DOI 10.34185/1562-9945-4-129-2020-06
УДК 004.738.52

Є.Р. Ковилін, О.С. Волковський

МОДЕЛЬ АВТОМАТИЧНОЇ ОЦІНКИ АДЕКВАТНОСТІ КОМП'ЮТЕРНИХ СИСТЕМ «ЗАПИТ-ВІДПОВІДЬ» З ВИКОРИСТАННЯМ ГЕНЕРАЦІЇ ТЕКСТІВ

Анотація. У статті розглянуто питання оцінки результатів роботи систем запит-відповідь з використанням IR-based архітектури, а саме системи з використанням генерації текстів, яка була розроблена на основі алгоритму побудови семантичної моделі документа. Оскільки створені алгоритми є інноваційними, розробка методів автоматичного тестування адекватності системи, побудованої на їх основі, є дуже актуальною темою для досліджень. Було сформульовано два методи досліджень результатів роботи системи - на основі бального методу, для якого були описані критерії оцінювання отриманих відповідей, і алгоритм на основі значення коефіцієнта семантичної відповідності, який дозволив організувати автоматичне оцінювання результатів роботи системи. Отримані оцінки дозволяють стверджувати про адекватність як розробленої моделі системи запит-відповідь на основі генерації текстів, так і про адекватність створеної підсистеми оцінювання IR-based систем.
Ключові слова: система запит-відповідь, IR-based архітектура, генерація текстів, критерії оцінювання.

Постановка проблеми. Перевірка адекватності роботи систем вирішення складних інтелектуальних завдань галузі АОТ, до яких належать системи запит-відповідь, є окремим науковим завданням, оскільки вона так само вимагає включення процесу розуміння семантичних властивостей тексту як і процес основної роботи системи. Саме тому, одним з найпопулярніших підходів до організації процесу оцінки систем запит-відповідь є залучення відповідних експертів, що у мануальному режимі протестують систему і дадуть своє заключення. Проте головною проблемою, яка лежить в основі такого класичного методу оцінки відповідей експертом-людиною є суб'єктивність і ненадійність судження, не кажучи вже про необхідність власне пошуку таких експертів. Окрім того, реалі-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

зація IR-based архітектури [1] систем запит-відповідь, може вимагати впровадження доступного інструменту самотестування додатку [2], що має забезпечити виконання критерію безпеки системи. Ці фактори схиляють до необхідності додаткової автоматизації процесу оцінок системи, що, по-перше, виключить людський фактор, а по-друге, дозволить гнучко сигналізувати про проблеми семантичних невідповідностей у процесі роботи системи.

Аналіз останніх досліджень і публікацій. Не дивлячись на те, що система запит-відповідь є різновидом пошукової системи, розповсюджений спосіб автоматизації тестування систем із використанням еталонних таблиць-множин запиту і відповіді, який добре зарекомендував себе при вирішенні класичних завдань пошуку колекцій відповідних до запиту документів [3], не підходить у випадку генерації текстової відповіді. Справа в тому, що результатом кожного пошукового завдання є не просто коротка відповідь, але і фрагмент тексту з конкретного документа, що явно підтверджує цю відповідь. Таких фрагментів може бути в колекції багато, і система може зробити висновок як на підставі якогось одного з них, так і кількох різних фрагментів в різних документах, що містять одну й туж саму відповідь. При цьому в множині відповідності, на відміну від класичного пошуку, міститься не тільки ідентифікатор документа, але і фрагмент тексту, і коротка відповідь з цього фрагмента. Складання подібних таблиць, де конкретна відповідь прив'язана до конкретного питання, на великому корпусі знань цілком можливо не буде являться єдино вірним рішенням. Тому, область дії автоматизації тестування повинна стосуватися всієї відповіді в цілому, а не конкретного її фрагменту на рівні одного документа і одного питання, що вимагає створення окремої підсистеми програмної оцінки згенерованих відповідей.

Мета дослідження. Розробити систему автоматичного оцінювання результатів роботи системи «запит-відповідь» з використанням автоматичної генерації текстів і перевірити за її допомогою систему, що представлена в роботі [2].

