УДК 004.8
DOI https://doi.org/10.32851/tnv-tech.2023.1.2

ЗАГАЛЬНОДОСТУПНІ НАБОРИ ДАНИХ ТА МЕТРИКИ ДЛЯ СПРИЯННЯ ДОСЛІДЖЕННЯМ СИСТЕМ НАДАННЯ ВІДПОВІДЕЙ

Вишняк М. Ю. *– кандидат технічних наук, доцент,*
професор кафедри системотехніки
Харківського національного університету радіоелектроніки
ORCID ID: 0000-0002-3240-139X

Пироженко М. Ю. *– аспірант*
Харківського національного університету радіоелектроніки
ORCID ID: 0000-0003-0785-1273

Системи надання відповідей – це інформаційні системи, призначені для відповідей на запитання, поставлених природною мовою. Останніми роками збільшилася увага дослідників до розробки систем надання відповідей та, разом з цим, збільшилася кількість загальнодоступних наборів тестових даних, які публікуються для сприяння дослідженням у цій галузі обробки текстів природних мов. Дослідження відкритих наборів даних є важливим, оскільки вони дозволяють розробляти кращі системи, які можуть точно відповідати на широкий спектр питань. У цій статті розглядаються загальнодоступні, великі, оригінальні та активно використовувані набори навчальних даних, які застосовуються при дослідженнях систем надання відповідей, а також надані метрики, що застосовуються для порівняння моделей цих систем. Дослідження проводилось з використанням системного підходу, методів абстрагування, системного аналізу, порівняння та синтезу. У результаті роботи було вирішено актуальне наукове завдання, що стосується розвитку методологічної бази розробки інформаційних систем, які здатні відповідати на питання користувача на базі інформації, представленої неструктурованими текстовими колекціями даних, оглянуті наявні загальнодоступні набори даних та метрики оцінки моделей систем надання відповідей, враховуючи останні публікації в цій галузі. Практична значущість такого дослідження полягає в можливості застосування його результатів 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 пошукових систем та систем надання відповідей.

Ключові слова: *системи надання відповідей, текстова аналітика, обробка природної мови, навчальні набори даних, метрики оцінки.*

Vyshniak M. Yu., Pyrozhenko M. Yu. Publicly available datasets and metrics to advance research on question-answering systems

Question-Answering systems are information systems designed to answer questions in natural languages. In recent years, the attention of researchers to the development of QA systems has increased and with it, the number of publicly available test datasets researches to facilitate research in this area of natural language text processing. Research on open datasets is important because it enables the development of better systems that can accurately answer a wide range of questions. In this paper, we have reviewed publicly available, large, original, and widely used training datasets that used in research on QA systems, and provided metrics that used to compare models of these systems. The research was conducted using a systemic approach, methods of abstraction, systemic analysis, comparison and synthesis. As a result of the work, an actual scientific task was solved, which consists in determining the current state of development of the methodological base for the development of information systems capable of answering user questions on the basis of information represented by unstructured textual data collections, reviewed existing publicly available datasets and evaluation metrics of QA system models, taking into account recent publications in this field. The practical significance of the research lies in the possibility of applying scientific provisions and conclusions for the development and imple-

mentation of response systems; in the process of teaching natural language processing disciplines in higher educational institutions; when writing manuals for natural language processing; during applied research of search engines and question-answering systems.

Key words: *question-answering systems, text analytics, natural language processing, training datasets, evaluation metrics.*

Постановка задачі. Системи надання відповідей – це інформаційні системи, призначені для формування відповідей на запитання, поставлених природною мовою. Ці системи зазвичай складаються з кількох компонентів, таких як підсистеми аналізу питання, пошуку документів, ідентифікації сутностей та формування відповідей. У той час, як традиційні інформаційно-пошукові системи використовуються для пошуку посилань на документи, які відповідають ключовим словам запиту, сучасні системи надання відповідей зосереджені на представлені кінцевих відповідей, на чітко сформульовані людиною запитання. Таким чином, кінцевим користувачам залишається лише ознайомитись зі сформованою відповіддю.

Останніми роками системам надання відповідей приділяється багато уваги з боку дослідників [1; 2]. Розробка якісних систем надання відповідей передбачає використання великих навчальних наборів даних, складність побудови яких обмежує розвиток галузі. Дослідження відкритих наборів даних є важливим, оскільки вони дозволяють розробляти кращі системи, які можуть точно відповідати на широкий спектр питань. Таке дослідження полягає в пошуку та створені ефективних засобів оцінки різних типів систем надання відповідей.

