

УДК 004.8

DOI [https://doi.org/10.15589/znп2022.2\(489\).9](https://doi.org/10.15589/znп2022.2(489).9)**MODERN TRENDS IN THE DEVELOPMENT OF SYSTEMS
FOR THE NATURAL LANGUAGE RESPONSES FORMATION
WHEN SEARCHING IN THE TEXT DOCUMENT DATABASES****СУЧАСНІ ТЕНДЕНЦІЇ РОЗРОБКИ СИСТЕМ ФОРМУВАННЯ
ПРИРОДНОМОВНИХ ВІДПОВІДЕЙ ПРИ ПОШУКУ
В БАЗАХ ТЕКСТОВИХ ДОКУМЕНТІВ**

Mykhailo Yu. Vyshniak
mykhailo.vyshniak@nure.ua
ORCID: 0000-0002-3240-139X
Mykhailo Yu. Pyrozhenko
mykhailo.pyrozhenko@nure.ua
ORCID: 0000-0003-0785-1273

М. Ю. Вишняк,
канд. техн. наук, доцент
М. Ю. Пироженко,
аспірант

Kharkiv National University of Radio Electronics, Kharkiv
Харківський національний університет радіоелектроніки, м. Харків

Abstract. Question-Answering systems are information systems for finding answers to questions that users ask in natural language. This discipline is one of the main areas of natural language text processing and information retrieval, which has practical applications in various fields, such as education, tourism, health care, and personal assistance. Robust text analysis techniques give us a way to understand or synthesize large amounts of text without reading it. Research in this field saves users efforts and enables them to make quick decisions.

Purpose. Analysis of trends of development of the methodological base for the development of information systems capable of answering factual questions based on information represented by unstructured text data collections.

Method. The research was carried out using a system approach, methods of abstraction, system analysis, comparison, and synthesis.

Results. An analysis of the current situation with the development of systems for the formation of natural language responses was carried out. The scheme of operation of such systems is defined. Modern trends in the improvement of the analyzed systems have been revealed.

Scientific novelty. The actual scientific task was solved, which consists in the theoretical and methodological substantiation of the approaches and principles of building natural language text processing systems, taking into account the latest trends.

Practical importance. It consists in the possibility of applying scientific provisions and research conclusions for the development and implementation of question-answering systems; in the process of teaching natural language processing disciplines in higher educational institutions; for writing manuals of natural language processing; during applied research.

Key words: question-answering systems; text analytics; information search methods; named entity recognition; natural language processing; multiple choice problems.

Анотація. Система надання відповідей – це інформаційна система для пошуку відповідей на запитання, які користувач ставить природною мовою. Ця дисципліна є одним з головних напрямів обробки текстів природних мов та інформаційного пошуку, що має практичне застосування в різних сферах, таких як освіта, туризм, охорона здоров'я та особиста допомога. Надійні методи аналізу тексту дають нам спосіб зрозуміти або синтезувати великі обсяги тексту, не читаючи їх. Дослідження цієї галузі дозволяють економити зусилля користувачів та надають їм можливість приймати швидкі рішення.

Мета. Аналіз стану розвитку методологічної бази розробки інформаційних систем, що здатні відповідати на фактографічні питання, на базі інформації, представленої неструктурованими текстовими колекціями даних.

Методика. Дослідження проводилось з використанням системного підходу, методів абстрагування, системного аналізу, порівняння та синтезу.

Результати. Проведено аналіз сучасної ситуації з розробкою систем формування природномовних відповідей.

Визначено схему роботи таких систем. Виявлено сучасні тенденції вдосконалення аналізованих систем. Наукова новизна. Було вирішено актуальне наукове завдання, що полягає в теоретико-методологічному обґрунтуванні підходів та принципів побудови систем обробки текстів природних мов з урахуванням останніх тенденцій.

Практична значимість. Полягає в можливості застосування наукових положень і висновків дослідження для розробки та впровадження систем надання відповідей; у процесі викладання дисциплін з обробки природних мов у вищих навчальних закладах; під час написання посібників з обробки природних мов; під час проведення прикладних досліджень.