Викладення основного матеріалу дослідження. Докладний алгоритм роботи системи «запит-відповідь», на основі якої проведено оці-

нювання за створеним алгоритмом, наведено в [2]. Зазначмо, що створена система являє собою інтелектуальний пошуковий сервіс, заснований на IR-based архітектурі побудови таких систем [1]. Сенс створеного алгоритму перевірки результатів роботи IR-based систем виходить з ідеї про те, що відповіді системи, за умови роботи із адекватною повнотекстовою базою знань, не можуть бути однозначно визначені як «правильні» або «неправильні», оскільки навіть фрагмент семантично зв'язного документа містить у собі деяку кількість корисної для користувача інформації. Проте, якщо результуюча відповідь містить семантично розірвані фрагменти із різних текстів, то її сенсова значимість значно зменшується. У нашому випадку, таким критерієм семантичної зв'язності служить жанрова відповідність фрагменту тексту, що є кандидатом до включення у результуючу відповідь, та набору термінів із запиту користувача. Тому, розроблений алгоритм роботи системи тестування (рис. 1) був побудований на оцінці залежностей між тематикою множин кандидатів-документів до включення у результуючу відповідь та тематикою множин автоматично сформованих баз знань запитів, що заздалегідь були назначені кожному документу у колекції текстів. Роздивимось етапи роботи алгоритму більш детально.

Перш за все, оскільки система тестування аналізує кожен наявний документ у корпусі, найпершим кроком є отримання документу із робочої бази знань і витяг усіх пов'язаних із ним стем, оскільки саме стеми являють собою терміни-запити до системи. Для кожної отриманої стеми знаходиться її вага, після чого вибирається 8 (кількість отримана емпіричним шляхом) найважчих стем для поточного документа, що формують множину запитів для системи. Такий підхід до створення запиту зумовлений двома факторами: по-перше, він дозволяє створити множину запитів для кожного документа і, відповідно, для кожної тематичної галузі у базі знань, що дозволяє максимально повно оцінювати процес роботи системи у полістилістичній колекції документів; по-друге, умова отримання найважчих стем дозволяє вибирати максимально узагальнені поняття, що семантично пов'язані із темою тексту, що дозволить додавати у множину документів-кандидатів до включення у результуючу відпо-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

від достатню для подальшої оцінки кількість документів – наприклад, отримання терміну «космос» із тексту про телескопи також межує із філософією, терміну «клас» із інформаційних технологій межує із іншими темами (класова думка, класи зірок) та ін.

Рисунок 1 - Алгоритм роботи системи автоматичного тестування якості запитів

Для отриманих таким чином стем знаходяться їх першочергові лексичні форми – терміни, що відправляються на вхід пошукової системи, у результаті чого формується множина речень-кандидатів до включення у результуючу відповідь та множина відповідних до них документів, кожному з яких відповідає заздалегідь заданий жанр. Оскільки усі терміни до запиту відбираються із одного документа, то вважається, що усі вони мають ту саму тематичну спрямованість, що і їх батьківський документ. Для кожної множини текстів-кандидатів розраховується кількість документів, тема яких співпала із темою запиту (ситуативно коректні) та кількість документів, чия тема не співпала із темою запиту (ситуативно некоректні).

Важливим питанням цього підходу стає, власне, аналіз отриманих таким чином даних – множини кандидатів ще не є результуючою відповіддю, оскільки отримані документи повинні пройти через операцію відсікання речень із найбільшою вагою. Тому, для розуміння ступеню впливу семантичної моделі на процес генерації тексту необхідно розрахувати коефіцієнт коректності формування множини документів Q_D за формулою (1):

$$Q_D = \frac{\Sigma N_C}{\Sigma N_W}, \quad (1)$$

де $\sum N_C$ – загальна кількість ситуативно коректних документів для кожного запиту, $\sum N_W$ – загальна кількість ситуативно некоректних документів для кожного запиту.

Це значення необхідно отримувати для двох режимів функціонування системи – для «семантично відповідного» та «семантично невідповідного. Вони являють собою етап ситуативної обробки семантичної мітки документу, та складаються з безпосереднього отримання запиту і перевірці його входження у кластери-стеми бази знань системи. Головна відмінність режимів полягає у створенні множин кандидатів-кластерів до формування відповіді користувачеві – у семантично відповідному режимі кластер-стема вважається кандидатом якщо він є семантично сильним для конкретного документу, тобто має найбільшу сумарну вагу перетинів із семантичними контурами кластерів-речень, тоді як у семантично невідповідному режимі роботи системи така перевірка не відбувається - стема вважається кандидатом якщо було знайдене хоча б одне входження терміну із вхідного запиту, що є по-суті, звичайним прямим рішенням пошукової задачі.

