УДК 378:001.891

Комарницька О.І. (НАДПСУ)
Комарницька І.І. (НАДПСУ)

МЕТОД СЕМАНТИЧНОГО ПОРІВНЯННЯ НЕЧІТКОЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Розроблено алгоритм семантичного порівняння нечіткої текстової інформації – відповіді на запитання, що подані студентом природною мовою, з варіантами правильних відповідей, в якому формалізовано опис лінгвістичної структури навчального контенту та відповіді. Розроблений алгоритм передбачає автоматизоване перетворення відповіді студента природною мовою до внутрісистемного вигляду, формування лексичних одиниць тексту з подальшим здійсненням морфологічного, синтаксичного, семантичного та прагматичного аналізу. Для формування частотної матриці індексованих слів удосконалено алгоритм нечіткого латентно-семантичного порівняння текстової інформації. Застосування запропонованих моделей та алгоритмів надає можливості такого: застосування морфології з метою виправлення слів, що написані студентом з помилками; врахування порядку слів у документах та їх значень для предметної сфери; синтаксичні відносини, логіку побудови терм у контексті предметної сфери; врахування омонімії, синонімії та полісемії.

Ключові слова: алгоритм; аналіз; знання; лексична одиниця; лінгвістична підсистема; метод; модель; морфологія; оцінювання; прагматика; семантика; текст.

Вступ. Розвитку моделей, методів, інформаційних технологій оцінювання знань в автоматизованих системах тестування присвячене широке коло робіт таких вчених:

В. П. Авраменко [1], І. А. Метешкина [5], Д. Г. Поспелова [3], А. А. Харламова [7], С. В. Штангей [8] та інших.

Одним із перспективних методів, що надають можливості порівняння за змістом текстів, є метод латентного семантичного аналізу [2; 3; 9; 10]. Принцип дії методу полягає у тому, що на підставі оцінки кореляції між словами та фрагментами текстів формулюється висновок щодо ступеня близькості змісту цих слів чи групи слів. Такий підхід дозволятиме лише зробити припущення щодо належності чи неналежності відповіді деякому тексту, а не щодо правильності відповіді за змістом. Крім того, для методу латентного семантичного аналізу існують певні обмеження. У ньому не використовується інформація щодо порядку слів у тексті, і, як наслідок, метод не враховує синтаксичні відношення, логіку та морфологію.

Метою дослідження є удосконалення алгоритму методу порівняння за змістом розгорнутих відповідей студентів, що подані в електронному вигляді, з варіантами правильних відповідей, поданих у XML-форматі. Цей алгоритм у подальшому використовуватиметься при здійснені латентно-семантичного аналізу.

Викладення основного матеріалу дослідження. Семантичне порівняння нечіткої текстової інформації пропонується здійснювати за таким алгоритмом [4].

1. Здійснюється формування контрольних баз даних вихідної інформації. До них належать бази даних для певної предметної сфери: “Словник” – містить перелік слів в усіх відмінках, які можуть використовуватися для опису процесів і явищ предметної сфери; “Абревіатура” – містить перелік скорочень та значень абревіатур; “Фрейми” – містить перелік словосполучень, які часто вживаються у даній предметній сфері; “Ключ” – містить перелік ключових слів предметної сфери.

2. Здійснюється перетворення відповіді студента до внутрісистемного вигляду: заміна регістру, видалення службових символів, зайвих пробілів.

3. Розбиття тексту відповіді на окремі слова. Слова подаються як окремі лексичні одиниці. Ці лексичні одиниці мають властивість нечіткості, оскільки деякі слова у вихідному тексті можуть містити помилки, неправильні закінчення, нестандартне скорочення тощо. Тоді кожне *i*-те речення представлятиме вектор лексичних одиниць, а текст відповіді можна представити формалізовано у вигляді матриці лексичних одиниць $\left\| xv_{i,j} \right\|_{n,k}$, де $i = [1; n]$ – номер речення у відповіді; $j = [1; k]$ – номер лексичної одиниці у реченні.

4. Формування бази даних лінгвістичних змінних вихідного тексту, яку можна представити такою таблицею:

Таблиця 1

База даних лексичних одиниць вихідного тексту

Код запису	Номер речення	Значення лінгвістичних змінних						Кількість змінних
n	*i*	xv_{11}	xv_{12}	…	xv_{ij}	…	xv_{1k}	kv_i
...								

5. Формування бази даних лінгвістичних змінних тексту оригіналу, з яким порівнюватиметься текст відповіді – матриця $\left\| xk_{i,j} \right\|_{n,m}$, де $i = [1; n]$ – номер речення у відповіді; $j = [1; m]$ – номер лексичної одиниці у реченні. Цю інформацію можна представити такою таблицею:

Таблиця 2

База даних лексичних одиниць тексту оригіналу

Код запису	Номер речення	Значення лінгвістичних змінних						Кількість змінних
n	*i*	xk_{11}	xk_{12}	…	xk_{ij}	…	xk_{1k}	kk_i
...								

