УДК 004.415

к.пед.н., доц. **Ваколюк Т.В.** (НАДПСУ, м. Хмельницький)
Комарницька О.І. (НАДПСУ, м. Хмельницький)

АЛГОРИТМ НЕЧІТКОГО СЕМАНТИЧНОГО ПОРІВНЯННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ

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

Ключові слова: аналіз, алгоритм, граматика, лексична одиниця, морфологія, прагматика, слово, синтаксис, семантика, фрейм.

В статье описан алгоритм, который предоставляет возможность сравнивать по содержанию тексты – ответы на вопросы, представленные студентом, с вариантами правильных ответов. Разработанный алгоритм предусматривает автоматизированное формирование лексических единиц текста с последующим осуществлением морфологического, синтаксического, семантического и прагматического анализа. Для сравнения нечетких лексических единиц используется метрика Левенштейна.

Ключевые слова: анализ, алгоритм, грамматика, лексическая единица, морфология, прагматика, слово, синтаксис, семантика, фрейм.

The algorithm, which enables comparison the texts-answers given by students with the texts-true answers by implication is described in the article. Developed algorithm provides automated formation of the lexical units, which are analyzed morphologically, syntactically, semantic and pragmatically. The Levenstein metric is used to compare the indistinct lexical units.

Keywords: analysis, algorithm, grammar, lexical unit, morphology, pragmatics, word, syntax, semantics, frame.

Вступ. Одним із важливих завдань у системах, що використовують електронні інформаційні засоби навчання, є автоматизований контроль перевірки знань, які студенти подають у вигляді тексту. Для реалізації такої перевірки необхідні відповідні методи та алгоритми. У більшості із широко розповсюджених комп'ютерних систем при тестуванні використовуються запитання, що передбачають вибір відповіді із запропонованого переліку. Таке тестування має певні недоліки:

1) тести орієнтовані на короткі відповіді: число, рік, правило, перелік тощо. Не передбачено самостійної відповіді на запитання;

2) студенти мають підказку, тобто здійснюють вибір лише із запропонованих відповідей, що позбавляє їх творчого мислення;

3) утруднена підготовка тестів, які спрямовані на перевірку творчих здібностей та логічного мислення. Такі тести, навіть за умови, що вони будуть розроблені, то відповідь буде зведена до вибору, а не до творчості;

4) тестування не передбачає перевірки розгорнутих письмових робіт;

5) відповідь можна вгадати. Якщо, наприклад, на запитання пропонується чотири варіанти відповіді, то імовірність "вгадування" складає 1/4.

Розвитку моделей, методів, інформаційних технологій оцінювання знань в автоматизованих системах тестування присвячено широке коло робіт таких вчених: В.М. Глушкова, В.С. Аванесова, Д.Г. Поспелова, Т.П. Подчасової, І.А. Метешкина, В.Є. Ходакова, С.В. Штангей та інших.

Метою даної роботи є розроблення алгоритму порівняння за змістом розгорнутих відповідей студентів, що подані в електронному вигляді, з варіантами правильних відповідей, поданих у xml-форматі.

Розв'язання проблеми. Одним із перспективних методів, що надають можливості порівняння за змістом текстів є метод латентного семантичного аналізу [1; 2; 3; 4]. Принцип дії методу полягає у тому, що порівняння множини усіх контекстів, у яких слова або групи слів вживаються, та контекстів, у яких вони не вживаються, надає можливості на підставі оцінки кореляції між словами та текстами зробити висновок щодо ступеня близькості змісту цих слів чи групи слів. Отже, 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.

Не дивлячись на більш ніж 50 років досліджень в галузі розроблення методів аналізу текстів на природній мові, універсального вирішення більшості задач обробки текстів не існує. Це обумовлено проблемами формалізації природної мови як надскладної семіотичної системи, що складається з необмеженого числа підсистем, внаслідок чого вона не може бути остаточно формалізована.

При підготовці відповіді студент може зробити помилку у словах, неправильно побудувати речення, вживати нестандартні скорочення і абревіатуру тощо. Оскільки наша задача оцінити знання а не граматику, процес аналізу вихідної інформації включає декілька етапів, у ході яких усуваються зазначені помилки [5; 6]. Структурна схема процесу переведення запиту з природної мови (варіанту відповіді) у внутрісистемне представлення даних наведено на рис. 1.

