УДК 004.021: 81’32
DOI https://doi.org/10.32838/TNU-2663-5941/2020.6-1/16

Кузьма К.Т.
Миколаївський національний університет імені В.О. Сухомлинського

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОЦІНКИ РІВНЯ ПОДІБНОСТІ РЯДКІВ НА ОСНОВІ МЕТОДУ N-ГРАМ

Обробка текстів, поданих природною мовою, – одна з головних задач комп’ютерної лінгвістики. Наряду з підвищенням результативності, якості обробки текстів важливими питаннями є просторова й обчислювальна ефективність алгоритмів обробки. Тому дослідження методів, які легко масштабуються, мають лінійну складність, у задачах обробки текстів є актуальним.

За результатами дослідження запропоновано інформаційну технологію (ІТ) для попередньої обробки текстових відповідей в автоматизованих системах тестування з метою визначення подібності відповіді й еталону. ІТ базуються на використанні синтаксичного аналізу тексту, а саме на методі N-грам. N-грамами є групи символів (від двох до п’яти), які йдуть поспіль. Для N-грам обчислюється коефіцієнт подібності кожного слова відповіді й еталону, застосовуючи індекс Жаккара. Слова вважаються подібними, якщо коефіцієнт подібності не менше заданого граничного значення (від 0,3 до 0,5). На останньому етапі визначається загальний коефіцієнт подібності для всіх слів відповіді. Повторюючи процедури циклічно для усіх груп символів, визначаємо максимальне значення коефіцієнту подібності відповіді й еталону. Якщо це значення менше за граничний коефіцієнт подібності, то не потрібно виконувати подальший етап обробки відповіді, оскільки вона не відповідає еталону. Застосування N-грам для символів слова, а не для слів, які йдуть підряд, дозволяє підвищити результативність процедури зіставлення слів.

За підсумками порівняння ІТ із методом латентно-семантичного аналізу (LSA) встановлено, що результативність методу N-грам висока для синтаксично подібних структур відповіді й еталону. Якщо структури речень відповіді й еталону значно відрізняються, то метод N-грам порівняно з LSA дає нижчий результат, оскільки він не враховує семантичні класи, на відміну від LSA.

Визначено, що подальшого дослідження потребують питання застосування «нейронних мереж» у задачах семантичного аналізу подібності текстів, об’єднання технологій синтаксичного та семантичного аналізу рядків.

Ключові слова: *обробка природної мови, N-грами, подібність рядків, індекс Жаккара, синтаксичний аналіз.*

Постановка проблеми. Сьогодні застосування N-грам є актуальним для задач комп’ютерної лінгвістики (обробка природної мови (Natural language processing), виявлення плагіату, машинне навчання), обчислювальної біології (пошук генетичних послідовностей), стиснення даних тощо. Основними перевагами методів і моделей, які базуються на використанні N-грам, є простота їх реалізації та лінійна складність. Враховуючи ці особливості N-грам, у рамках дослідження задачі автоматизованої обробки текстів, поданих природною мовою [1–3], детального вивчення потребують питання застосування N-грам для попередньої оцінки подібності відповіді й еталону.

Аналіз останніх досліджень і публікацій. Метод N-грам – один із розповсюджених підходів, який застосовується для визначення показника схожості рядків. У роботах [4–8] досліджені різні аспекти використання цього методу в задачах обробки природної мови.

У [4] як N-грами розглядаються підрядки розміром два слова, які отримують із вхідного рядка із зсувом на одне слово.

У [5] N-грамами є підрядки, розмір яких змінюється від одного до чотирьох слів залежно від довжини вхідного рядка й еталону.