6. Здійснюється порівняння лексичних одиниць, що містяться у матриці $\left\| xk_{i,j} \right\|_{n,m}$ – базі даних вихідного тексту, зі словами, що містяться у базі даних “Словник”. Порівняння проводиться за морфологічними частинами слова. Метою цього порівняння є заміна слів, що написані з помилками у вихідному тексті, на правильні з бази даних “Словник”.

7. На цьому кроці здійснюється оцінка подібності між матрицями $\left\| xk_{i,j} \right\|_{n,m}$ та $\left\| xk_{i,j} \right\|_{n,m}$. Така оцінка передбачає пошук кількості лінгвістичних одиниць, що належать до обох матриць, та кількості ключових слів, які присутні у матриці відповідей та базі даних “Ключ”. Крім того, проводиться оцінка збіжності порядку слідування лексичних одиниць обидвох матриць матриць. При нечіткому порівнянні використовується метрика Левенштейна [6].

8. Здійснюється пошук кількості фреймів, що одночасно присутні у матриці відповідей та базі даних “Фрейм”. Метою цього кроку є визначення належності вихідного тексту до предметної сфери. Наприклад, фрейм “мова програмування”, який включає дві лексичні одиниці “мова” та “програмування”, кожна з яких може відноситися до різних предметних сфер: “мова” до лінгвістики або літератури; “програмування” до інформатики або прикладної математики – математичне програмування. Фрейм “мова програмування” однозначно відноситься до галузі інформатики.

На підставі оцінок, одержаних на 6-8 кроках, приймається рішення щодо ступеня відповідності тексту відповіді з текстом, що містися у базі даних предметної сфери. Для формування загальної оцінки відповіді на питання використовується комплексний показник, у якому враховується: наявність у відповіді слів присутніх у зразку (з урахуванням нечіткості), відповідність структур зразка і відповіді (порядку слідування слів). Кожен із часткових показників нормується і їм присвоюються вагові коефіцієнти.

Алгоритм аналізу рядків. Один з можливих підходів, який може бути використаний для нечіткого порівняння рядків передбачає визначення метрики і обчислення відстані між рядками [10]. Чим більша відстань, тим більшою є відмінність. Оскільки в комп’ютері текстова інформація кодується числами, кожний текстовий рядок представляє собою вектор в N-мірному просторі, де N – кількість символів в рядку.

Функція $d(x,y)$ для обчислення відстані (метрики) між двома векторами x та y має мати наступні властивості:

– невід’ємність: $d(x,y) \geq 0 \ \forall x,y$;

– властивість нуля: $d(x,y) = 0 \Leftrightarrow x = y$;

– симетричність: $d(x,y) = d(y,x) \ \forall \ x,y$;

– нерівність трикутника: $d(x,z) \leq d(x,y) + d(y,z) \ \forall \ x,y,z$.

Можна побудувати багато різних метрик, які б відповідали цим властивостям. Наприклад може бути використана Евклідова метрика:

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \ .$$

Однак при обробці текстової інформації така метрика не завжди є зручною. Звичайно, кількість символів, які будуть введені у відповідь на тестове питання не є константою. Тому необхідно мати можливість порівнювати рядки різної довжини і, відповідно, розмірності просторів, в яких вони знаходяться.

Однією з найпростіших є відстань (метрика) між рядками за Хеммінгом, яка визначається як число позицій в яких символи не співпадають. Більш складною є метрика Левенштейна, з використанням якої можливим є порівняння рядків різної довжини.

Однак, застосування метрики Левенштейна показали, що в задачі нечіткого порівняння відповіді зі зразком в ході тестування, є малоефективним. При незначних відхиленнях відповіді від зразка відстань за Левенштейном є невеликою. Проте, якщо в реченні присутній зайвий пропуск, зсув, перестановка слів або інші спотворення, несуттєві з точки зору змісту, інколи отримується значна відстань. У той же час на коротких текстових рядках (одне слово), даний підхід до порівняння показав задовільні результати. Все це свідчить про те, що безпосереднє застосування метрики Левенштейна для перевірки і оцінки відповідей у тестових системах не є ефективним.

Тому для нечіткого порівняння текстової інформації у відповідях в ході тестування було розроблено алгоритм, у якому і зразок і відповідь розбиваються на окремі слова. Після чого проводиться нечіткий пошук збігів за словами між зразком і відповіддю, для чого застосовується алгоритм Левенштейна. На основі інформації про збіжність слів будується оцінка в межах від 0 до 100. Значення 100 відповідає повній збіжності за словами, 0 коли жодного слова з оригіналу немає у відповіді.