Ключові слова: системи надання відповідей; текстова аналітика; методи інформаційного пошуку; ідентифікація іменованих об'єктів; обробка природної мови; проблема множинного вибору.

ПОСТАНОВКА ЗАДАЧІ

Пошукові системи представляють користувачеві ранжований список релевантних документів у відповідь на сформульовані користувачем слова, виходячи з різних факторів, таких як показники авторитетності джерела та відповідності документа ключовим словам. Однак це не вирішує глобальну проблему пошуку інформації, оскільки користувачам, як і раніше, доводиться вивчати кожен документ окремо.

Системи надання відповідей, на відміну, від звичайних традиційних систем інформаційного пошуку, аналізують тексти, а тому дозволяють зменшити час користувача на отримання потрібного документу. Користувач, при роботі з системою, ставить запитання та отримує сфокусовану відповідь. Побудова якісних систем надання відповідей та постійний їх розвиток потребує виявлення та аналізу тенденцій у цій галузі, насамперед, на основі провідних наукових публікацій.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ І ПУБЛІКАЦІЙ

Як свідчать публікації, що незважаючи на актуальність досліджень, не багато науковців займаються цими питаннями.

Слід відзначити працю «Recent Trends in Deep Learning Based Open-Domain Textual Question Answering Systems»[1]. У статті узагальнюється література з глибокого навчання та приводиться класифікація методів в системах надання відповідей, які працюють з неструктурованими джерелами даних, питаннями фактографічного типу, відповідями на основі вилучення даних та у загальному домені.

Важливою є також стаття «A survey on question answering systems over linked data and documents»[2]. Автори статті наводять приклади існуючих систем надання відповідей та роблять класифікацію методів та проблем, з якими стикаються розробники при побудові систем. Також наводять набори готових колекцій питань-відповідей для навчання нейронних мереж.

Стаття «Multiple choice questions: answering correctly and knowing the answer»[3] описує проблему множинного вибору та пропонує тести для оцінки якості відповідей. Результати дослідження можуть

використовуватися при побудові наборів питань-відповідей для навчання та контролю систем.

ВІДОКРЕМЛЕННЯ НЕВИРІШЕНИХ РАНІШЕ ЧАСТИН ЗАГАЛЬНОЇ ПРОБЛЕМИ

На сьогодні фактично відсутній огляд сучасних методів та моделей систем надання відповідей, а також тенденцій розвитку у цій галузі, який би обґрунтовано прогнозував подальшу еволюцію систем та відповідей на питання щодо доцільності використання тих чи інших підходів при їх побудові.

МЕТА ДОСЛІДЖЕННЯ

Визначення сучасного стану розвитку методологічної бази розробки інформаційних систем, що здатні відповідати на фактографічні питання на базі інформації, представленої неструктурованими текстовими колекціями даних, враховуючі останні тенденції розвитку в цій галузі обробки текстів природних мов.

МЕТОДИ, ОБ'ЄКТ ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

Дослідження проводилось з використанням системного підходу, методів абстрагування, системного аналізу, порівняння та синтезу. Об'єктом дослідження є моделі та методи систем надання відповідей. Предметом дослідження є тенденції в галузі обробки текстів природних мов.

Основний матеріал.

Сьогодні значна частина неструктурованої інформації представлена у вигляді текстів на природних мовах. За деякими винятками, тексти, з яким доводиться працювати, написані людьми. Письмовий текст, сам по собі є продуктом глибокого соціального явища, в основі якого стоять люди. Інформація в цих текстах містить багато корисних даних та описів знань, необхідних для вирішення практичних завдань.

З 1960-х років розроблено велику кількість систем надання відповідей, пов'язаних з різними форматами джерел даних, типами питань та формами відповіді[4]. Ці системи використовуються в різних предметних областях, таких як медицина[5][6], культура[7], освіта[8], туризм[9], комерція[10] та інших.