Авторами робіт [6; 7] метод N-грам застосовується для оцінки подібності текстів-перекладів (машинного перекладу та перекладу, здійсненого людиною). Розмір N-грам також змінюється від 1 до 4. Частота кожної N-грами тексту машинного перекладу (еталон) обмежується частотою, з якою вона зустрічається в текстах-перекладах (відповідях), які виконали люди. Відповідно, N-грама, яка зустрічається дуже часто в еталоні, не збільшує своє значення, якщо вона всього декілька разів зустрічається у відповідях. Остаточна оцінка подібності відповіді й еталону обчислюється як результат зваженої суми логарифмів різних значень N-грам.

Постановка завдання. Метою роботи є розробка інформаційної технології для попередньої обробки відповідей, поданих у довільній текстовій формі, в автоматизованих системах тестування з метою визначення подібності відповіді й еталону. Попередній етап обробки – це етап, який не вимагає значних обчислювальних і часових ресурсів, але дозволяє зробити висновок щодо необхідності подальшого, більш детального аналізу.

Виклад основного матеріалу дослідження. Нехай $X[n]$ – масив слів відповіді, поданої у текстовій формі, $Y[m]$ – масив слів еталону (правильної відповіді). Необхідно оцінити ступінь «схожості, подібності» масивів.

Для вирішення цієї задачі запропоновано інформаційну технологію, яка дозволяє попередньо оцінити ступінь подібності відповіді й еталону Передбачається виконання процедури синтаксичного аналізу у декілька етапів:

1) На першому етапі визначається коефіцієнт подібності кожного слова відповіді й еталону.

2) На другому етапі обчислюється загальний коефіцієнт подібності для всього масиву слів $X[n]$ та $Y[m]$.

Виконання першого етапу базується на використанні N-грам. N-грамами (N=2..5) є символи, які йдуть поспіль. Спочатку N-символів першого слова відповіді порівнюються з N-символами першого слова еталону. Символи здвигаються циклічно (якщо N=2, то символи перебираються наступним чином: два символи, починаючи з 1-го , на наступній ітерації – два символи, починаючи з другого, і так далі до кінця слова). Для порівняння N-грам доцільно застосовувати методи порівняння рядків певної мови програмування, якою реалізовується алгоритм. Наприклад, для мови C# це метод Equals().

Коефіцієнт подібності, який може набувати значення від 0 до 1, обчислюється за формулою Жаккара [8, с. 101]:

$$k = a / (b + c - a) \qquad (1)$$

Для слів a – кількість N-грам (групи символів), що збігаються, слова відповіді й еталону, b – довжина слова відповіді, c – довжина слова еталону.

Слова відповіді й еталону вважаються подібними, якщо коефіцієнт подібності більший за граничний коефіцієнт. Граничний коефіцієнт відображає вимоги до відсотка символів, що збігаються, слова еталону та відповіді. Якщо, наприклад, значення коефіцієнту дорівнює 0,3, це означає, що 30% символів слів еталону та відповіді однакові. Внаслідок тестування встановлено рекомендаційне значення граничного коефіцієнту від 0,3 до 0,5. Чим більше значення граничного коефіцієнту, тим більша частина, що збігається, слів відповіді й еталону.

Далі процедура повторюється для усіх слів відповіді й еталону. На останньому етапі обчислюємо значення подібності всього набору слів за формулою (1). Для всього набору слів a – кількість подібних слів відповіді й еталону, b – кількість слів відповіді, c – кількість слів еталону.

Повторюючи процедури циклічно для (N=2..5), визначаємо максимальне значення коефіцієнту подібності відповіді й еталону. Якщо це значення менше за граничний коефіцієнт подібності (0,3), то не потрібно виконувати подальший етап обробки відповіді, оскільки вона не відповідає еталону.

Застосування N-грам для символів слова, а не для слів, які йдуть поспіль, дозволяє підвищити результативність процедури зіставлення слів. Слова-синоніми записуються в дужках після основного слова в еталоні, що дозволяє скоротити час заповнення бази правильних відповідей і, відповідно, час перевірки наданої відповіді.