Тому, перш за все, потрібно оглянути поточний стан розробки загальнодоступних наборів даних та оглянути метрики, що дозволяють порівняти моделі систем надання відповідей.

Аналіз останніх досліджень і публікацій. Останніми роками дослідження систем надання відповідей були зосереджені на покращенні точності відповідей цих систем за рахунок підвищення їх здатності розуміти природну мову та опрацьовувати складні питання. Це потребувало створення великих навчальних наборів даних, які краще відображають реальні сценарії та виклики.

У праці [3] запропонована класифікація наборів даних, відповідно до типу завдань, на абстрагування, вилучення та пошук. Автори встановили, що метрики, які використовуються в оцінюванні, чітко розмежовуються залежно від типу завдання. Так було визначено, що для всіх абстрактних завдань, особливість яких полягає в тому, що відповідь формуються природною мовою та у вільній формі, використовуються версії метрик ROUGE, BLEU та METEOR. Для завдань на основі вилучення, що потребують визначити частини документа, які містять відповідь на запитання, покладаються на F1 та EM метрики. В задачах на основі пошуку, які передбачають лише ранжування певної кількості коротких текстових сегментів, використовуються такі метрики, як MAP та MRR.

В [4] запропонована класифікація наборів даних, де останні поділяються за стилем анотації на чотири категорії: закритий (такий, що передбачає доповнення речення питання словом або фразою), з вибором поміж кількома варіантами (такий, що передбачає вибір правильної відповіді серед кандидатів, які мають вводити в оману), з вилученням діапазону (такий, що передбачає пошук діапазону слів, які є відповіддю на запитання) та у довільній формі (такі, що передбачають генерувати будь-яку форму тексту у якості відповіді). Автори узагальнили тенденції у галузі та висловили власні погляди щодо майбутніх напрямків дослідження тестів до текстових систем надання відповідей.

В [5] представлено комплексне дослідження, що включає також набори даних до систем надання відповідей пов'язаних з мультимедійними формами джерел, насамперед зображеннями та відео. Також представлені статистичні показники та приклади використання.

Відокремлення невирішених частин загальної проблеми. Системи надання відповідей вимагають великої кількості анотованих даних. Набори даних не завжди доступні, особливо для спеціалізованих областей знань або мов, що негативно впливає на продуктивність таких системи. На сьогодні фактично відсутній огляд сучасних, великих, оригінальних та активно використовуваних наборів даних, який враховував би досягнення, які відбулись у цій галузі за останні роки.

Мета дослідження. Визначення сучасного стану розвитку загальнодоступних наборів даних та метрик оцінки інформаційних систем, які здатні відповідати на питання на базі інформації, представленої неструктурованими текстовими колекціями даних, враховуючи останні досягнення в цій галузі.

Методи, предмет та об'єкт дослідження. Дослідження проводилось з використанням системного підходу, методів абстрагування, системного аналізу, порівняння та синтезу. Предмет дослідження – поточний стан розробки загальнодоступних наборів даних та метрик для систем надання відповідей. Об'єкт дослідження – система надання відповідей.

Виклад основного матеріалу. Відповідно до мети розглянуто сучасні, великі, загальнодоступні набори даних. Цей огляд не включає набори: які передбачають лише підтвердження або спростування твердження, обмежені вузькою областю знань та непридатні до масового використання, з малим обсягом питань, представлені без контексту, створені на основі баз знань. Також ми не враховували набори, що були пов'язані з відмінними від англійської або української мовами.

Метрики оцінки відіграють важливу роль у аналізі систем надання відповідей, забезпечуючи спосіб кількісної оцінки та порівняння продуктивності систем, визначенні областей для покращення та оптимізації моделей, тому попередньо розглянемо їх.

Accuracy являє собою відсоток запитань, на які система дала правильну відповідь. Цей показник розраховується за наступною формулою:

$$Accuracy = \frac{M}{N} \quad (1)$$

де M є кількість запитань, на які надані правильні відповіді; N позначає загальну кількість запитань у наборі даних.

Exact Match вимірює відсоток передбачень. Передбачення вважається правильним лише тоді, коли воно точно збігається з будь-якою з еталонних відповідей на задане запитання. Exact Match розраховується за наступною формулою:

$$EM = \frac{M}{N} \quad (2)$$

де M позначає кількість правильних прогнозів; N позначає загальну кількість запитань у наборі даних.