Зрозуміло, що порівняння лише за нечітким збігом окремих слів не дає об'єктивної оцінки. Тому окрім порівняння за збігом слів, проводиться перевірка відповідності структури речення відповіді і зразка. З'ясовується, наскільки порядок слів у відповіді відповідає порядку слів у зразку і оцінюється в межах від 0 до 100.

На основі двох оцінок формується інтегральна зважена оцінка, яка зараховується як відповідь. Вагові коефіцієнти встановлюються експертами.

Дослідження розробленого алгоритму показали, що він дозволяє оцінити надану у текстовому вигляді відповідь на питання в тесті. При незначних відхиленнях у відповіді від зразка виставляється оцінка близька до максимальної. Навіть при незначних спотвореннях речення, коли його зміст не втрачається, оцінка відповіді є високою. З іншого боку, коли надана відповідь не відповідає зразку, виставляється низька оцінка, яка за 100 бальною шкалою прямує до 0.

Слід відмітити, що в окремих випадках, коли питання не пов'язане з використанням у відповіді чіткої термінології, надана текстова відповідь у довільній формі за змістом може відповідати зразку, однак використання алгоритму нечіткого порівняння приводить до виставлення низької оцінки.

Для покращення алгоритму потрібно врахувати всі можливі синоніми слів і різні підходи до конструювання речень (можливо, навіть і помилки побудови речень).

При перевірці відповіді, наданої у текстовому форматі на природній мові, використовується алгоритм нечіткого порівняння, що розглянутий вище, робота якого полягає у такому:

1) зразок і відповідь приводяться до одного регістру (верхнього) і видаляються службові символи;

2) здійснюється розбиття зразку і відповіді на окремі слова;

3) для формування загальної оцінки відповіді на питання використовується комплексний показник, в якому враховується: наявність у відповіді слів присутніх у зразку (з урахуванням нечіткості), відповідність структур зразку і відповіді (порядку слідування слів).

Кожен з часткових показників є нормованим (в діапазоні 0-100) і з урахуванням вагових коефіцієнтів включається до узагальненого показника – нормовану оцінку за відповідь (в діапазоні 0-100). За необхідності вагові коефіцієнти можуть корегуватися з допомогою сторінки налаштувань sa.aspx (за замовченням перший показник враховується з коефіцієнтом 75, другий – 25 (значення показників було визначено на основі досліджень)).

При обчисленні першого показника, обчислюється відсоток наявності слів зі зразку у відповіді. Для цього проводиться нечіткий пошук слів у відповіді по кожному слову, яке присутнє у зразку. При нечіткому пошуку використовується метрика Левенштейна. У випадку, коли відстань між словами, що обчислена за метрикою Левенштейна, є нижчою за порогове значення (визначене експериментально), слова вважаються однаковими. У випадку

присутності всіх слів зразку у відповіді перший показник дорівнює 100. Коли у відповіді немає жодного слова зі зразку, значення показника – 0.

При обчисленні другого показника, визначається наскільки порядок слів у відповіді (структура речення) збігається з порядком слів у зразку. У випадку повного збігу структури речення, значення показника – 100.

У випадку, коли при формулюванні питання тесту вказано декілька варіантів зразку відповіді, проводиться нечітке порівняння з кожним зразком і виставляється максимальне число балів. Це дає можливість для питань, на які можуть надаватись різні можливі варіанти вірних відповідей, внести всі ці варіанти, як зразкові.

Висновки. Розроблений алгоритм надає можливості порівнювати за змістом тексти – відповіді на запитання, що подані студентом, з варіантами правильних відповідей та передбачає автоматизоване формування лексичних одиниць тексту з подальшим здійсненням морфологічного, синтаксичного, семантичного та прагматичного аналізу. Для порівняння нечітких лексичних одиниць використовується метрика Левенштейна.

Удосконалений алгоритм аналізу рядків надає можливість порівнювати за змістом текст поданої студентом відповіді на запитання з варіантами правильних відповідей.

Напрямом подальших досліджень є визначення порогових значень показників, на підставі яких прийматиметься рішення щодо збігу текстів відповіді та зразка й загальна оцінка за відповідь на запитання.

ЛІТЕРАТУРА:

1. Авраменко В. П. Моделирование процесса контроля знаний в системе дистанционного обучения / В. П. Авраменко, С. В. Штангей, Е. Н. Артемов // АСУ и приборы автоматики, – 2001. – Вып. 117. – С. 14 – 18.

2. Заболеева-Зотова, А. В. Латентный семантический анализ : новые решения в Internet / А. В. Заболеева-Зотова, А. Ю. Пастухов, П. В. Сердюков, Н. А. Козлова, С. А. Чернов // Информационные технологии. – 2001, № 6. – С. 67-82.

