УДК 004.58+004.94

О.О. Кряжич, О.В. Васенко, Л.М. Ісак, О.А. Бабак, В.О. Грицишин

МЕТОД ПОБУДОВИ ЗАПИТІВ ДО ЧАТ-БОТІВ НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ

Кряжич Ольга Олександрівна

Інститут телекомунікацій і глобального інформаційного простору НАН України, м. Київ,
orcid: 0000-0003-1845-5014

economconsult@gmail.com

Васенко Олександр Васильович

Університет Григорія Сковороди в Переяславі,
orcid: 0000-0002-8895-4284

vasenko.olexandr@gmail.com

Ісак Людмила Марківна

Університет Григорія Сковороди в Переяславі,
orcid: 0000-0001-7466-8757

isakluda@ukr.net

Бабак Олександр Анатолійович

Університет Григорія Сковороди в Переяславі,
orcid: 0000-0003-1657-4132

babak1109@gmail.com

Грицишин Володимир Омелянович

Східноукраїнський національний університет імені Володимира Даля, м. Київ,
orcid: 0000-0003-3674-4848

hrytsyshynvo@gmail.com

В Інтернет все більшої популярності набувають чат-боти з породжувальним штучним інтелектом (ШІ) — ChatGPT, Gemini, GitHub Copilot та ін. Вони використовуються для навчання та інтелектуального пошуку інформації. Правильно побудований запит дає чіткий результат-відповідь від чат-бота і полегшує процес взаємодії з ним. У роботі математично обґрунтовано метод побудови запитів до чат-ботів з урахуванням особливостей, які дозволяють формувати запит таким чином, щоб у наданій відповіді задовольнялися вимоги до основних властивостей інформації. Крім того, в роботі проаналізовано можливість формування запитів до чат-бота щодо перетворення з використанням нечіткої логіки і побудовою рішень у вигляді лінгвістичних правил-продукцій та подальшої формалізації із

 ISSN 2786-6491

застосуванням одного з базових методів наближення для отримання результату за максимумом відповідності. Зокрема, запропоновано алгоритм з десяти базових кроків, за яким реалізується можливість вибору термів лінгвістичних змінних, що з різним ступенем відповідають одній з двох сформованих множин. Терм — це деякий логічний вираз, що стосується об'єкта дослідження. Попередні результати тестування запропонованого методу виявили можливість на 14 % надавати більш точну, достовірну, повну, корисну та зрозумілу відповідь. Витрати часу на підготовку запитів можна мінімізувати шляхом розробки вебсервісів за заданим алгоритмом, які дозволятимуть користувачам отримувати інформаційні масиви для запиту та будувати з них запити за продукційними правилами, що є перспективним напрямом подальших досліджень. Зазначене може бути використане при розробці вебсервісів обробки інформації.

Ключові слова: породжувальний алгоритм, мовна модель, продукційні правила, масив, формування підказок, показник ефективності відповіді, онтологія.

Вступ

Чат-боти з породжувальним штучним інтелектом (ШІ), такі як ChatGPT, Gemini, GitHub Copilot та інші, можна розглядати як інструменти ШІ [1], що за допомогою вдосконаленої мовної моделі здатні генерувати творчі продукти за технологіями, подібними до людських [2]. Робота з ШІ передбачає формування запитів для отримання бажаних відповідей, що дозволяє користувачам взаємодіяти з чат-ботом для отримання пояснень за запитом, фрагментів програмного коду, пропозицій щодо дизайну та розробки зображень і допомоги в оформленні документації.

Правильно побудований запит дозволяє отримати від ChatGPT чітку відповідь. Для складання якісного тексту запиту необхідно враховувати діалогічні, структурні та лінгвістичні аспекти аргументації. Як зазначається в роботі [3], для цього може бути недостатньо як базових знань, так і знань предмета дослідження та навичок попередньої роботи з довідковими та пошуковими системами в мережі Інтернет.

Цифрова трансформація суспільства призвела до активізації запитів з використанням сервісів ШІ, а також до потреб щодо визначення оптимальних комбінацій людського та генеративного ШІ для виконання різних завдань і визначення способів оцінки точності тексту, створеного генеративним ШІ [4]. З досліджень цих потреб випливає однозначна відповідь щодо аргументованого запиту до системи [3]. Варто зауважити, що у зазначеному випадку предмет пізнання, який у контексті пошуку виступає об'єктом дослідження, для надання повної відповіді на запит повинен бути представлений через причинно-наслідкові зв'язки, які характеризують предмет дослідження як складну систему або компонент складної системи [5] на деякому часовому проміжку функціонування. З цього випливає, що запити користувача до інструментів ШІ, таких як ChatGPT, Gemini, GitHub Copilot та інші, потребують певного методу структурування тексту запиту та використання певної моделі тексту, яка дозволить інструментам ШІ надати точну, достовірну, повну, корисну та зрозумілу відповідь з мінімальними витратами часу на обробку запиту.

Основна проблема дослідження полягає у тому, що серед наукових праць публікації з промпт-інжинірингу та наукового обґрунтування методів та підходів до створення запитів при роботі з чат-ботами становлять лише незначну частку. Здебільшого розглядаються лінгвістичні особливості складання запитів природною мовою при навчанні школярів [3] та студентів [6] для перевірки своїх знань та самопідготовки з вдосконалення власних навичок, включаючи творчі [7], а також

мовні моделі при формуванні підказок [8]. У публікаціях, розглянутих у даній роботі, математичне обґрунтування методів формування підказок наведене здебільшого фрагментарно, що викликає певні складності розуміння у користувачів чат-ботів, які застосовують дані технології у професійній діяльності або для вирішення завдань прикладного характеру з використанням системного підходу.