F1 є середнім гармонійним значенням точності (precision) та повноти (recall). У системах надання відповіді, точність вимірює відношення кількості лексем у прогнозі, які збігаються з правильною відповіддю, до загальної кількості лексем у прогнозі. Під повнотою розуміється відношення кількості лексем у правильній відповіді, які були охоплені прогнозом, до загальної кількості лексем у правильній відповіді.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

де Precision є відношенням кількості лексем у прогнозі, які збігаються з правильною відповіддю, до загальної кількості лексем у прогнозі; Recall є відношення кількості лексем у правильній відповіді, які були охоплені прогнозом, до загальної кількості лексем у правильній відповіді.

BLEU [6] порівнює лексичні особливості двох послідовностей зі словами та широко застосовується для генеративного текстового оцінювання якості.

ROUGE[7] працює шляхом порівняння автоматично створеного тексту з довідковим текстом, створеним людиною. Найпопулярнішими показниками ROUGE у текстовій перевірці якості є ROUGE-N і ROUGE-L, які представляють порівняння текстів із різною деталізацією. ROUGE-N вимірює відношення кількості n-грам, що перекриваються між згенерованим та еталонним текстом, до загальної кількості n-грамів у еталонному тексті. Подібним чином ROUGE-L вимірює найдовшу відповідну послідовність слів, використовуючи найдовшу загальну підпослідовність (LCS), яка може автоматично включати найдовші n-грами в послідовності.

Meteor[8] – ще одна метрика, що використовується в системах надання відповідей. Попередньо була запропонована для машинного перекладу. Стверджується, що має кращу кореляцію з людським судженням.

HEQ був запропонований разом із набором даних QuAC[9]. HEQ має два варіанти: HEQ-Q (відсоток питань, для яких відповідь вірна) і HEQ-D (відсоток діалогів, для яких відповідь вірна для кожного запитання в діалогу). Розробники набору QuAC стверджують, що оцінка F1 може вводити в оману для питань із кількома відповідями.

RACE використовується як метрика оцінки іспитів середньої школи КНР, але ці набори питань також використовуються в дослідженнях систем надання відповідей[10]. RACE-M та RACE-H, кожна з яких позначають різні ступені складності питань іспиту.

В процесі дослідження оглянуто близько 30 наборів даних, що широко використовуються дослідниками та розробниками систем надання відповідей. Вони надають колекції різноманітних запитань, які можна використовувати для навчання, оцінки та порівняння різних систем. Далі будуть оглянуть особливості цих наборів.

ARC [11] – це набір даних, що містить наукові запитання із варіантами відповідей. Набір даних включає окремий набір складних завдань та загальний набір. Набір складних завдань містить лише ті запитання, на які пошукові алгоритми дали неправильну відповідь.

Набір даних AdversarialQA [12] створено за допомогою змагання, де задаються запитання до трьох різних моделей. Набір складається з питань, на які вони не змогли правильно відповісти. Це гарантує, що набір складається з питань, які сучасні моделі вважають складними. Питання поділено на групи, що містять навчальні приклади, приклади для перевірок та тестів.

Набір даних Children's Book Test (CBT) [13] призначений для безпосереднього вимірювання того, наскільки добре мовні моделі можуть використовувати широкий лінгвістичний контекст для встановлення відповідей. Набір даних складається з книг, що є у вільному доступі. CBT містить чотири різні конфігурації, в залежності від типу відповіді: де відповідями на питання є дієслова, відповідями

на питання є займенники, відповідями на питання є назви сутностей, відповідями на питання є загальні іменники.

Набір даних CNN / Daily Mail [14] було згенеровано з унікальних статей новин, написаних журналістами, та опублікованих на веб-сайтах CNN та Daily Mail. Набір даних складається з запитань (з прихованою однією з сутностей), а також історій, з яких система повинна отримати відповідь.

ComQA [15] – це набір даних, які були зібрані з веб-сайту для відповідей на запитання спільноти. Таким чином вони стосуються інформаційних потреб справжніх користувачів, на які вони попередньо не змогли знайти правильну відповідь за допомогою пошукових систем. Набір даних містить запитання з різними складними явищами.

Набір даних CoQA [16] – це великомасштабний набір даних для побудови систем відповідей на розмовні запитання. Мета тестів CoQA – виміряти здатність машин розуміти окремі уривки тексту та відповідати на низку взаємопов'язаних запитань, які з'являються під час розмови.

Набір даних Cosmos QA [17] – це великомасштабний набір із завдань, які сформульовані у вигляді запитань із варіантами відповідей. Цей набір складається з колекції повсякденних розповідей людей та запитань щодо ймовірних причин або наслідків подій, які вимагають міркування поза межами контексту.

