвального пошуку відповідей

Висновки. Розроблено алгоритм полілінгвального інформаційного пошуку для визначення релевантних відповідей у багатомовній базі знань, орієнтованій на інформаційну підтримку іноземних студентів. Запропонований алгоритм належить до правилових методів інформаційного пошуку та не потребує навчання на попередньо розмічених наборах запитів. Його перевагою є можливість кількісного оцінювання відповідності за визначеними правилами, а також адаптація процедури зіставлення до особливостей різних мов. Результат пошуку залежить від повноти множин ключових фраз: невідомі синоніми, орфографічні помилки, словозміна та неоднозначні формулювання можуть ускладнювати вибір релевантного запису. Дана процедура також не забезпечує повноцінного семантичного аналізу запитів.

Подальше дослідження передбачає формування незалежного тестового набору запитів, що міститиме точні формулювання, перефразування, часткові збіги та запитання, для яких у базі знань відсутні відповіді.

Використання інструментів ШІ. Ghat GPT використано для перевірки орфографії та пунктуації, усунення мовних неточностей і перевірки термінологічної узгодженості українськомовного тексту та англомовної анотації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., & Lin, J. (2023). MIRACL: A multilingual retrieval dataset covering 18 diverse languages, *Transactions of the Association for Computational Linguistics*, 11, 1114-1131. https://doi.org/10.1162/tacl_a_00595.

[2] ang, J., Jiang, F., & Baldwin, T. (2024). *Language bias in multilingual information retrieval: The nature of the beast and mitigation methods. IV Workshop on Multilingual Representation Learning*, 280-292. https://doi.org/10.18653/v1/2024.mrl-1.23.

[3] Kamalloo, E., Zhang, X., Ogundepo, O., Thakur, N., Alfonso-Hermelo, D., Rezagholizadeh, M., & Lin, J. (2023). *Evaluating embedding APIs for information retrieval, 61st Annual Meeting of the Association for Computational Linguistics*, 5: Industry Track, 518-526. https://doi.org/10.18653/v1/2023.acl-industry.50.

Zaika A.R., Bulgakova O.S.

Multilingual answer retrieval algorithm for information support of international students

Abstract. *This paper addresses the problem of multilingual information retrieval in an information support system for international students. An algorithm is proposed for matching natural language queries with records in a multilingual knowledge base, taking into account the linguistic characteristics of Ukrainian, English, and Chinese.*

Keywords: *multilingual information retrieval; search algorithm; multilingual knowledge base; natural language processing; relevance assessment; virtual assistant.*