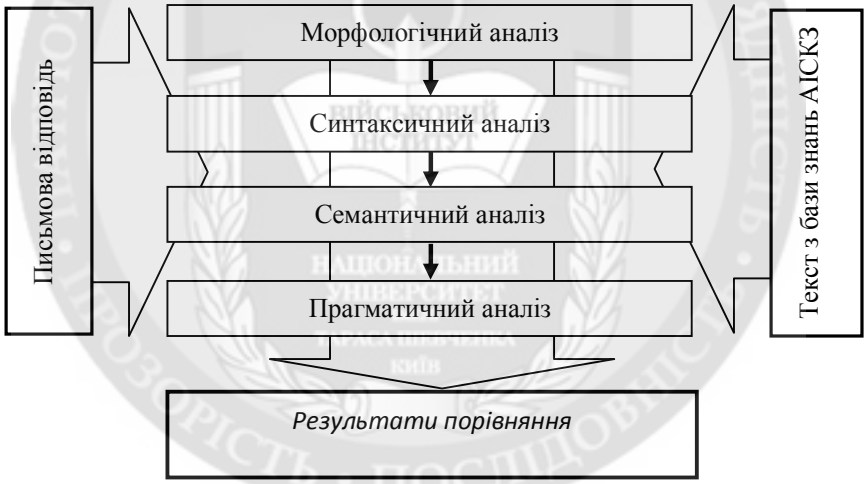

Рис. 1. Структурна схема процесу переведення запиту з природної мови у внутрісистемне представлення даних

На виході *морфологічного блоку* матимемо правильно побудовану послідовність лексичних одиниць. Неправильно написані у варіанті відповіді слова, абревіатури тощо, будуть замінені правильними словами (фреймами), одержаних із бази даних "Словник". Ця послідовність поступає далі на вхід блоку *синтаксичного аналізу*, метою якого є отримання синтаксичної правильно побудованої структури фрази. При використанні природної мови однакове за змістом висловлювання може бути описане різними способами. Тому структура текстових представлень може суттєво відрізнятись від зразка. Отже, для порівняння за змістом текстової відповіді зі зразком потрібно виділити ці знання. Вирішення даної задачі

можливе з проведенням *семантичного аналізу*. Саме семантичний аналіз дає можливість з довільного тексту на природній мові виділити змістовну структуру (знання). Порівняння поданої відповіді і зразка проводитиметься за декількома ознаками: за кількістю слів, кількістю ключових слів і фреймів, порядком слів, значимістю слів і фреймів. У ході *прагматичного аналізу* визначається, чи належить відповідь до визначеної предметної сфери.

Розглянемо детальніше роботу алгоритму нечіткого семантичного порівняння текстової інформації.

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 та зайвих пробілів.

3. Розбиття тексту відповіді на окремі слова. Слова подаються як окремі лексичні одиниці. Ці лексичні одиниці володіють властивістю нечіткості, оскільки деякі слова у вихідному тексті можуть містити помилки, неправильні закінчення, нестандартне скорочення тощо. Тоді кожне *i*-те речення представлятиме вектор лексичних одиниць

$$XV_i = \left\langle xv_1; xv_2; \ldots, xv_j; \ldots, xv_k \right\rangle \quad . \tag{1}$$

Тоді текст відповіді можна представити формалізовано у вигляді матриці лексичних одиниць:

$$XV = \left\| \begin{matrix} xv_{11}; xv_{12}; \ldots, xv_{1j}; \ldots, xv_{1k} \\ xv_{21}; xv_{22}; \ldots, xv_{2j}; \ldots, xv_{2k} \\ \ldots \\ xv_{n1}; xv_{n2}; \ldots, xv_{nj}; \ldots, xv_{nk} \end{matrix} \right\|, \tag{2}$$

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 (2). 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
...								

Текст оригіналу відповіді представлятиме матрицю лексичних одиниць:

$$XK = \left\| \begin{matrix} xk_{11}; xk_{12}; \ldots, xk_{1j}; \ldots, xk_{1k} \\ xk_{21}; xk_{22}; \ldots, xk_{2j}; \ldots, xk_{2k} \\ \ldots \\ xk_{n1}; xk_{n2}; \ldots, xk_{nj}; \ldots, xk_{nk} \end{matrix} \right\|, \quad (3)$$

де $i = [1; n]$ – номер речення тексту; $j = [1; k]$ – номер лексичної одиниці у реченні.

6. Здійснюється порівняння лексичних одиниць, що містяться у матриці XV(2) – 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 *XV*(2) та *VK*(3). Така оцінка передбачає пошук кількості лінгвістичних одиниць, що належать до обох матриць, та кількості ключових слів, які присутні у матриці (2) та базі даних “Ключ”. Крім того, проводиться оцінка збіжності порядку слідування лексичних одиниць матриць (2) та (3). При нечіткому порівнянні використовується метрика Левенштейна [7].

8. Здійснюється пошук кількості фреймів, що одночасно присутні у матриці (2) та база 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.