Постановка задачі

У статті представлено математично обґрунтований метод побудови запитів до чат-ботів з урахуванням особливостей, які дозволяють формувати запит таким чином, щоб у наданій відповіді дотримувалися вимоги до основних властивостей інформації.

Варто зазначити, що навіть англомовна версія статті Вікіпедії «Prompt engineering» (конструювання підказок) надає загальні користувальницькі уявлення про створення запитів, базуючись на абстрактних керівництвах для користувачів чат-ботів. Відсутність фундаментальних досліджень, які мають доказову базу, вказує на новизну даної теми, а зростання запитів користувачів щодо формування запитів до ШІ на основі різноманітних мовних моделей при роботі з чат-ботами [9] свідчить про зростання її актуальності.

Формування запиту на пошук інформації та відповідь на запит можна розглянути через представлення опису та актуальної інформації про стан деякої складної системи [10]. Наприклад, створюючи запит про отримання довідки щодо стану довкілля, варто згадати, що довкілля — це складна система з живих і неживих компонентів. Об’єкт запиту, «довкілля», не існує сам по собі. Деяке підприємство в межах довкілля виступає як самостійна складна система. Людину або людський організм теж можна представити як складну систему. Перетин таких систем збільшує перелік компонентів, процеси взаємодії та зворотні зв’язки, тобто ускладнює представлення терміну «довкілля» у межах простого запиту без відповідної мовної конструкції, вимагаючи одночасно лаконічного та змістовного вираження [3].

Якщо чат-бот розглядати як інструмент, що дозволяє реалізувати низку алгоритмів переробки інформації в потрібні формати даних за заданими параметрами та з урахуванням існуючих обмежень, то, за визначеннями академіка В.М. Глушкова [5], подібна система може виступати:

— інформаційно-пошуковою системою (ІПС), тобто сукупністю мовно-алгоритмічних і технічних засобів, призначених для зберігання, пошуку та видачі необхідної інформації; ІПС можуть бути документальними (для видачі запитів з технічних документів — статей, патентів, звітів та ін.) та фактографічними (призначеними для видачі відповідей на інформаційні запити щодо якихось фактів);

— інформаційно-довідковою системою (ІДС), яка є системою реєстрації, переробки і зберігання інформації, призначеної для забезпечення користувачів інформацією довідкового характеру.

Аналізуючи наведені визначення, а також роботи [11–13] щодо конструювання підказок при формуванні запитів до моделей породжувального ШІ, можна зазначити, що, знов-таки, за визначенням В.М. Глушкова [5], для створення повноцінного запиту необхідно мати уявлення та повний набір знань про об’єкт запиту користувача. У користувача такий набір знань відсутній, а безпосередньо чат-бот має доступ до досліджуваного об’єкта запиту та відповідні алгоритми. За умов невизначеності з позиції користувача для формування такого набору можна (і слід) використовувати будь-яку наявну формалізовану і неформалізовану інформацію, яка після обробки за допомогою алгоритмів чат-бота дасть мінімально повні знання на запит користувача. Тому запит до чат-бота можна описати так:

$$Q = F + N + \Delta\varphi, \tag{1}$$

де Q — знання про об'єкт дослідження у відповідь на запит користувача; F — формалізовані знання про складну систему, яка є об'єктом дослідження користувача; N — неформалізовані знання про складну систему (наявність передумов до визначення відповідності цільової функції, структури, складу ресурсів і критеріїв функціонування вимогам навколишнього середовища; нагальні потреби; можливість виникнення кризи; та ін.); $\Delta\varphi$ — визначник рівня погрішності інформації при її обробці за умов невизначеності, який висвітлює ті дані, яких не вистачає для опису складної системи і серед яких, можливо, є відомості про вирішення системних проблем.

Для того щоб описати систему відповідно до стану в умовах реального часу, необхідно прагнути до

$$F + N > \Delta\varphi. \tag{2}$$

Іншими словами, знання, що відсутні серед формалізованих та неформалізованих знань, не повинні перевищувати межу, яку визначає відповідність складної системи її цільовій функції. Саме $\Delta\varphi$ корелює до міри невизначеності інформації в системі управління E [14], що дозволяє при наявності на вході менш імовірних значень (поєднання окремих вибраних слів у запиті користувача) отримати більш імовірні значення на виході (точна відповідь системи).

Логіка вирішення поставленої задачі

Інформація в загальному розумінні [5] є науковою категорією, що визначає деякі відомості, сукупність деяких знань або даних за певними властивостями. З усіх властивостей інформації з погляду математичної логіки можна представити повноту та несуперечність, а своєчасність — як математичну модель події у якійсь точці часового відрізку, що важливо у порівнянні алгоритмів роботи чат-ботів та ІПС/ІДС. Поняття переробки великих обсягів інформації та формування чітко структурованого набору розподілених даних можна пояснити через переробку інформації в масив даних за адаптивною схемою за двома паралельними напрямами. Перший — кількість інформації розглядається саме як число, що відображає важливість одержуваних відомостей з їх семантичними та прагматичними аспектами. Другий — обробка інформації виконується з використанням визначення загальної міри кількості інформації К. Шеннона та ентропії [14], що дозволяє зрозуміти процеси кодування, передачі та зберігання. Завдяки цьому аспекту можна зрозуміти обсяг одержуваної інформації та пов'язати його з поведінкою отримувача, який вирішує якісь завдання, що приводить до розуміння кількості семантичної інформації для оцінки прагматичної цінності для переробки в масив даних [5].