Набір даних DROP [18] – це тест із запитань, у якому система повинна опрацьовувати посилання в питанні, можливо, на кілька вхідних позицій, та виконувати над ними окремі операції (такі як додавання, підрахунок або сортування). Ці операції вимагають набагато більш повного розуміння змісту уривків, ніж те, що було необхідним для більшості наборів даних.

Набір даних DuoRC [19] розроблено спеціально для того, щоб містити велику кількість питань із низьким лексичним перекриттям між запитаннями та відповідними уривками Це вимагає, щоб моделі виходили за межі змісту уривка та лише так могли прийти до відповіді. У наборі використовуються розповіді по сюжетам фільмів, які описують складні події. DuoRC потребує складних міркувань у кількох реченнях, щоб отримати відповідь на запитання. Також перевіряє, щоб модель виявила відсутність відповіді на запитання.

ELI5 [20] – це набір даних для розгорнутих відповідей на запитання. Набір даних містить складні, різноманітні запитання, які вимагають пояснювальних відповідей з кількох речень. Результати веб-пошуку використовуються у якості джерела для формування відповідей на кожне запитання.

HotpotQA [21] – являє собою набір даних для відповідей на питання, що включає природні, багатоступеневі питання, з контролем підтверджених фактів. Питання вимагають доступу до декількох документів для отримання відповіді, а також підтверджень певних фактів для отримання правильної відповіді.

Набір даних MS MARCO [22] орієнтовано на глибоке навчання під час пошуку. Перша версія набору даних складається з відповідей на запитання, що містили понад 100 тисяч реальних запитань, поставлених до пошукової системи Bing, та відповідей, створених людиною. Згодом колекція була розширена набором даних із понад мільйона запитань.

MultiRC [23] – це набір даних з коротких абзаців та запитань, на які відповісти можна зі змісту абзацу. Кількість правильних варіантів відповідей на кожне запитання заздалегідь не вказується. Правильна відповідь не обов'язково має бути проміжком у тексті. Уривки тексту в наборі мають різне походження.

Набір даних NarrativeQA [24] містить книги та сценарії фільмів, зібрані з різних сайтів, а також коротких описів сюжетів, що були отримані з Вікіпедії. В результаті роботи було отримано 1567 пар оповідань та резюме, які були перевірені редакторами. Питання складені з урахуванням представлених коротких описів сюжетів. Питання складені з урахуванням представлених коротких описів сюжетів, причому таким чином, щоб на них могли відповісти люди, які прочитали повні версії оповідань. Відповіді складені на основі змісту анотацій.

Набір даних NaturalQuestions [25] містить реальні запитання користувачів, які ставляться до пошукової системи Google, та відповіді на них, знайдені людьми у Вікіпедії. Таким чином, Natural Questions призначено для оцінки систем у реальному середовищі.

Набір даних NewsQA [26] був випущений з метою створення більш природного та складного набору даних, ніж існуючи на той час набори. У наборі відсутні варіантів відповідей, з яких можна обрати. Відповіддю у NewsQA є проміжки довільної довжини. На деякі питання немає відповіді у відповідній статті.

Набір даних NLGEN [22] містить сформовані людьми відповіді після проведення спеціального аналізу попередньо згенерованих відповідей. Відповідь створювалася людиною лише для питання, на яке згенерована відповідь мала проблеми з граматикою, була створена шляхом копіювання тексту з одного зі знайдених уривків, або розуміння поточної відповіді потребувало доступу до контексту питання та уривка.

Набір даних PR [22] – набір, що був випущений на основі анонімних запитів, зроблених у пошуковій системі. Для набору були відібрані запитання, що містять не менше восьми слів та подані кількома користувачами за короткий проміжок часу. Відповіді сформовані з повних сторінок Вікіпедії, замість окремих уривків.

Набір даних QnA [22] був випущений компанією Microsoft. Набір містить вибірку запитів, отриману з Bing та Cortana. Запити, що не містять питань видалені з вибірки на етапі пост-обробки за допомогою класифікатора на основі машинного навчання. У середньому, кожне питання пов'язане з десятьма уривками, отриманими з веб-сторінок, знайдених за допомогою Bing. Кожне питання пов'язане з нулем, однією або кількома відповідями, які були створені після вивчення змісту уривків, знайдених для цього питання, та обмежені інформацією, доступною у знайдених уривках.