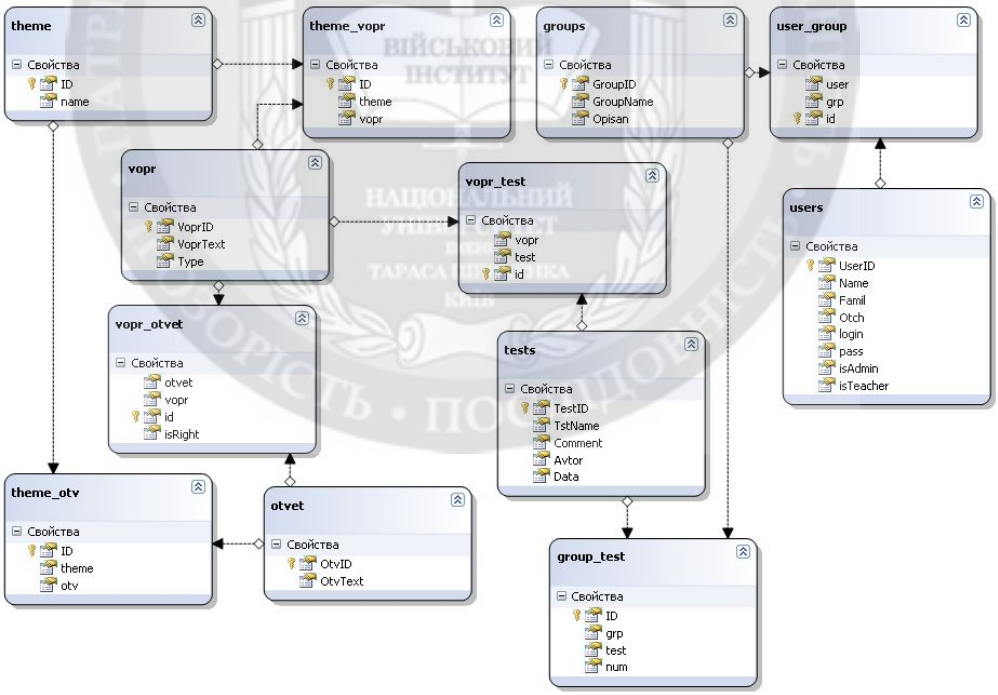

Рис. 2. Відношення між таблицями бази даних з інформацією про тести

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 зразку і відповіді (порядку слідування слів). Кожен із часткових показників нормується і їх присвоюються вагові коефіцієнти.

Розробка інформаційного забезпечення системи тестування передбачає розробку реляційної бази даних для зберігання інформації необхідної для проведення тестування (списки користувачів, назви тестів, перелік питань, варіантів відповідей, інформація про предмети, їх тематику, навчальні заняття, викладачів тощо). Структура баз даних представлена на рис. 2. У якості платформи для побудови інформаційного центру інформаційно-телекомунікаційної системи обрано СКБД MS SQL Server Express Edition.

В інтелектуальній системі оцінювання знань використовується реляційні бази даних, що передбачає формування таблиць з визначенням зв'язків (відношень) між ними. Для забезпечення більшої гнучкості пропонується розділити зберігання інформації в декількох базах даних. Перша база даних використовуватиметься для зберігання баз: списки користувачів, назви тестів, перелік питань, варіантів відповідей тощо. У другій базі даних зберігатиметься інформація про предмети, їх тематику, навчальні заняття, викладачів та їх закріплення за навчальними групами тощо.

Висновок. Таким чином, описаний алгоритм надає можливості порівнювати за змістом тексти відповіді на запитання подані студентом 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. Заболеева-Зотова А.В., Пастухов А.Ю., Сердюков П.В., Козлова Н.А., Чернов С.А. Латентный семантический анализ: новые решения в Internet. Информационные технологии. 2001, № 6.

2. Марченко О.О. Алгоритми семантичного аналізу природно-мовних текстів: Дис... канд. фіз.-мат. наук: 01.05.01 / Київський національний ун-т ім. Тараса Шевченка. – К., 2005. – 150 арк. – Бібліогр.: арк. 142-150.

3. Лесько О.М., Рогушина Ю.В. Использование онтологий для анализа семантики естественно-языковых текстов / Проблеми програмування. 2009. № 3. – С. 59-65.

4. Малащук Е.В. Средства семантического поиска близких по общему контексту документов / Штучний інтелект, 2006, № 4. – С. 613-618.

5. Катеринчук І.С., Комарницька О.І. Новітні інформаційні технології оцінювання знань у вищих навчальних закладах. Збірник наукових праць № 51. Частина ІІ. – Хмельницький: НАДПСУ, 2010. – С. 56-59.

6. Інтелектуальна системи автоматизованого оцінювання знань у вищих навчальних закладах // Звіт про НДР/ НАДПСУ, ХДЦНТіЕІ (№ 0109V005890). – Хмельницький, 2008. – 120 с.

7. Расстояние Левенштейна. [Електронний ресурс]. – Режим доступу: http://ru.wikipedia.org/wiki/Расстояние_Левенштейна.

Рецензент: д.т.н., проф. Сбітнєв А.І.