Перетворювачами інформації можуть виступати різні сервіси, реалізовані на вебплатформах, які отримують інформацію від інших користувачів і переробляють її для подальшого використання в процесах перетворення інформації. Набір записів даних щодо якогось об'єкта запиту в чат-ботах являє собою масив даних (інформаційний масив) [15] великого обсягу [16, 17]. Ці тези можна вважати основоположними в інформаційному підході В.М. Глушкова, оскільки вони розкривають суть трансформації інформації в дані і побічно вказують на той факт, що дані можуть бути не просто великими, а безмежними.

Користувач чат-бота, стикаючись з невизначеністю при створенні запиту до чат-бота, може реалізувати дві основні можливості. По-перше, він може спробувати отримати додаткову інформацію і ще раз проаналізувати проблему [18]. Таким чином часто вдається зменшити новизну і складність питання, що виникло. Ко-

ристувач поєднує цю додаткову інформацію та аналіз з накопиченим досвідом, здатністю до судження або інтуїцією, щоб надати відповіді від чат-бота суб'єктивну або передбачувану ймовірність. По-друге, користувач чат-бота може діяти у точній відповідності до минулого досвіду, судження або інтуїції і зробити припущення про ймовірність подій з використанням обмеженої інформації [19] для запиту до чат-бота. Це необхідно, коли не вистачає часу на збір додаткової інформації або створити точний запит занадто складно. Такі часові та інформаційні обмеження можуть призвести до великого відсотка нерелевантної інформації.

Перебіг часу обумовлює зміни ситуації в динамічних системах. Якщо вони значні, ситуація може змінитися настільки, що наявна інформація не відповідатиме критеріям. Тому слід знову досягти якоїсь межі у відборі, обробці та видачі користувачу тієї інформації, яка є релевантною і точною. Часто це є важким завданням, оскільки час між отриманням інформації з якогось її джерела і початком дії, коли є запит на цю інформацію, занадто великий. Крім того, зазвичай інформація потрібна користувачу для прийняття якогось рішення, і це рішення слід прийняти швидко, щоб бажана дія не втратила актуальності. Тому іноді для врахування фактору часу користувачі, формуючи запит до чат-бота, змушені спиратися на судження або створювати запит інтуїтивно. Подібним чином слід враховувати ймовірність того, що інформація може випереджати свій час. Це стосується здебільшого прогнозних даних, які є передбаченням майбутнього стану предмета чи події на основі минулих або теперішніх даних. Запит користувача може бути створений таким чином, що виникатимуть інформаційні обмеження і відповіддю від чат-бота будуть згенеровані дані на основі різних джерел, включаючи ті, що несуть у собі неперевірену, неточну інформацію або вигадані дані.

Аналітико-синтетична обробка інформації для створення запиту користувача до чат-бота

Виконання алгоритмів чат-бота після отримання запиту від користувача можна описати за В.М. Глушковим: «виконання комплексу операцій над даними… з метою перетворення різноманітних відомостей і фактів у відомості, що мають цінність з певного погляду» [5]. Алгоритми чат-бота працюють з формалізованими даними. Формалізація ігнорує деяку частину доступної інформації [3], перетворюючи її на систему формул як інформацію про об'єкт дослідження. А потім ця система формул знову трансформується для надання відповіді природною мовою, зрозумілою користувачу. І в цьому процесі дуже важливо не втратити основні властивості інформації, надаючи перевагу аксіомам і правилам без урахування семантики.

При перетворенні інформації будь-яку подію в процесі функціонування складної системи на певному проміжку часу $[t; t']$ можна розглядати як деяку точку виникнення ситуації в точці M з умовними координатами $[t_n; t_{n+1}]$. У цій точці M виникають певні вимоги до властивостей інформації, яка необхідна для виконання системою своїх функцій, а саме:

— актуальність: забезпечення істотних даних для події у точці M $[t_n; t_{n+1}]$;

— вірогідність: наявні дані про стан системи відповідають саме події у вказаний момент часу;

— об'єктивність: на отриману інформацію не вплинули жодні сторонні чинники;

— повнота: отриманої інформації достатньо для того, щоб оцінити подію в точці M $[t_n; t_{n+1}]$ та прийняти відповідне рішення;

— адекватність: прийняте за наведеними ознаками рішення відповідає реальним потребам складної системи в точці $M\,[t_n;\, t_{n+1}]$.

Зазвичай при розробці інформаційних технологій управління складними системами [16] системний підхід використовується безпосередньо як для управління, так і для аналізу процесів та подій, що відбуваються в системі. Запит користувача до чат-бота має ситуаційний характер. Без уточнення системних вимог на запит користувача надається неточна або невідповідна інформація. Вийти з подібної ситуації можна через конкретизацію запиту, яка може бути представлена, наприклад, онтологічною залежністю у вигляді впорядкованої множини довільних графів, у якій ребра графа визначають властивості між концептами [20]. За допомогою онтології можна прослідкувати чіткі зв'язки відношення між об'єктами та зберегти ієрархію між об'єктами при побудові запиту, що забезпечить дотримання структури інформації під час її обробки та формалізації, включаючи велику кількість різнорідних даних. Тобто фактично створення запиту до чат-бота є проходженням за ієрархічними відношеннями між різними класами об'єктів з переходами за заданими зв'язками.

Варто згадати, що користувач може не мати відповідних знань за тематикою свого запиту і не уявляти ієрархічної структури скінченної множини понять щодо об'єкта, який досліджує, а також мати вільну інтерпретацію понять і відношень між ними та не мати можливості визначити функції інтерпретації. Тобто з погляду створення запиту користувач має необроблений масив різнорідної інформації.