QuAC [9] – це набір даних для моделювання, розуміння та участі в діалозі з метою пошуку інформації. Набір даних складено у діалозі пов’язаному з прихованим текстом у Вікіпедії. Запитання QuAC часто не мають відповіді або мають значення лише в контексті діалогу.

RACE [10] – це набір даних (з такою ж назвою, як і метрика), зібраний на основі тестів з англійської мови, призначений для учнів середньої та старшої китайської школи. Моделі оцінюються на основі точності іспитів середньої школи та загального набору даних.

ReClor [27] – це набір даних, отриманий із стандартизованих іспитів GMAT та LSAT, що потребує логічних міркувань. Цей набір даних перевіряє різні типи логічних міркувань: необхідність та достатність припущень, посилення та послаблення важливості, синтез інформації, оцінка наслідків, висновок та іншими.

Набір даних SearchQA [28] було створено на основі архіву J!Archive, де зберігаються запитання та відповіді з телевізійного шоу Jeopardy. Кожна пара була пов'язана з набором уривків, отриманих під час пошуку відповідей у Google. Отримані уривки пройшли певну фільтрацію таким чином, що відповіді не можуть бути

знайдені просто збігом слів питання. Пара питання-відповідь видалялася, якщо відповідь складалася більш ніж з трьох слів або у пошукових уривках не містилося відповіді.

SQuAD [29] – це набір даних, що складається із запитань, поставлених до набору статей з Вікіпедії. Перша версія набору SQuAD, містила близько 100 тисяч пар запитань-відповідей до 500 статей. Друга об'єднує запитання у першому наборі із понад 50 тисяч запитань, на які немає відповіді. Питання без відповіді написані таким чином, щоб виглядали схожими на ті, на які можна відповісти. Цей набір було перекладено українською.

Story Cloze Test [30] – це набір даних, який призначений для систем надання відповідей задля оцінки можливостей аналізу. Набір даних надає історії з чотирьох речень та двома можливими закінченнями. Системи повинні вибрати правильний кінець історії.

Набір даних TriviaQA [31] включає питання та відповіді зібрані з 14 різних сайтів, присвячених вікторинам, з яких видалено питання, що містять менше чотирьох лексем. Питання в наборі в основному складні та композиційні, з великою синтаксичною та лексичною варіативністю між питаннями. Відповідь на питання вимагає глибшого осмислення перехресних пропозицій порівняно з іншими наборами даних.

Набір даних TWEETQA [32] є першим великомасштабний набором, який фокусується на текстах соціальних мереж. Оскільки соціальні медіа стають дедалі популярнішими, розробка систем надання відповідей має важливе значення для тих систем, які покладаються на знання в реальному часі. Саме завдання вимагає від системи проаналізувати запитання, короткий твіт та вивести текстову фразу як відповідь.

Результати досліджень. Проведено огляд загальнодоступних, великих, оригінальних та активно використовуваних наборів даних, які використовуються в дослідженнях, спрямованих на покращення якості систем надання відповідей.

Ці та інші набори допомагають покращити продуктивність систем надання відповідей, зробивши їх більш кориснішими та доступнішими для ширшого кола застосувань. Основні характеристики наборів представлено у таблиці 1.

Обговорення результатів. Набори даних запитань-відповідей є важливими для розробки систем надання відповідей, оскільки вони відіграють вирішальну роль у визначенні точності відповідей та якості системи загалом.

Створення власного набору даних вимагає значних зусиль та часу. Добре розроблений набір даних має містити різноманітний діапазон питань з широкого спектру тем відповідних областей знань. Система надання відповідей повинна пристосовуватися до реальних питань, з якими стикаються під час фактичного використання. Важливо щоб набір був репрезентативним. Це має важливе значення, насамперед для того, щоб моделі систем могли узагальнювати данні. Це може також бути складним завданням, оскільки вимагає глибокого розуміння тем та областей знань, які охоплює набір даних.

Проблема, з якою стикаються розробники при створенні системи надання відповідей, полягає в забезпеченні відповідності систем вимогам та забезпеченні корисності для користувача. Це вимагає глибокого розуміння запитань, які ставляться, та інформації, яку шукають. Ретельний аналіз запитань та відповідей дозволяє виявити шляхи покращення таких систем. Використання вільних наборів даних пропонують численні переваги. Надаючи точні та актуальні навчальні дані, ці набори допомагають створювати системи наближені до реальних потреб.