Системи формування природньомовних відповідей працюють за схемою, що включає в себе аналіз

питання, пошук документів та уривків, ідентифікацію сутностей та представлення відповіді.

Першим етапом у процесі роботи будь якої системи надання відповіді є аналіз запитання. Системі надання відповідей необхідно проаналізувати запитання, введене користувачем, та перетворити його у форму або форми, які потрібні наступним частинам системи. Вхідними даними на цьому етапі є питання користувача, що представлені у вигляді рядка тексту.

На етапі аналізу запитання, до системи надання відповідей можуть застосовуватися обмеження. Система може попросити користувача уточнити запитання[11]. Від користувача може знадобитися використовувати підмножину природної мови, яка обмежена з точки зору лексики та синтаксису. Можливо, користувач буде використовувати певний інтерфейс формування запитань[12][13], що значно спрощує завдання інтерпретації запитання.

Система має отримати тип питання, тип відповіді та типи відносин, про які йдеться у запитанні. Для цього система надання відповідей вдається до аналізу питання, який зазвичай включає визначення семантичного типу сутності, який відповідає класу запитання та визначити додаткові обмеження для об'єкта відповіді[14].

Виключенням є IBM Watson[15]. Ця система не володіла інформацією про очікуваний тип відповіді. Стандартними питаннями в телеграмі, для якої була розроблена система, були два твердження про неназваний об'єкт, який потрібно було визначити гравцям. IBM Watson генерував відповіді для різних типів та підраховував бали для кожного з кандидатів на відповідь.

Для різних систем застосовуються різні ієрархії типів запитань на основі типів відповідей. В роботах [14] та [16] були запропоновані критерії класифікації фактографічних питань.

Можна стверджувати, що побудова ієрархії типів запитань для систем надання відповідей не є тривіальним завданням. Різні питання потребують різної стратегії пошуку відповіді, а системи надання відповідей можуть значно відрізнятися між собою. Побудова ієрархії типів запитань залежить виключно від вимог, що ставиться до такої системи.

На другому етапі роботи, системі потрібно проаналізувати документи. Вхідними даними є векторні представлення питання користувача та колекції документів, які представлені у вигляді набору текстів природної мови.

Джерела інформації можуть бути різними. Найчастіше це статті з Вікіпедії[2]. Це великі обсяги неструктурованих даних, у яких міститься багато фактів та інформації про об'єкти нашого світу.

Системі надання відповідей потрібно відібрати підмножину документів із загальної колекції документів, зазвичай на кілька порядків менше початкової, що включає ті документи, які, як вважається, найімовірніше можуть містити правильну відповідь на

запитання. Пошук релевантних документів передусім потрібен для того, щоб відсіяти загальну колекцію до меншого набору, який необхідно детально розглянути вже більш точними методами. Це дозволяє раціональним чином використовувати наявні ресурси.

Процес пошуку документів передбачає виділення ключових слів із запитання, які будуть використані для порівняння з словами документів колекції. Цей набір ключових слів можна розширити, використовуючи синоніми або морфологічні варіанти слів[17].

Для пошуку потрібних документів використовують математичні методи, що дозволяють прогнозувати, які з документів найбільш релевантні по відношенню до запиту користувача. На сьогодні найбільш поширеними моделями інформаційного пошуку є LMIR[18], TF-IDF[19], Okapi BM25[20].

В декількох наукових роботах були порівняні моделі інформаційного пошуку[16][21]. Результати досліджень продемонстрували, що ефективність окремих моделей може суттєво розрізнятися в залежності від умов самих експериментів. В середньому, TF-IDF виявився найбільш точним.

До колекцій документів також можуть застосовуватися обмеження. Припускаючи, що система має доступ до великої колекції документів, як до ресурсу знань, для відповідей на можливі запитання, цю колекцію треба обробити, щоб перетворити її у форму, яка більш підходить для відповідей на запитання в режимі реального часу.