Гіпотеза перевірки полягає у тому, що значення коефіцієнту коректності формування множини документів для семантично відповідного режиму має бути більшим, ніж для семантично невідповідного, що покаже на позитивний вплив використання створених семантичних моделей у основі систем генерації текстових відповідей. Для її підтвердження було автоматично сформовано і виконано 520 запитів до системи, у резуль-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

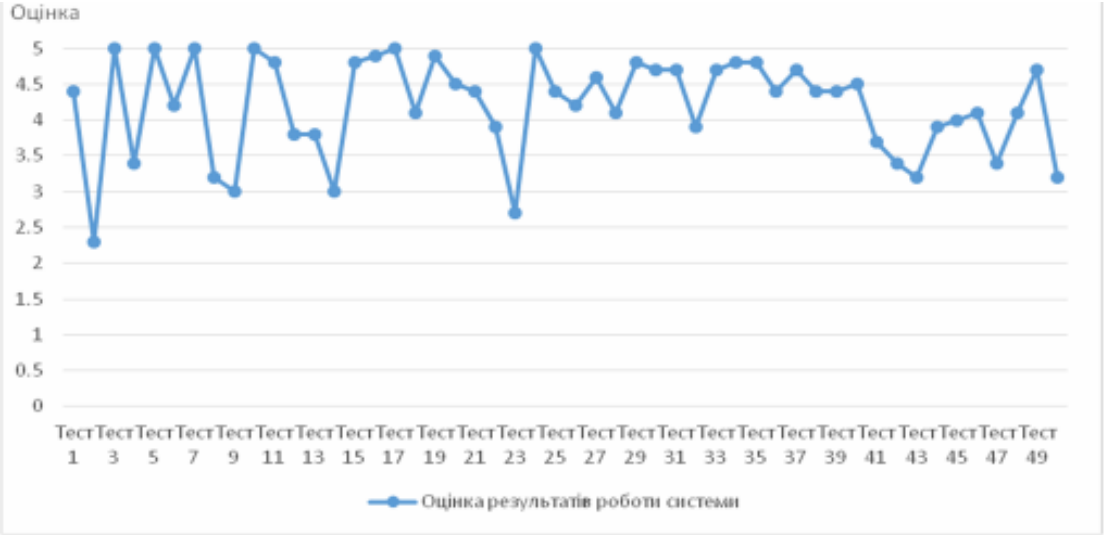
таті чого значення коефіцієнту коректності формування множини документів для семантично відповідного режиму склало 0.67 проти значення у 0.49 для семантично невідповідного режиму, що вказує на доцільність застосування підходу заснованого на семантичних моделях у системі генерації тексту відповіді на поставлене питання. Крім того, описаний алгоритм являє собою автоматичний інструмент самотестування системи, що, на відміну від підходу із множинами відповідностей, орієнтується у своїй роботі на аналіз відповіді в цілому і надає більш надійний і гнучкий процес оцінки якості системи запит-відповідь.

Для перевірки доцільності використання розробленої системи автоматичного тестування систем запит-відповідь було організоване тестування результатів людиною - кінцевим користувачем системи, що базується на індивідуальних оцінках відповідей системи за бальним методом, і являє собою сукупність оцінок вимог до згенерованої відповіді від 0 до 1 із кроком 0,1. Усього таких вимог 5, тобто кожна відповідь сумарно може получити від 0 до 5 балів, а саме: присутність відповіді – чи зміг користувач отримати необхідну інформацію, чи була вона достатньо корисною для користувача; ступень збігу із тематикою запиту – вказує на кореляцію тематики інформації, отриманої у відповіді і тематики, що цікавила користувача у запиті; повнота викладу – вказує на присутність додаткової корисної для користувача інформації, а також ступінь розкриття теми запиту; присутність тематичних розривів – пов’язана із ступенем семантичної зв’язності згенерованого тексту і вказує на неконсистентну присутність фрагментів тексту на іншу тематику; присутність сенсових розривів – пов’язана із ступенем стилістичної зв’язності згенерованого тексту і вказує на присутність абзаців або речень, що відносяться до однієї теми, проте, не зв’язані із попередніми частинами документа чим порушують його когерентність.

Окрім критеріїв оцінки якості відповіді , існує набір вимог для поставлення запиту до системи, а саме: запит має бути тематично цільним і містити терміни що семантично не суперечать одне одному, сукупно чітко описуючи певну предметну галузь. Запит має відносну семантичну силу, яка не може дорівнювати 0, та обчислюється як відсоток кількості

документів у базі знань системи на пов'язану із темою запиту тематику. Запит має містити хоча б один термін, присутній у базі знань системи.

Усього у ході дослідження було проведено 100 тестів із різною складністю питань і різним тематичним напрямом запитів, на кожен з яких була отримана відповідна оцінка, фрагмент графіку значень якої зображено на рис. 2.