3. Кандрашина Е. Ю., Литвинцева Л. В., Поспелов Д. А. Представление знаний о времени и пространстве в интеллектуальных системах / Под ред. Д.А. Поспелова.– М. : Наука. – 328 с.

4. Комарницкая О. И. Совершенствование алгоритма латентно-семантического анализа нечеткой текстовой информации / Современный научный вестник. № 29 (225). Серия: Филологические науки. – Белгород: Руснаучкнига, 2014. – С. 58-62.

5. Метешкин К. А Искусственный интеллект в современных образовательных системах // Новий колегіум, 2001. – №5/6. – С. 20-24.

6. Расстояние Левенштейна. – [Електронний ресурс]. – Режим доступу : http://habrahabr.ru/post/114997/ Нечёткий поиск в тексте и словаре.

7. Харламов А.А., Єрмаков А.Э., Кузнєцов Д.М. Технологія обробки текстової інформації з опорою на семантичні представлення на основі ієрархічних структур з динамічних нейронних мереж, керованих механізмом уваги. // Інформаційні технології, 1998, №2. – С. 26-32.

8. Штангей С.В. Алгоритм контроля и оценивания частично правильных ответов в ходе дистанционного тестирования знаний / С.В. Штангей // 8-я Международная конференция Украинской ассоциации дистанционного образования «Образование и виртуальность – 2004». Сборник научных трудов. – Харьков-Ялта : УАДО, ХНУРЭ, 2004. – С. 348-354.

9. Gladun V., Velichko V., Svyatogor L. Hierarchical Three-level Ontology for Text Processing. International Book Series "INFORMATION SCIENCE & COMPUTING", №. 7 – FOI ITHEA Sofia, Bulgaria, 2008. – P. 11-17.

10.Kittredge, K. Synthesizing Whether Forecasts from Formatted data / K. Kittredge, A. Polguere, E. Goldberg // Proceedings of the 11 th International Conference on Computational Linguistics (COLING-86). Bonn, Germany, 1986. – P. 563-565.

Рецензент: д.т.н., проф. Замаруєва І.В.

Комарницкая О.И., Комарницкая И.И.

МЕТОД СЕМАНТИЧЕСКОГО СРАВНЕНИЯ НЕЧЕТКОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

Разработан алгоритм семантического сравнения нечеткой текстовой информации – ответов на вопросы, представленные студентом на естественном языке, с вариантами правильных ответов, в котором формализовано описание лингвистической структуры учебного контента и ответов. Разработанный алгоритм предусматривает автоматизированное преобразование ответа студента на естественном языке во внутрисистемный вид, формирование лексических единиц текста с последующим осуществлением морфологического, синтаксического, семантического и прагматического анализа. Для формирования частотной матрицы индексированных слов используется усовершенствованный алгоритм нечеткого латентно-семантического сравнения текстовой информации.

Применение предложенных новых и усовершенствованных методов, моделей и алгоритмов предоставляет возможность следующего: применение морфологии с целью исправления слов, написанные студентом с ошибками; учета порядка следования слов в документах и их значений для предметной области; синтаксические отношения, логику построения терм в контексте предметной области; учета омонимии, синонимии и полисемии.

Ключевые слова: алгоритм; анализ; знания; лексическая единица; лингвистическая подсистема; метод; модель; морфология; оценивания; прагматика; семантика; текст.

Komarnicki O.I., Komarnicki I.I.

COMPARISON OF FUZZY METHOD OF SEMANTIC TEXTUAL INFORMATION

The algorithm of fuzzy semantic comparison of textual information - answers to questions submitted by the student in natural language, with options of correct answers, which formalizes description of linguistic structure of the study content and answers has been elaborated for the first time. The algorithm provides automatic conversion of student's responses from a natural language into an intersystem form, the formation of lexical units of the text, followed by the implementation of the morphologic, syntactic, semantic and pragmatic analysis. In order to form a frequency matrix of the indexed words there has been improved an algorithm of fuzzy latent-semantic comparison of textual information.

The application of the proposed new and improved methods, models and algorithms provides the possibility of the following: an application of morphology for correction the words with mistakes written by students; taking into account the order of words in documents and their meanings for the subject area; syntactic relations, logic of the term in the context of the subject area; taking notice of homonymy, synonymy and polysemy. Scientific significance of the results obtained in the dissertation is creation of the automated data processing intellectual tools, applied in the natural language and based on fuzzy data, semantics and pragmatics of lexical units of the subject area texts.

Keywords: algorithm; analysis; knowledge; lexical unit; linguistic subsystem; method; model; morphology; evaluation; pragmatics; semantics; text.