Для деяких випадків є потреба використовувати пошукову систему засновану на рейтингу документів, а також прийняти рішення щодо того, скільки документів буде розглянуто системою. За умови, якщо на запитання потрібно відповідати в режимі реального часу, системи роблять попередню індексацію документів[22].

Знайдені документи сортируються відповідно до ймовірності того, що моделі цих документів побудували модель запитання. Відібрані документи розбиваються на уривки, часто на окремі речення, та знов сортируються до потрібного рівня декомпозиції.

Наступним етапом, після звуження кількості документів, є глибокий аналіз. З цих речень, що знаходяться у неструктурованому вигляді, потрібно витягти певну частину інформації в структурованому вигляді. Це означає, що на базі відібраних уривків, системі необхідно проаналізувати та вилучити іменовані об'єкти та зв'язки між ними.

Системі надання відповідей потрібно розпізнати та класифікувати слова, інколи багатослівні рядки, відповідно семантичному класу імен[23]. Для класифікації використовуються бібліотеки ідентифікації іменованих об'єктів, найбільш поширеними інструментами є CoreNLP[24], NLTK[25] та Spacy[26].

Останнім етапом роботи системи надання відповіді є представлення відповіді користувачу. Щоб відповісти на запитання, системам надання відповідей

потрібно, звернувшись до наявних у неї ресурсів, знайти одну або кілька можливих за семантичним типом сутностей. Виявлення кількох сутностей потрібного типу породжує проблему множинного вибору.

Система сортирує кандидатів на відповідь відповідно до ймовірності їх правильності. Кінцеві відповіді на запитання користувача, при достатній вірогідності правильності, надаються або процедурою вилучення, або шляхом генерації.

Ознайомлення з проблемами та аналіз рішень з розробки систем дав можливість визначити низку тенденцій щодо подальшого вдосконалення систем.

Перша тенденція: пошук стратегій вирішення проблеми множинного вибору відповіді. Проблема з множинним вибором відповідей полягає у тому, що кілька варіантів відповідей надаються разом із контекстом[3][27]. Авторами [28] пропонується використовувати модель seq2seq для навчання систем обирати правильну відповідь. Засновуючись на положенні, що відповідь на одне запитання може одночасно міститись у декількох різних джерелах, нами пропонується додатково робити аналіз по всій колекції документів.

Наступна тенденція: профілювання користувачів. Реальні системи надання відповідей не існують незалежно, зазвичай вони є лише частинами машин інформаційного пошуку або діалогових систем, відтак питання можна інтерпретувати або в контексті діалогу з персональними помічниками, або з урахуванням попередніх запитів до пошукової системи. Профілюючи вподобання окремих користувачів можна враховувати їх індивідуальні інформаційні потреби.

Також є тенденція поширення технологій автоматичного машинного навчання. Автори [1] роблять висновок, що в міру збільшення навчальних наборів даних, міркування, алгоритми, архітектури систем надання відповідей будуть ставати все більш складними, з інтеграцією технологій від інших галузей, та пропонують прискорити обчислення за допомогою технологій автоматичного машинного навчання.

Розробка якісних систем надання відповідей на базі моделей штучного інтелекту потребує великих навчальних наборів даних, складність побудови яких обмежує розвиток галузі.

Тенденцією стає і рішення проблеми локалізації, або орієнтації на одну мову [29][30]. Більшість проаналізованих систем надання відповіді працюють лише з однією мовою, найчастіше з англійською.

В колекціях для менш поширених мов може не бути потрібних для відповіді документів. Це спонукає використовувати сервіси машинного перекладу для конвертації тексту питання до більш поширених мов.

ОБГОВОРЕННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ.

В результаті численних досліджень останніх років, присвячених системам надання відповіді, можна відмітити, що вони знайшли застосування в багатьох областях, таких як медицина, культура, туризм та інші. Показовим прикладом успішної системи є IBM Watson, яка не була обмежена певною областю, будучи націленою на широкий спектр питань.