Рисунке 2 - Фрагмент графіку значень експертних оцінок системи для 50 тестів

Висновки. Середнє значення усіх проведених експертних оцінок склало 0,839, що вказує на задовільні результати роботи системи запит-відповідь і підтверджує висновок, який було отримано завдяки проведення автоматичного тестування системи. Виконання мануального тестування показало наявність проблем у присутності сенсових розривів та необхідності обробки виключних ситуацій із відсутністю необхідних знань у системі, що і знизило сумарно оцінку роботи системи, проте ці факти не мають критичного впливу на алгоритм генерації відповідей. Виконання автоматичних і мануальних тестів системи запит-відповідь із використанням автоматичної генерації текстів показали сукупну адекватність роботи додатку як з точки зору використання семантичної моделі як основи для побудови інтелектуальних пошукових інструментів, так і з точки зору доцільності застосування семантичних моделей у ситуативній генерації текстів. Розроблений підхід автоматичного тестування системи

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

доцільно використовувати для перевірки роботи IR-based систем запит-відповідь із використанням автоматичної генерації текстів.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. P. Gupta. A Survey of Text Question Answering Techniques // International Journal of Computer Applications – 2012. № 53(4) - p. 1-8.

2. O.S. Volkovsky, Y. R. Kovylin. Computer system of intellectual semantic search with the text generation using// Bulletin of the Kherson National University - 2018. №3 (66). -p. 238-245.

3. А.С. Кулешов. Формирование вопросно-ответной системы в условиях ограниченного объема семантически размеченного корпуса// Программные продукты, системы и алгоритмы. № 4, 2016.

REFERENCES

1. P. Gupta. A Survey of Text Question Answering Techniques // International Journal of Computer Applications – 2012. № 53(4) - p. 1-8.

2. O.S. Volkovsky, Y. R. Kovylin. Computer system of intellectual semantic search with the text generation using// Bulletin of the Kherson National University - 2018. №3 (66). -p. 238-245.

3. A.S. Kuleshov. Formation of a question-answer system in a limited volume of a semantically labeled case // Software products, systems and algorithms. No. 4, 2016.

Received 28.02.2020.
Accepted 02.03.2020.

Модель автоматической оценки адекватности запросно-ответных компьютерных систем с использованием генерации текстов

Рассматриваются вопросы внедрения процесса автоматического тестирования запросно-ответных систем с использованием автоматической генерации текстов. С помощью разработанных методов оценена адекватность запросно-ответной системы, работающей на основе построения семантической модели документа.

Automatic assessment model of the adequacy of the request-response computer systems with using of the text generation

The article deals with the question of evaluation of the results of the work of the request-response systems with the use of IR-based architecture, namely the system with the use of text generation, which was developed on the basis of algorithm of construction of semantic model of the document. Because the algorithms created are innovative, the development of methods for automatically testing the adequacy of a system based on them is a very relevant topic for research. Despite the fact that the query-response system is a kind of search engine, a common way to automate the testing of systems using reference tables-sets of query and answer, which is well proven in solving the classic problems of searching collections of query-relevant documents, is not suitable in the case of generation text response. Therefore, two methods of researching the results of the system operation were formulated. The first is an algorithm based on the value of

the semantic correspondence coefficient, which allowed to organize the automatic evaluation of the results of the system. 520 system queries were automatically generated and executed, resulting in a value of the multiplicity of document formation correctness for the semantically appropriate mode was 0.67 versus the value of 0.49 for the semantically inappropriate mode, which indicates the feasibility of applying a semantic model-based approach to the request-response systems. The second one is based on the scoring method, for which the criteria for evaluating the answers received were described, namely: presence of the answer, degree of coincidence with the subject of the request, completeness of presentation, presence of thematic breaks and presence of meaning breaks. A total of 100 tests were conducted in the course of the study with different complexity of questions and different thematic direction of inquiries. The average of all expert assessments was 0.839, which indicates satisfactory results of the system of inquiry-response and confirms the conclusion that was obtained through the automatic testing of the system. The obtained estimates allow us to confirm the adequacy of both the developed model of the system of inquiry-response on the basis of text generation, and the adequacy of the created subsystem of evaluation of IR-based systems. It is advisable to use the developed system testing approach to test the operation of IR-based query response systems using automatic text generation.

Волковский Олег Степанович – к.т.н., доцент, Днепропетровский национальный университет имени Олеся Гончара, ДНУ, доцент кафедры автоматизированных систем обработки информации.
Ковылин Егор Романович - Днепропетровский национальный университет имени Олеся Гончара, ДНУ, аспирант кафедры автоматизированных систем обработки информации, Днепр.

Волковський Олег Степанович – к.т.н., доцент, Дніпропетровський національний університет імені Олеся Гончара, ДНУ, доцент кафедри автоматизованих систем обробки інформації.
Ковилін Єгор Романович - Дніпропетровський національний університет імені Олеся Гончара, ДНУ, аспірант кафедри автоматизованих систем обробки інформації, Дніпро.

Volkovskiy Oleg - candidate of technical science, associate professor, Oles Honchar Dnipropetrovsk National University, DNU, Senior Lecturer of department automated information processing systems, Dnepr, Ukraine.
Kovylin Egor - Oles Honchar Dnipropetrovsk National University, DNU, postgraduate student of department automated information processing systems, Dnepr, Ukraine.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)