Таблиці 1

Характеристика наборів даних систем надання відповідей

Колекція	Джерело документів	Кількість питань	Метрики
ARC	Наука	7787	Accuracy
AdversarialQA	Вікіпедія	36000	EM, F1
Children Book Test	Книги	687343	Accuracy
CNN/DailyMail	Новини	311672	Accuracy
ComQA	Форум	11214	Accuracy, F1
CoQA	Вікіпедія	127000	F1
Cosmos QA	Блог	35600	Accuracy
DROP	Вікіпедія	96567	F1
DuoRC	Кіносценарії	186089	Accuracy, F1
ELI5	Форум	270000	ROUGE-L, ROUGE-1, ROUGE-2
HotpotQA	Вікіпедія	112779	EM, F1
MS MARCO	Інтернет	1010916	ROUGE-L, BLEU-1
MultiRC	Інтернет	9700	F1, EM
NarrativeQA	Вікіпедія	46765	ROUGE-L, BLEU-1, BLEU-4, METEOR
Natural Questions	Вікіпедія	323045	F1
NewsQA	Новини	119633	EM, F1
NLGEN	Інтернет	182669	ROUGE-L, BLEU-1
PR	Вікіпедія	1010916	MRR@10
QnA	Інтернет	1026758	ROUGE-L, BLEU-1
QuAC	Вікіпедія	98407	HEQ, F1
RACE	Екзамен	97867	RACE, RACE-H, RACE-M
ReClor	Екзамен	6138	Accuracy
SearchQA	Інтернет	140461	Accuracy, F1
SQuAD	Вікіпедія	151054	EM, F1
StoryCloze Test	Книги	101901	Accuracy
TriviaQA Web	Інтернет	95956	EM, F1
TWEETQA	Твіти	13757	ROUGE-L, BLEU-1, METEOR

Оскільки системи надання відповідей продовжують розвиватися та стають все більш складними, а також враховуючи складність розробки власних наборів, можна стверджувати, що важливість загальнодоступних наборів даних продовжуватиме зростати та відіграватиме дедалі важливішу роль у дослідженнях систем надання відповідей.

Висновки. У цьому дослідженні були розглянуті великі, оригінальні та активно використовувані набори даних, що застосовуються в дослідженнях систем надання відповідей. Також розглянуто метрики, які використовують для оцінки якості моделей систем надання відповідей. У результаті цієї роботи представлено 27 наборів даних, включно з найновітнішими у цій галузі, що були створені для розробки текстових систем надання відповідей за останні роки, та надані їх характеристики.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ:

1. Recent trends in deep learning based open-domain textual question answering systems / Z. Huang et al. IEEE access. 2020. Vol. 8. P. 94341–94356. URL: https://doi.org/10.1109/access.2020.2988903 (date of access: 03.03.2023).

2. Dimitrakis E., Sgontzos K., Tzitzikas Y. A survey on question answering systems over linked data and documents. Journal of intelligent information systems. 2019. Vol. 55, no. 2. P. 233–259. URL: https://doi.org/10.1007/s10844-019-00584-7 (date of access: 03.03.2023).

3. A review of public datasets in question answering research / B. B. Cambazoglu et al. ACM SIGIR Forum. 2020. Vol. 54, no. 2. P. 1–23. URL: https://doi.org/10.1145/3483382.3483389 (date of access: 03.03.2023).

4. Wang Y. B. More Than Reading Comprehension: A Survey on Datasets and Metrics of Textual Question Answering. Computation and Language. 2021.

5. Wang Z. Modern Question Answering Datasets and Benchmarks: A Survey. CoRR. 2022.

6. Bleu / K. Papineni et al. The 40th annual meeting, Philadelphia, Pennsylvania, 7–12 July 2002. Morristown, NJ, USA, 2001. URL: https://doi.org/10.3115/1073083.1073135 (date of access: 03.03.2023).

7. Lin C. Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. 2004.

8. Banerjee S. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments / S. Banerjee, A. Lavie. // workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005. C. 65–72.

9. QuAC: question answering in context / E. Choi et al. Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, Belgium. Stroudsburg, PA, USA, 2018. URL: https://doi.org/10.18653/v1/d18-1241 (date of access: 03.03.2023).

10. RACE: large-scale reading comprehension dataset from examinations / G. Lai et al. Proceedings of the 2017 conference on empirical methods in natural language processing, Copenhagen, Denmark. Stroudsburg, PA, USA, 2017. URL: https://doi.org/10.18653/v1/d17-1082 (date of access: 03.03.2023).

11. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge / P. Clark et al. ArXiv, 3 March 2018.