В останні роки набули поширення персональні помічники на основі систем надання відповідей, включаючи Siri від Apple, Alexa від Amazon, Cortana від Microsoft, Bixby від Samsung та Google Assistant. Сучасні пошукові машини також рухаються в напрямку прямого надання відповідей. Крім того, що сучасні системи інформаційного пошуку знаходять релевантні запити матеріали, їх цінність також у тому, що у відібраних документах вони виділяють фрагменти тексту, які можуть містити потрібну інформацію.

Незважаючи на те, що сучасні системи надання відповідей досягли досить значного прогресу, вони все ще далекі від застосування у повсякденні та потребують розробки нових моделей та методів для вирішення визначених проблем. Оскільки кількість цифрованих текстів постійно зростає, можна прогнозувати збільшення потреби у системах текстової аналітики в майбутньому.

ВИСНОВКИ

В роботі проведено дослідження стану розвитку методологічної бази розробки інформаційних систем, що здатні відповідати на фактографічні питання користувачів, на базі інформації, отриманої з неструктурованих текстових колекцій даних, проблем та останніх тенденцій, що відбуваються у галузі.

Як наслідок, на основі дослідження проблем та аналізу рішень з розробки систем, було встановлено, що основними тенденціями у галузі розробки систем надання відповідей є побудова багатомовних систем, пошук стратегій вирішення проблеми множинного вибору відповіді, профілювання користувачів та поширення технологій автоматичного машинного навчання для прискорення систем.

REFERENCES

- [1] Huang, Z., Xu, S., Hu, M., Wang, X., Qiu, J., Fu, Y., ... & Wang, C. (2020). Recent trends in deep learning based open-domain textual question answering systems. *IEEE Access*, 8, 94341-94356.
- [2] Dimitrakis, E., Sgontzos, K., & Tzitzikas, Y. (2020). A survey on question answering systems over linked data and documents. *Journal of intelligent information systems*, 55(2), 233-259.
- [3] McKenna, P. (2019). Multiple choice questions: answering correctly and knowing the answer. *Interactive Technology and Smart Education*.
- [4] Mishra, A., & Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3), 345-361.