У цьому разі обробку інформації та формалізацію можна провести за допомогою методу, який дозволить подавати та структурувати інформацію за певними правилами. Для цього спочатку виконується вибірка інформації, яка явно чи неявно стосується теми запиту. Потім проводяться функціональні перетворення з використанням нечіткої логіки та побудовою рішень у вигляді лінгвістичних правил-продукцій та подальшої формалізації із застосуванням одного з базових методів наближення для отримання результату за максимумом відповідності.

Метод формування запиту

Використовуючи вирази (1) та (2), показник ефективності відповіді на запит користувача S можна представити рівнянням

$$S = 1 - \Delta\varphi. \tag{3}$$

До $\Delta\varphi$ можна додати не лише частину умовно невідомих знань, а й знань, визначених через $(F + N)$, які є неповними (I), суперечними (K) та застарілими (T), але присутні для обробки алгоритмами чат-бота і можуть бути наявними в сформованій відповіді користувачу. Тоді з використанням виразу (2) можна отримати вираз:

$$F + N > \Delta\varphi + I + K + T. \tag{4}$$

З урахуванням (4) показник ефективності відповіді користувачу буде представляти наступний вираз:

$$\Omega = 1 - (\Delta\varphi + I + K + T). \tag{5}$$

При розрахунку показника ефективності відповіді користувачу (Ω) слід розуміти, що це показник, до забезпечення якого варто прагнути не лише при розробці алгоритмів обробки інформації на отриманий запит, а й при формуванні запиту до чат-бота. Комплекс показників ($I + K + T$) може призвести до проблеми ще на етапі формування запиту, коли в так званій «підказці» чат-бота, що закладена в запиті користувача, є вираз чи термін, який скеровує на використання неповної, суперечної чи застарілої інформації. При цьому природний показник ($\Delta\varphi$) буде

підсилено, бо його неможливо уникнути або зменшити без надання запиту акцентів щодо використання конкретних джерел інформації, про які користувач чат-бота може не мати інформації. І в цьому разі варто підкреслити, що невизначеність може бути обумовлена запитом щодо отримання максимально повної інформації про об'єкт дослідження в одному запиті до чат-бота ($\Delta\varphi \to \max$) або при вимозі забезпечити можливий мінімум ($\Delta\varphi \to 0$) такої інформації (наприклад, запит до чат-бота надано одним словом природної мови користувача). Числове значення показника ($\Delta\varphi$) означає відхилення від заданої цільової функції та може вважатися системною похибкою. Тоді загальною умовою побудови запиту на обробку інформації з метою отримання відповіді від чат-бота виступить рівняння

$$Q-(F+N)=\Delta\varphi+I+K+T \text{ при } (\Delta\varphi+I+K+T)\to 0. \quad (6)$$

Модель (6) можна розглядати як загальну для представлення запиту природною мовою користувача до чат-бота. З неї випливають дві базові вимоги до формування запиту користувача: повнота і своєчасність; інакше будуть забезпечені вимоги виразу (3), що є загальними для опису діалектики системи обробки інформації для управління будь-якою складною системою. З неї видно, що тільки повнота знань про систему (1), їхня несуперечність і своєчасність визначають критерії (3) створення живучої складної системи.

Чат-бот за своїми алгоритмами звертається до наявної у мережі Інтернет статистичної та аналітичної інформації, на основі якої формується масив для вибору оптимальної відповіді на запит користувача. Через показник ефективності відповіді ($\Delta\Omega$) з позиції користувача чат-бота зазначене може бути описане наступним чином:

$$\Delta\Omega = \Omega\text{чат-бот} - \Omega, \quad (7)$$

де Ω — показник ефективності наявної у користувача інформації, на основі якої він має змогу самостійно, без допомоги чат-бота дати відповідь на запит; Ωчат-бот — показник ефективності інформації, яку користувач отримує за допомогою вебсервісу для більш точної відповіді на запит.

Водночас слід враховувати, що після отримання відповіді від чат-бота ефективність отриманої відповіді (інформації) починає зменшуватися:

$$\Delta\Omega \to 0 \text{ при } \Omega \to \Omega\text{чат-бот}. \quad (8)$$

Тому під час роботи з чат-ботом перед користувачем постає проблема взаємодії, яка реалізується не через вираз (8), а через розуміння необхідного/достатнього для отримання відповіді на запит. Наприклад, якщо інформація Б, отримана на запит А, задовольняє користувача, вона вважається ефективною.

Описане виразами (3)–(8) можна представити через алгоритм за правилами нечіткої логіки. Припустимо, що користувач природною мовою ввів у запит до чат-бота термін, щодо якого слід здійснити пошук тлумачення з відповідними варіантами використання при описі події чи явища. У цьому разі формування запиту до чат-бота повинно бути сформоване за наступними кроками.

Крок 1. Початковою точкою пошуку обираємо термін або ключове слово, що вводить користувач, наприклад terminus.

Крок 2. Обираємо терми лінгвістичних змінних, які відповідають підмножині А. Термом вважаємо деякий логічний вираз, що стосується об'єкта дослідження. Тобто проводимо пошук компонентів (слів, виразів), які природною мовою описують цей термін. Це може бути пошук слів та виразів, які найчастіше вживаються при описі цього терміну в публікаціях, спілкуванні, наявних базах даних та ін.

Крок 3. Для кожного терму обираємо значення, яке характеризує його найточніше.

 ISSN 2786-6491

Крок 4. З відсортованих термів формуємо підмножину Б та знаходимо в ній ідентичні компоненти, які за різних обставин також можуть описати заданий у пошуку термін, чітко дотримуючись правила комутативності щодо термів з множин А та Б.

Крок 5. Здійснюємо пошук відповідних значень з обраних на кроці 4 характеристик у підмножині Б з присвоєнням 1 або 0 кожному з отриманих значень.