12. Beat the AI: investigating adversarial human annotation for reading comprehension / M. Bartolo et al. Transactions of the association for computational linguistics. 2020. Vol. 8. P. 662–678. URL: https://doi.org/10.1162/tacl_a_00338 (date of access: 03.03.2023).

13. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. / F. Hill et al. ICLR.

14. Teaching Machines to Read and Comprehend Tibetan Text / Y. Sun et al. Journal of Computer and Communications. 2021. Vol. 09, no. 09. P. 143–152. URL: https://doi.org/10.4236/jcc.2021.99011 (date of access: 03.03.2023).

15. ComQA: question answering over knowledge base via semantic matching / H. Jin et al. IEEE access. 2019. Vol. 7. P. 75235–75246. URL: https://doi.org/10.1109/access.2019.2918675 (date of access: 03.03.2023).

16. Reddy S., Chen D., Manning C. D. CoQA: a conversational question answering challenge. Transactions of the association for computational linguistics. 2019. Vol. 7. P. 249–266. URL: https://doi.org/10.1162/tacl_a_00266 (date of access: 03.03.2023).

17. Cosmos QA: machine reading comprehension with contextual commonsense reasoning / L. Huang et al. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), Hong Kong, China. Stroudsburg,

PA, USA, 2019. URL: https://doi.org/10.18653/v1/d19-1243 (date of access: 03.03.2023).

18. Dua et al. Proceedings of the 2019 conference of the north, Minneapolis, Minnesota. Stroudsburg, PA, USA, 2019. URL: https://doi.org/10.18653/v1/n19-1246 (date of access: 03.03.2023).

19. DuoRC: towards complex language understanding with paraphrased reading comprehension / A. Saha et al. Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), Melbourne, Australia. Stroudsburg, PA, USA, 2018. URL: https://doi.org/10.18653/v1/p18-1156 (date of access: 03.03.2023).

20. ELI5: long form question answering / A. Fan et al. Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy. Stroudsburg, PA, USA, 2019. URL: https://doi.org/10.18653/v1/p19-1346 (date of access: 03.03.2023).

21. HotpotQA: a dataset for diverse, explainable multi-hop question answering / Z. Yang et al. Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, Belgium. Stroudsburg, PA, USA, 2018. URL: https://doi.org/10.18653/v1/d18-1259 (date of access: 03.03.2023).

22. MS MARCO: benchmarking ranking models in the large-data regime / N. Craswell et al. SIGIR '21: the 44th international ACM SIGIR conference on research and development in information retrieval, Virtual Event Canada. New York, NY, USA, 2021. URL: https://doi.org/10.1145/3404835.3462804 (date of access: 03.03.2023).

23. Looking beyond the surface: a challenge set for reading comprehension over multiple sentences / D. Khashabi et al. Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long papers), New Orleans, Louisiana. Stroudsburg, PA, USA, 2018. URL: https://doi.org/10.18653/v1/n18-1023 (date of access: 03.03.2023).

24. The narrativeqa reading comprehension challenge / T. Kočiský et al. Transactions of the association for computational linguistics. 2018. Vol. 6. P. 317–328. URL: https://doi.org/10.1162/tacl_a_00023 (date of access: 03.03.2023).

25. Natural questions: a benchmark for question answering research / T. Kwiatkowski et al. Transactions of the association for computational linguistics. 2019. Vol. 7. P. 453–466. URL: https://doi.org/10.1162/tacl_a_00276 (date of access: 03.03.2023).

26. NewsQA: a machine comprehension dataset / A. Trischler et al. Proceedings of the 2nd workshop on representation learning for NLP, Vancouver, Canada. Stroudsburg, PA, USA, 2017. URL: https://doi.org/10.18653/v1/w17-2623 (date of access: 03.03.2023).

27. Jiang Y. W., Dong Z., Feng J. ReClor: a reading comprehension dataset requiring. In international conference on learning representations.

28. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine / M. Dunn et al. CoRR. 2017.

29. SQuAD: 100,000+ questions for machine comprehension of text / P. Rajpurkar et al. Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas. Stroudsburg, PA, USA, 2016. URL: https://doi.org/10.18653/v1/d16-1264 (date of access: 03.03.2023).

30. LSDSem 2017 shared task: the story cloze test / N. Mostafazadeh et al. Proceedings of the 2nd workshop on linking models of lexical, sentential and discourse-level semantics, Valencia, Spain. Stroudsburg, PA, USA, 2017. URL: https://doi.org/10.18653/v1/w17-0906 (date of access: 03.03.2023).

31. TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension / M. Joshi et al. Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), Vancouver, Canada. Stroudsburg, PA, USA, 2017. URL: https://doi.org/10.18653/v1/p17-1147 (date of access: 03.03.2023).

32. TWEETQA: a social media focused question answering dataset / W. Xiong et al. Proceedings of the 57th annual meeting of the association for computational linguistics,

Florence, Italy. Stroudsburg, PA, USA, 2019. URL: https://doi.org/10.18653/v1/p19-1496 (date of access: 03.03.2023).

REFERENCES:

1. Huang, Z., Xu, S., Hu, M., Wang, X., Qiu, J., Fu, Y., … & Wang, C. (2020). Recent trends in deep learning based opendomain 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. Cambazoglu, B. B., Sanderson, M., Scholer, F., & Croft, B. (2021). A Review of Public Datasets in Question Answering Research. SIGIR Forum, 54(2).

4. Wang, Y. B. A. D. (2021). More Than Reading Comprehension: A Survey on Datasets and Metrics of Textual Question Answering. Computation and Language.

5. Wang, Z. (2022). Modern Question Answering Datasets and Benchmarks: A Survey. CoRR, doi:10.48550/arXiv.2206.15030.

6. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

7. Lin, C.-Y. (2004, July). ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out, 74–81.

8. Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72).

9. Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-T., Choi, Y., … Zettlemoyer, L. QuAC: Question Answering in Context. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2174–2184. (pp. 311–318).

10. Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017, September). RACE: Large-scale ReAding Comprehension Dataset from Examinations. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 785–794.

11. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. ArXiv, abs/1803.05457.

12. Bartolo, M., Roberts, A., Welbl, J., Riedel, S., & Stenetorp, P. (2020). Beat the ai: Investigating adversarial human annotation for reading comprehension. Transactions of the Association for Computational Linguistics, 8, 662–678.

13. Hill, F., Bordes, A., Chopra, S., & Weston, J. (2016). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In Y. Bengio & Y. LeCun (Eds.), ICLR.

14. Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. Advances in Neural Information Processing Systems, 28.

15. Abujabal, A., Roy, R. S., Yahya, M., & Weikum, G. (2019). ComQA: A Community-sourced Dataset for Complex Factoid Question Answering with Paraphrase Clusters. In Proceedings of NAACL-HLT (pp. 307–317).

16. Reddy, S., Chen, D., & Manning, C. D. (2019). CoQA: A Conversational Question Answering Challenge. Transactions of the Association for Computational Linguistics, 7, 249–266.

17. Huang, L., Le Bras, R., Bhagavatula, C., & Choi, Y. (2019). Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2391–2401.

18. Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., & Gardner, M. (2019, June). DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning

Over Paragraphs. Proceedings of the 2019 Conference of the North AMerican Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 2368–2378.

19. Saha, A., Aralikatte, R., Khapra, M. M., & Sankaranarayanan, K. (2018, July). Duorc: Towards complex language understanding with paraphrased reading comprehension. In Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (ACL).

20. Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., & Auli, M. (2019, July). ELI5: Long Form Question Answering. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3558–3567.

21. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2369–2380).

22. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. choice, 2640, 660.

23. Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., & Roth, D. (2018, June). Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. Proceedings of the 2018 Conference of the North AMerican Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 252–262.

24. Kocisky, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., & Grefenstette, E. (2018). The narrativeqa reading comprehension challenge. Transactions of the Association for Computational Linguistics, 6, 317–328.

25. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., ... & Petrov, S. (2019). Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453-466.

26. Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., & Suleman, K. (2017, August). NewsQA: A Machine Comprehension Dataset. Proceedings of the 2nd Workshop on Representation Learning for NLP, 191–200.

27. Yu, W., Jiang, Z., Dong, Y., & Feng, J. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In International Conference on Learning Representations.

28. Dunn, M., Sagun, L., Higgins, M., G"uney, V. U., Cirik, V., & Cho, K. (2017). SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. CoRR, abs/1704.05179.

29. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2383–2392.

30. Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., & Allen, J. (2017). Lsdsem 2017 shared task: The story cloze test. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (pp. 46-51).

31. Joshi, M., Choi, E., Weld, D., & Zettlemoyer, L. (2017). TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 1, 1601–1611.

32. Xiong, W., Wu, J., Wang, H., Kulkarni, V., Yu, M., Chang, S., … Wang, W. Y. (2019, July). TWEETQA: A Social Media Focused Question Answering Dataset. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5020–5031.