- [5] Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., & Socher, R. (2021). COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ digital medicine*, 4(1), 1-9.
- [6] Mutabazi, E., Ni, J., Tang, G., & Cao, W. (2021). A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12), 5456.
- [7] Yuan, Z., Sun, S., Duan, L., Li, C., Wu, X., & Xu, C. (2020). Adversarial multimodal network for movie story question answering. *IEEE Transactions on Multimedia*, 23, 1744-1756.
- [8] Goel, A. K., & Polepeddi, L. (2018). Jill Watson: A virtual teaching assistant for online education. In *Learning engineering for online education* (pp. 120-143). Routledge.
- [9] Scheider, S., Nyamsuren, E., Krüger, H., & Xu, H. (2021). Geo-analytical question-answering with GIS. *International Journal of Digital Earth*, 14(1), 1-14.
- [10] Gao, S., Chen, X., Ren, Z., Zhao, D., & Yan, R. (2021). Meaningful answer generation of e-commerce question-answering. *ACM Transactions on Information Systems (TOIS)*, 39(2), 1-26.
- [11] Guo, D., Tang, D., Duan, N., Zhou, M., & Yin, J. (2018). Dialog-to-action: Conversational question answering over a large-scale knowledge base. *Advances in Neural Information Processing Systems*, 31.
- [12] Morton, K., Wang, P., Bizon, C., Cox, S., Balhoff, J., Kebede, Y., ... & Tropsha, A. (2019). ROBOKOP: an abstraction layer and user interface for knowledge graphs to support question answering. *Bioinformatics*, 35(24), 5382-5384.
- [13] Wenyan, L., Hao, T., Chen, W., & Feng, M. (2009). A web-based platform for user-interactive question-answering. *World Wide Web*, 12(2), 107-124.
- [14] Moldovan, D., Pasca, M., Harabagiu, S., & Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS)*, 21(2), 133-154.
- [15] IBM Watson: Ramprasath, M., & Hariharan, S. (2012). A survey on question answering system. *International Journal of Research and Reviews in Information Sciences*, 2(1).
- [16] Deepak Kumar Gupta, Rajkumar Pujari, Asif Ekbal, Pushpak Bhattacharyya, Anutosh Maitra, Tom Geo Jain, & Shubhashis Sengupta (2018). Can Taxonomy Help? Improving Semantic Question Matching using Question Taxonomy. In *COLING*.
- [17] Chen, X., Chen, C., Zhang, D., & Xing, Z. (2019). Sthesaurus: Wordnet in software engineering. *IEEE Transactions on Software Engineering*, 47(9), 1960-1979.
- [18] Sherman, G. (2019). Document expansion and language model re-estimation for information retrieval (Doctoral dissertation, University of Illinois at Urbana-Champaign).
- [19] Kaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1), 25-29.
- [20] Tinega, G. A., Mwangi, W., & Richard, M. R. (2019). Text Mining in Digital Libraries using OKAPI BM25 Model. *International Journal of Computer Applications Technology and Research*, 7(10), 398-406.
- [21] Shao, Y., & Ye, Z. (2019). THUIR@ AILA 2019: Information Retrieval Approaches for Identifying Relevant Precedents and Statutes. *FIRE (Working Notes)*, 2517, 46-51.
- [22] Gupta, S., Yadav, S., & Prasad, R. (2018). Document Retrieval using Efficient Indexing Techniques: A Review. *Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications*, 1745-1764.
- [23] Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29, 21-43.
- [24] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261-266.
- [25] Farkiya, A., Saini, P., Sinha, S., & Desai, S. (2015). Natural language processing using NLTK and wordNet. *IJCSIT Int J Comput Sci Inf Technol*, 6(6), 5465-5454.
- [26] Nasiboglu, R., & Gencer, M. (2021). Comparison of spacy and stanford libraries pre-trained deep learning models for named entity recognition. *Journal of Modern Technology and Engineering*, 6(2), 104-111.
- [27] Alqifari, R. (2019, September). Question answering systems approaches and challenges. In *Proceedings of the Student Research Workshop Associated with RANLP 2019* (pp. 69-75).
- [28] Hao, T., Li, X., He, Y., Wang, F. L., & Qu, Y. (2022). Recent progress in leveraging deep learning methods for question answering. *Neural Computing and Applications*, 1-19.
- [29] Asai, A., Yu, X., Kasai, J., & Hajishirzi, H. (2021). One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, 34, 7547-7560.
- [30] Loginova, E., Varanasi, S., & Neumann, G. (2021). Towards end-to-end multilingual question answering. *Information Systems Frontiers*, 23(1), 227-241.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] Huang, Z., Xu, S., Hu, M., Wang, X., Qiu, J., Fu, Y., ... & Wang, C. (2020). Recent trends in deep learning based open-domain textual question answering systems. *IEEE Access*, 8, 94341-94356.
- [2] Dimitrakis, E., Sgontzos, K., & Tzitzikas, Y. (2020). A survey on question answering systems over linked data and documents. *Journal of intelligent information systems*, 55(2), 233-259.
- [3] McKenna, P. (2019). Multiple choice questions: answering correctly and knowing the answer. *Interactive Technology and Smart Education*.