У табл. 1. частково представлено результати тестування. Порівняння результатів здійснювалося з коефіцієнтом латентно-семантичного аналізу (LSA), для визначення якого застосовувався ресурс [9]. Додатнє значення («+» в таблиці) показує, наскільки отриманий результат більше (менше, якщо «-») за LSA.

Порівнюючи з коефіцієнтом LSA, можна зробити висновок, що ІТ забезпечує достатній рівень зіставлення. Технологія латентно-семантичного аналізу дає кращі результати (другий, п'ятий рядки), коли відповідь синтаксично відрізняються від еталону. Це зумовлено тим, що метод N-грам є синтаксичним методом, він не враховує семантичні класи, на відміну від LSA.

Оскільки метод має лінійну складність, його застосування ефективне під час попередньої обробки вхідної відповіді. Перевагою ІТ є можливість врахування синонімів і випадкового порядку розташування слів у відповіді й еталоні.

Висновки. Дослідження методів порівняння рядків є актуальною задачею у сфері обробки текстів, поданих природною мовою. Оскільки процедура визначення подібності відповіді й еталону як на основі синтаксичних, статистичних методів, так і на основі семантичних класів, фонетичних методів потребує значних обчислювальних і часових ресурсів, доцільно застосовувати швидкий, із лінійною складністю, легко масштабований алгоритм попередньої оцінки рівня подібності

Таблиця 1

Результати випробування інформаційної технології

Відповідь	Еталон	*k*	LSA	Δ
Сервер-проксі	Проксі-сервер	1	1	0
одиночний потік команд та одиничний потік даних	один потік команд і один потік даних	0,86	0,88	-0,02
createthraed	CreateThread()	1	0,65	+0,35
узгодження роботи потоків під час звернення до загальних ресурсів,	організація узгодженої (дружньої) роботи із загальними ресурсами,	0,5	0,75	-0,25
всесвітня мережа, до складу якої входять локальні, глобальні та інші мережі	глобальна (всесвітня) мережа, яка є сполученням (об'єднанням) локальних, регіональних і глобальних мереж	0,5	0,4	+0,1
Одночасне виконання двох і більше задач у рамках операційної системи. Класифікація видів багатозадачності: кооперативна, істина	властивість(можливість) операційної системи або середовища програмування одночасного (паралельного) виконання декількох (двох і більше) потоків (процесів, задач). Види (класифікація) багатозадачності: істинна, невитісняюча (кооперативна), витісняюча	0,49	0,43	+0,06

k (діапазон [0; 1]) – коефіцієнт подібності відповіді й еталону; LSA (діапазон [-1; 1]) – коефіцієнт узгодженості відповіді й еталону за технологією латентно-семантичного аналізу; =k–LSA.

рядків. На основі методу N-грам і коефіцієнту Жаккара запропоновано ІТ попередньої оцінки рівня подібності рядків. Застосування N-грам до груп символів, а не для слів дозволяє ефективно оцінити синтаксичну подібність відповіді й еталону. Подальші дослідження будуть спрямовані на вивчення питань застосування «нейронних мереж» у задачах семантичного аналізу подібності текстів, об'єднання технологій синтаксичного та семантичного аналізу рядків.

Список літератури:

1. Кузьма К.Т. Аналіз методів перевірки відповіді в системах тестування, поданої в текстовій формі». *Вчені записки ТНУ імені В.І. Вернадського. Серія: Технічні науки»*. 2019. Т. 29 (68) № 1, Ч. 1. С. 163–167. URL: http://www.tech.vernadskyjournals.in.ua/journals/2018/1_2018/part_1/30.pdf (дата звернення: 08. 11. 2020).
2. Кузьма К.Т., Мельник О.В. Обчислювальна технологія перевірки відповідей у системах тестування. *Вчені записки ТНУ імені В.І. Вернадського. Серія: Технічні науки»* 2020. Т. 31 (70) № 1. Ч. 1. С. 85–88. DOI: https://doi.org/10.32838/2663-5941/2020.1-1/15.
3. Кузьма К.Т. Інформаційна технологія перевірки відповідей в інтелектуальній автоматизованій системі контролю знань. *Вісник Вінницького політехнічного інституту*. 2020. № 4. С. 58–66. DOI: https://doi.org/10.31649/1997-9266-2020-151-4-58-66.
4. Вихтенко Э.М., Карманов Д.А., Син Д.З. Информационная система «Плагиат в программах студентов». *Вестник ТОГУ*. 2019. № 3 (54). С. 25–34.
5. Kumar, Praveen & Narendra, & Vimal, Bibhu & Islam, Md & Shashank, Bhardwaj. Approximate string matching Algorithm. *(IJCSE) International Journal on Computer Science and Engineering*. Vol. 02. № 03. 2010. P. 641–644. URL: https://www.researchgate.net/publication/49617308_Approximate_string_matching_Algorithm ().
6. Perez D., Gliozzo A., Strapparava C., Alfonseca E., Rodriguez P., Magnini B. Automatic assessment of students' free-text answers underpinned by the combination of a bleu-inspired algorithm and latent semantic analysis. *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference. Ed. Ingrid Russell and Zdravko Markov. California: AAAI Press*. 2005. P. 358–363. URL: https://www.aaai.org/Papers/FLAIRS/2005/Flairs05-059.pdf (дата звернення: 4.11.2020).
7. Perez Diana, Alfonseca Enrique. Application of the Bleu algorithm for recognising textual entailments. *PASCAL First Challenges Workshop, Southampton.* 2005. URL: http://u.cs.biu.ac.il/~nlp/RTE1/Proceedings/perez_and_alfonseca.pdf (дата звернення: 4.11.2020).
8. Словник-довідник з екології / уклад. О.Г. Лановенко, О.О. Остапішина. Херсон : ПП Вишемирський В.С., 2013. 226 с.
9. Latent Semantic Analysis @ CU Boulder. Sentence Comparison [Online]. URL: http://lsa.colorado.edu/(Sentence Comparison).

Kuzma K.T. INFORMATION TECHNOLOGY FOR ASSESSING THE LEVEL OF STRING MATCHING BASED ON THE N-GRAM METHOD

Natural language processing is one of the main tasks of computational linguistics. Along with improving the efficiency, quality of word processing, important issues is the time and computational efficiency of processing algorithms. Therefore, a research of methods that are easily scalable, have a linear complexity in word processing problems is relevant.

According to the results of the research, information technology (IT) was proposed for pre-processing of text answers in automated testing systems in order to determine the similarity of the answer and the standard. IT is based on the use of text parsing, named the N-gram method. N-grams are groups of characters (from two to five characters) that follow in a row. For N-grams, the similarity coefficient of each answer word and standard is calculated using the Jacquard index. Words are considered similar if the similarity coefficient is not less than the specified limit value (from 0.3 to 0.5). At the last stage the general similarity coefficient for all words of the answer is defined. Repeating the procedure cyclically for all groups of symbols determines the maximum value of the matching of the answer and the standard. If this value is less than the marginal similarity factor, then no further processing of the response is required, as it does not match the standard. Applying N-grams to word symbols, rather than words that follow in a row, can increase the effectiveness of the word matching procedure.

Based on the results of comparing IT with the method of latent-semantic analysis (LSA), it was found that the effectiveness of the method of the N-gram is high for syntactically similar answer structures and standards. If the sentence structures of the answer and the standard are significantly different, then the N-gram method in comparison with LSA gives a lower result, because it does not take into account the semantic classes in contrast to LSA.

It is determined that further research is needed on the application of "neural networks" in the tasks of semantic analysis of text similarity, combining technologies of parsing and semantic analysis of strings.

Key words: *natural language processing, N-gram, string matching, Jacquard index, parsing.*