Крок 6. Після отримання екстремальних значень обираємо проміжні значення, кроки до яких можуть не виконуватися по прямій, а описуватися різноманітними функціями.

Крок 7. Присвоюємо 1 або 0 кожному з отриманих проміжних значень.

Крок 8. Всім значенням присвоюємо відповідні функції стандартних приналежностей.

Крок 9. Формуємо інформаційний масив для створення запиту до чат-бота.

Крок 10. Визначаємо продукційні правила створення запиту до чат-бота.

Під час роботи з чат-ботом продукційні правила можуть бути задані наступним чином:

— якщо використано термін terminus, може бути характеристика *C* чи *D* або уточнююче слово *G*;

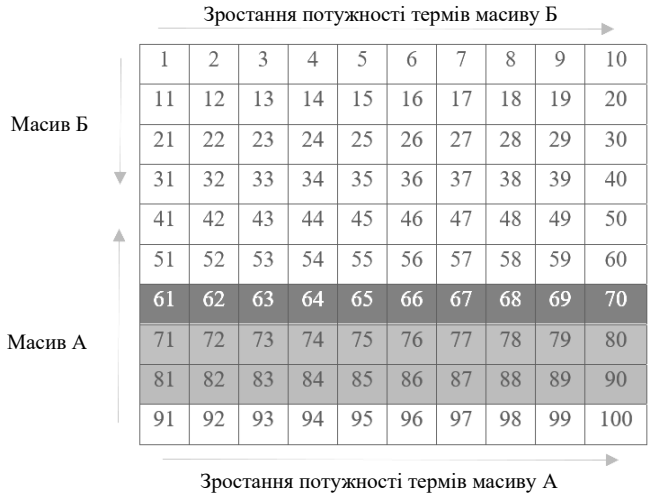
— якщо використано уточнююче слово *G*, — характеристика *Q*;

— якщо використано характеристику *Q*, також може бути характеристика *W* або уточнююче слово *U* чи інші, які виконуються послідовно, включаючи слова із заміщенням лівої та правої частин.

Результат пошуку буде обумовлено відповідністю хоча б за одним продукційним правилом або частин продукційних правил пошуку з виконанням (8). Тоді буде отримано низку термінів з характеристиками, які при формуванні запиту до чат-бота дозволять отримати найбільш ефективну відповідь $\Delta\Omega$.

Результати комп'ютерного експерименту

Для проведення комп'ютерного експерименту за допомогою доступних в мережі тезаурусів за предметною областю та за вмістом або із застосуванням програм аналізу текстів сформовані підмножини термінів А та Б з використанням кроків 2–4 наведеного вище алгоритму. З отриманих термінів сформовано інформаційний масив, а загальна кількість значень упорядкованого набору елементів обмежена діапазоном (1–100) (рис. 1).

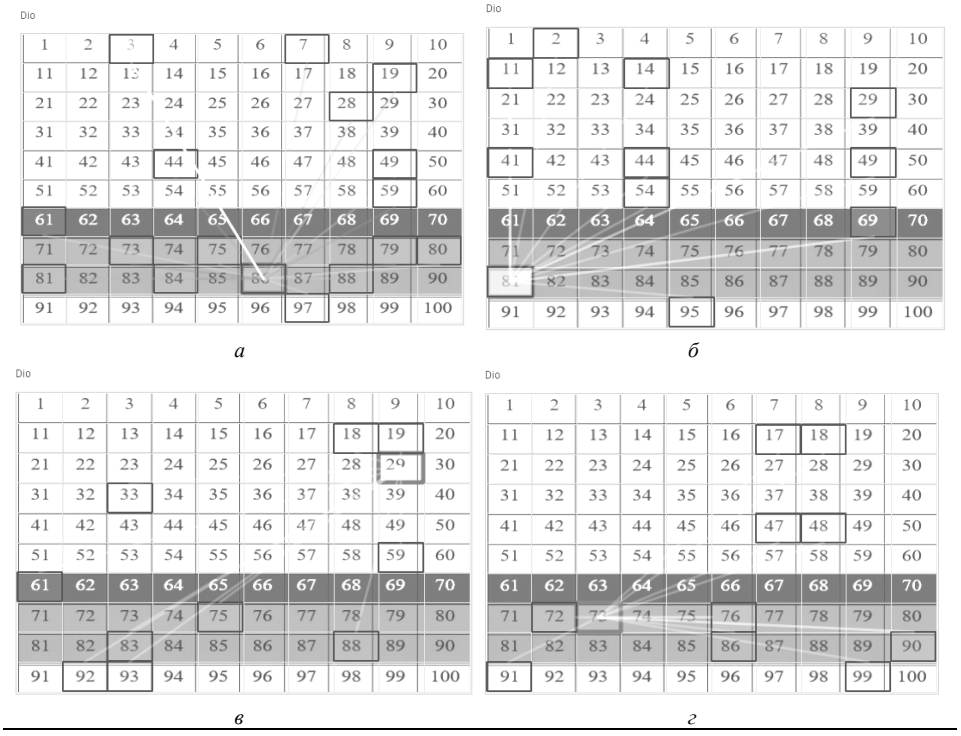


1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Рис. 1

Масиви наведені таким чином, що терми в кожному вузлі підсилюють ключові слова в порядку зростання. Розмірність масивів може бути різною. Початковий термін, який узагальнює запит, також вноситься до масиву А — він виступає одним з ключових слів для пошуку. Прийнято, що ключових слів для пошуку може бути декілька. Застосовано алгоритм, викладений у роботі [21], та його дещо видозмінену програмну реалізацію з метою полегшення та візуалізації розрахунків при вирішенні поставленої задачі. На відміну від [21], початкова точка пошуку не обирається довільно, а призначається за одним з ключових слів. Усі ключові слова за пошуковим запитом розташовано у другому рядку знизу на рис. 1. В останньому рядку отримано терми, що надають надлишкову інформацію. Для полегшення розуміння алгоритму кожен терм має свій порядковий номер.