- [4] Mishra, A., & Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3), 345-361.
- [5] Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., & Socher, R. (2021). COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ digital medicine*, 4(1), 1-9.
- [6] Mutabazi, E., Ni, J., Tang, G., & Cao, W. (2021). A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12), 5456.
- [7] Yuan, Z., Sun, S., Duan, L., Li, C., Wu, X., & Xu, C. (2020). Adversarial multimodal network for movie story question answering. *IEEE Transactions on Multimedia*, 23, 1744-1756.
- [8] Goel, A. K., & Polepeddi, L. (2018). Jill Watson: A virtual teaching assistant for online education. In *Learning engineering for online education* (pp. 120-143). Routledge.
- [9] Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2021). Geo-analytical question-answering with GIS. *International Journal of Digital Earth*, 14(1), 1-14.
- [10] Gao, S., Chen, X., Ren, Z., Zhao, D., & Yan, R. (2021). Meaningful answer generation of e-commerce question-answering. *ACM Transactions on Information Systems (TOIS)*, 39(2), 1-26.
- [11] Guo, D., Tang, D., Duan, N., Zhou, M., & Yin, J. (2018). Dialog-to-action: Conversational question answering over a large-scale knowledge base. *Advances in Neural Information Processing Systems*, 31.
- [12] Morton, K., Wang, P., Bizon, C., Cox, S., Balhoff, J., Kebede, Y., ... & Tropsha, A. (2019). ROBOKOP: an abstraction layer and user interface for knowledge graphs to support question answering. *Bioinformatics*, 35(24), 5382-5384.
- [13] Wenyan, L., Hao, T., Chen, W., & Feng, M. (2009). A web-based platform for user-interactive question-answering. *World Wide Web*, 12(2), 107-124.
- [14] Moldovan, D., Pasca, M., Harabagiu, S., & Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS)*, 21(2), 133-154.
- [15] IBM Watson: Ramprasath, M., & Hariharan, S. (2012). A survey on question answering system. *International Journal of Research and Reviews in Information Sciences*, 2(1).
- [16] Deepak Kumar Gupta, Rajkumar Pujari, Asif Ekbal, Pushpak Bhattacharyya, Anutosh Maitra, Tom Geo Jain, & Shubhashis Sengupta (2018). Can Taxonomy Help? Improving Semantic Question Matching using Question Taxonomy. In *COLING*.
- [17] Chen, X., Chen, C., Zhang, D., & Xing, Z. (2019). Sthesaurus: Wordnet in software engineering. *IEEE Transactions on Software Engineering*, 47(9), 1960-1979.
- [18] Sherman, G. (2019). Document expansion and language model re-estimation for information retrieval (Doctoral dissertation, University of Illinois at Urbana-Champaign).
- [19] Kaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1), 25-29.
- [20] Tinega, G. A., Mwangi, W., & Richard, M. R. (2019). Text Mining in Digital Libraries using OKAPI BM25 Model. *International Journal of Computer Applications Technology and Research*, 7(10), 398-406.
- [21] Shao, Y., & Ye, Z. (2019). THUIR@ AILA 2019: Information Retrieval Approaches for Identifying Relevant Precedents and Statutes. *FIRE (Working Notes)*, 2517, 46-51.
- [22] Gupta, S., Yadav, S., & Prasad, R. (2018). Document Retrieval using Efficient Indexing Techniques: A Review. *Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications*, 1745-1764.
- [23] Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29, 21-43.
- [24] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261-266.
- [25] Farkiya, A., Saini, P., Sinha, S., & Desai, S. (2015). Natural language processing using NLTK and wordNet. *IJCSIT) Int J Comput Sci Inf Technol*, 6(6), 5465-5454.
- [26] Nasiboglu, R., & Gencer, M. (2021). Comparison of spacy and stanford libraries pre-trained deep learning models for named entity recognition. *Journal of Modern Technology and Engineering*, 6(2), 104-111.
- [27] Alqifari, R. (2019, September). Question answering systems approaches and challenges. In *Proceedings of the Student Research Workshop Associated with RANLP 2019* (pp. 69-75).
- [28] Hao, T., Li, X., He, Y., Wang, F. L., & Qu, Y. (2022). Recent progress in leveraging deep learning methods for question answering. *Neural Computing and Applications*, 1-19.
- [29] Asai, A., Yu, X., Kasai, J., & Hajishirzi, H. (2021). One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, 34, 7547-7560.
- [30] Loginova, E., Varanasi, S., & Neumann, G. (2021). Towards end-to-end multilingual question answering. *Information Systems Frontiers*, 23(1), 227-241.