Далі реалізуються кроки 5–8 наведеного алгоритму. На цих кроках за методом можливих напрямків, зокрема відшукання можливого напрямку з використанням методу Зойтендейка, серед термів відшукується той, який з ключовим словом дозволяє отримати максимальне значення під час пошуку в мережі. Наприклад, комбінація термінів *N* та *L* дозволяє отримати найбільшу кількість результатів при заданні їх як слів для пошуку в будь-якій пошуковій системі мережі Інтернет. На рис. 2 наведено візуалізацію результатів розрахунків за описаним методом: *а* — ключове слово з комірки 85 та комірки зі словами, що під час пошуку дозволяють отримати найбільшу кількість результатів; *б* — такий самий розрахунок з іншим ключовим словом з незмінною темою запиту; *в* — використання як початкової точки неключового слова з масиву Б надає результати про необхідність застосування ключових слів для точності запиту; *г* — початкова точка з використанням неключового слова дає найкращі результати із застосуванням ключових слів, слів з масиву А та термів з надлишковою інформацією.



а *б* *в* *г*

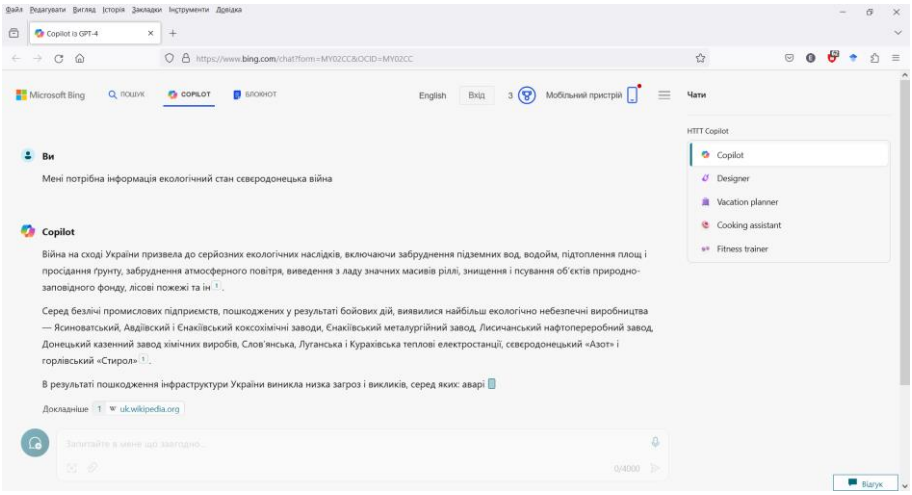
Рис. 2

Під час комп’ютерного експерименту проведено понад 1,2 тис. розрахунків з метою формування масиву слів для пошуку інформації (крок 9 наведеного вище алгоритму) за допомогою чат-ботів та шляхом створення 100 пошукових запитів.

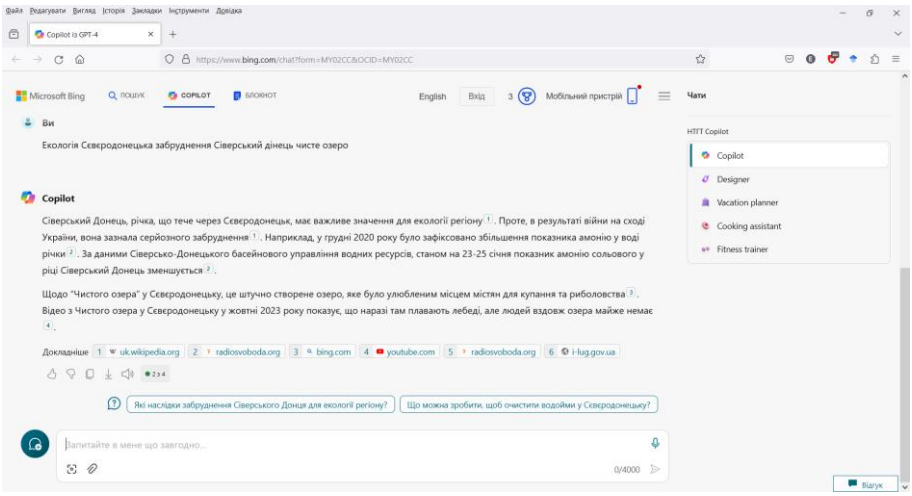
 ISSN 2786-6491

Продукційні правила для пошуку, зазначені на кроці 10 алгоритму, формувалися за надлишковістю результату — при задаванні як початкової точки іншого ключового слова деякий терм потрапляв до інформаційних масивів і першого, і другого ключових слів.

На рис. 3 (*а* — введення запиту до чат-бота природною мовою без обробки інформації та визначення ключових слів; *б* — введення запиту до чат-бота з використанням слів сформованого інформаційного масиву) наведено приклади побудови запитів до чат-бота GitHub Copilot з використанням слів зі сформованих інформаційних масивів за представленим методом для отримання інформації про екологічний стан м. Сєвєродонецька (Луганська область) під час воєнних дій. Таке формулювання запиту цікаве тим, що Сєвєродонецьк окуповано з 2022 року, тому інформація про екологічний стан території обмежена, неточна і несвоєчасна, а метою звернення до чат-бота є саме отримання інформації із забезпеченням її властивостей.



а



б

Рис. 3

Аналізуючи відповідь чат-бота, наведену на рис. 3, *а*, можна помітити абстрактність висловлювань з використанням узагальнень і одним посиланням на ресурс, на базі якого формувалася відповідь. Наведена інформація стосується Луганської та Донецької областей, тому відповідь неточна. Час реакції системи чат-бота на запит складає 4 с.

Відповідь на запит користувача до чат-бота, наведена на рис. 3, *б*, більш точна та має шість посилань на джерела, які дозволили сформувати наведений текст. Відповідь так само має незначну неточність, яка стосується Сіверського Донця (річки, що розмежовує Сєвєродонецьк та Лисичанськ). Проте ця неточність обумовлена обмеженою інформацією від зазначених ресурсів, бо територія окупована, більшість дописувачів ніколи там не перебували, а у дописах є наявні логічні помилки — Сєвєродонецьк ототожнюється з Сіверським Донцем. Час реакції системи чат-бота на запит складає 7 с.

Для запиту, який разом з відповіддю наведено на рис. 3, *б*, використовувалися розрахунки, візуалізацію яких наведено на рис. 2. Обумовлена неточність відповіді пов'язана з обмеженням, в якому для візуалізації кінцевого результату використаний не інформаційний масив, а вибірка, що складалася з термів, які мали порядкові номери 44, 49, 73, 86, 88. У термінах це відповідно «Сіверський Донець», «Чисте озеро», «екологія», «забруднення Сєвєродонецьк». Продукційне правило задавалося наступним чином:

— якщо як ключове вживається слово «забруднення», то використовується уточнення «де» — «Сєвєродонецьк»; уточнення може виступати одним з ключових слів, звужуючи сферу пошуку відповіді на запит;

— якщо як ключове вживається слово «забруднення», то характеристика — «екологія»;

— якщо як ключове вживається слово «Сєвєродонецьк», також уточнюючим може бути слово «Сіверський Донець»;

— якщо як ключове вживається слово «Сєвєродонецьк», також уточнюючим може бути слово «Чисте озеро».

Економічне обґрунтування запропонованого методу дозволяє констатувати збільшення часу на підготовку запиту до чат-бота та отримання відповіді приблизно на 75 % проти запиту користувача природною мовою без підготовки та підказок у вигляді ключових слів та термів предметної області. Проте ефективність отриманої відповіді (інформації) за моделлю (7) зростає з 0,44 до 0,89. Таке зростання дає підстави говорити, що в цілому запропонований метод дозволяє на 14 % надавати більш точну, достовірну, повну, корисну та зрозумілу відповідь. Витрати часу на підготовку запитів можна мінімізувати шляхом розробки вебсервісів за заданим алгоритмом, які дозволятимуть користувачам отримувати інформаційні масиви для запиту та будувати з них запити за продукційними правилами, що є перспективним напрямом подальших досліджень.

Висновок

У роботі представлено один з можливих методів формування запитів до чат-ботів з породжувальним ШІ, які зараз набирають популярності в Інтернет-користувачів (ChatGPT, Gemini, GitHub Copilot та ін.). Метод заснований на забезпеченні подачі та структуруванні інформації за певними продукційними правилами. Для цього на перших етапах реалізації методу здійснюється вибірка інформації та інформація розподіляється за критерієм зв'язку з темою запиту (точно чи опосередковано). Після цього проводяться функціональні перетворення з використанням нечіткої логіки та побудовою рішень у вигляді лінгвістичних правил-продукцій і подальшої формалізації із застосуванням одного з базових методів наближення для отримання максимально відповідного результату. Метод націлений на подолання розриву між системним підходом при розробці алгоритмів чат-бота та ситуаційним підходом при створенні запиту користувачем. У кінцевому підсумку користувач отримує сформовану вибірку термінів для створення аргументованого запиту та ймовірну схему представлення термінів у запиті за продукційними правилами з можливістю конструювання у запиті підказок для отримання точної, достовірної, повної, корисної та зрозумілої відповіді.

Зазначене можна використовувати для розробки вебсервісів, які дозволятимуть користувачам отримувати інформаційні масиви для запиту та будувати з них запити з забезпеченням можливості отримати найбільш ефективну відповідь.

O. Kryazhych, O. Vasenko, L. Isak, O. Babak, V. Grytsyshyn

METHOD OF CONSTRUCTING REQUESTS TO CHAT-BOTS ON THE BASE OF ARTIFICIAL INTELLIGENCE

Olha Kryazhych

Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, Kyiv,

economconsult@gmail.com

Oleksandr Vasenko

Hryhorii Skovoroda University in Pereiaslav, Ukraine,

vasenko.olexandr@gmail.com

Ludmyla Isak

Hryhorii Skovoroda University in Pereiaslav, Ukraine,

isakluda@ukr.net

Oleksandr Babak

Hryhorii Skovoroda University in Pereiaslav, Ukraine,

babak1109@gmail.com

Volodymyr Grytsyshyn

Volodymyr Dahl East Ukrainian National University, Kyiv,

hrytsyshynvo@gmail.com

On the Internet, chatbots — ChatGPT, Gemini, GitHub Copilot and others — with generative artificial intelligence are gaining more and more popularity. A chatbot is used for training and intelligent search for information. A correctly constructed query gives a clear result of the answer from the chatbot and facilitates the interaction process. The task of the work is to present a mathematically based method of building queries to chatbots, taking into account the features that allow you to form a query in such a way that the provided answer meets the requirements for the main properties of information. The paper analyzes the possibility of generating queries to the transformation chatbot using fuzzy logic and building solutions in the form of linguistic rules-products and further formalization using one of the basic approximation methods to obtain a result based on maximum correspondence. In particular, the implementation of an algorithm of ten basic steps is proposed, which implements the possibility of choosing the terms of linguistic variables that correspond with different degrees of correspondence to one of the two formed sets. The term in the work is understood as a logical expression related to the object of research. Preliminary results of testing the proposed method revealed a 14 % ability to provide a more accurate, reliable, complete, useful and understandable answer. The time spent on query preparation can be minimized by developing web services according to a given algorithm, which will allow users to receive information arrays for a query and build queries from them according. This can be used in the development of information processing web services.

Keywords: generative algorithm, language model, production rules, array, prompt engineering, indicator of response efficiency, ontology.

ПОСИЛАННЯ

1. Baker P. Chatgpt. Hoboken NJ: John Wiley & Sons, 2023. 159 p. ISBN. 9781394204632.
2. Ferguson M. Prompt engineering: the future of language generation (1). North Charleston : Independently published, 2023. 79 p. ISBN: 9798215905739.
3. Su Y., Lin Y., Lai C. Collaborating with chatGPT in argumentative writing classrooms. *Assessing Writing*. 2023. Vol. 57. P. 100752. DOI: https://doi.org/10.1016/j.asw.2023.100752
4. Dwivedi Y.K., Kshetri N., Hughes L., Slade E.L., Jeyaraj A., Kar A.K., Baabdullah A.M., Koohang A., Raghavan V., Ahuja M., Albanna H., Albashrawi M.A., Al-Busaidi A.S., Balakrishnan J., Barlette Y., Basu S., Bose I., Brooks L., Buhalis D. et al. Opinion paper: «So what if chatGPT wrote it?» Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management.* 2023. Vol. 71. P. 102642. DOI: https://doi.org/10.1016/j.ijinfomgt.2023.102642
5. Енциклопедія кібернетики: у 2 т. / за ред. В.М. Глушкова та ін. Київ : Головна редакція української радянської енциклопедії, 1973. 1228 с.
6. Darvishi A., Khosravi H., Sadiq S., Gašević D., Siemens G. Impact of AI assistance on student agency. *Computers & Education*. 2024. Vol. 210. P. 104967. DOI: https://doi.org/10.1016/j.compedu.2023.104967
7. Kang J., Yi Y. Beyond ChatGPT: multimodal generative AI for l2 writers. *Journal of Second Language Writing*. 2023. Vol. 62. P. 78–92. DOI: https://doi.org/10.1016/j.jslw.2023.101070
8. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A. Language models are few-shot learners. *Advances in Neural Information Processing Systems*. 2020. Vol. 33. P. 1877–1901.
9. Phoenix J., Taylor M. Prompt engineering for generative AI : future-proof inputs for reliable AI outputs at scale (1st Edition). Sebastopol : O'Reilly Media, 2024. 396 p. OCLC: 1420047512.
10. Волошин О.Ф., Мащенко С.О. Моделі та методи прийняття рішень: навч. посіб. для студ. вищ. навч. закл. Вид. 2-ге, перероб. і допов. Київ : Видавничо-поліграфічний центр «Київський університет», 2010. 336 с.
11. Lester B., Al-Rfou R., Constant N. The power of scale for parameter-efficient prompt tuning. *Proceedings of the 2021. Conference on Empirical Methods in Natural Language Processing*. 2021. С. 3045–3059. DOI: https://doi.org/10.18653/V1/2021.EMNLP-MAIN.243
12. Shin T., Razeghi Y., Robert L. Logan IV, Wallace E., Singh S. AutoPrompt: eliciting knowledge from language models with automatically generated prompts. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics*. 2020. С. 4222–4235. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.346
13. Pikies M., Ali J. Analysis and safety engineering of fuzzy string matching algorithms. *ISA Transactions*. 2021. Vol. 113. P. 1–8. ISSN 0019-0578. DOI: https://doi.org/10.1016/j.isatra.2020.10.014.
14. Nahin P.J. The logician and the engineer: how George Boole and Claude Shannon created the information age. Princeton : Princeton University Press, 2017. 248 p. DOI: https://doi.org/10.23943/princeton/9780691176000.001.0001
15. Глушков В.М. Кибернетика. Вопросы теории и практики. М. : Наука, 1986. 488 с.
16. Глушков В.М., Стогний А.А., Афанасьев В.Н. Автоматизированные информационные системы. М. : Знание, 1973. 64 с.
17. Глушков В.М. Введение в АСУ. 2-е изд. испр. и дополн. Киев : Техника, 1974. 320 с.
18. Myerson R.B. Game theory: analysis of conflict. Cambridge, Massachusetts : Harvard University Press, 1991. P. 600. ISBN 0-674-34116-3. DOI: https://doi.org/10.2307/j.ctvjsf522
19. Ramin A., Md Muzahid K., Haque M.H., Rahman M.H. An approach to develop a dynamic job shop scheduling by fuzzy rule based system and comparative study with the traditional priority rules. *American Journal of Engineering and Applied Sciences*. 2016. N 9. P. 202–212. DOI: https://doi.org/10.3844/ajeassp.2016.202.212
20. Стрижак О.Є. Засоби онтологічної інтеграції і супроводу розподілених просторових та семантичних інформаційних ресурсів. *Екологічна безпека та природокористування*. 2013. № 12. С. 166–177.
21. Трофимчук О.М., Кряжич О.О., Коваленко О.В. Алгоритм визначення початкової точки при моделюванні за методом можливих напрямків. *Радіоелектроніка, інформатика, управління.* 2019. № 3. С. 40–46. DOI: https://doi.org/10.15588/1607-3274-2019-3-5

Отримано 07.03.2024