

МЕТОД ОЦІНЮВАННЯ ДОВІЛЬНИХ ВІДПОВІДЕЙ У КОМП'ЮТЕРНИХ СИСТЕМАХ ТЕСТУВАННЯ ЗНАНЬ

Abstract: *The article describes students' knowledge method by comparison analysis of the textual answer with given standart text identifying their correspondence. There are given practical settlements. Particular attention is paid to the theoretical analysis on terms synonymity basis by operation with sets key words' assesment method, comparison analysis.*

Key words: *assessment, synonymity basis, analysis of the textual answer, sets.*

Анотація: *У статті розглянутий метод оцінювання знань студентів шляхом порівняльного аналізу тексту відповіді з заданим еталонним текстом та визначення їх релевантності, наведені практичні розрахунки. Особлива увага приділена дослідженню теоретико-множинної моделі уявлення знань на основі синонімії термінів предметної області за допомогою операцій над множинами.*

Ключові слова: *релевантність, синонімія термінів, аналіз тексту, множини.*

Аннотация: *В статье рассмотрен метод оценки знаний студентов с помощью сравнительного анализа текстового ответа с заданным эталонным текстом и определение их релевантности, приведены практические расчеты. Особое внимание уделено исследованию теоретико-множественной модели представления знаний на основе синонимии терминов предметной области с помощью операций над множествами.*

Ключевые слова: *релевантность, синонимия терминов, анализ текста, множества.*

1. Введення

Впровадження прогресивних форм навчання та розвиток сучасних інформаційних технологій створює об'єктивну необхідність переходу до комп'ютерного тестування знань студентів. При цьому на перший план встає задача автоматичного оцінювання знань студентів. Ця задача є достатньо простою, якщо студенту пропонується вибрати одну чи декілька правильних відповідей з заданого набору варіантів, але значно ускладнюється, якщо процедура тестування передбачає формулювання розгорнутої відповіді у довільній формі, тобто власними словами на природній мові. В останньому випадку оцінювання знань студентів можливо лише шляхом порівняльного аналізу тексту відповіді з заданим еталонним текстом та визначення їх релевантності. При цьому повинні враховуватись усі можливі словотворні форми, терміни предметної області та граматичні структури висловлювання відповіді з використанням синонімів. Це, у свою чергу, вимагає розробки теоретико-множинної моделі уявлення знань на основі синонімії термінів предметної області за допомогою операцій над множинами [1].

Метод, що пропонується, дозволяє визначати релевантність еталонного визначення і відповіді студента з урахуванням синонімії понять.

Підхід заснований на припущенні, що знання предметної області визначається її тематичним словником, умінням правильно розставити слова, тобто дати коректне формулювання понять предметної області, що має саме значеннєвий збіг або семантично близьке значення з тлумаченням даного поняття в тезаурусі предметної області. При цьому враховується, що фактичне формулювання терміна предметної області може бути не у виді однозначної послідовності термінів мови предметної області (твердого формулювання), а з використанням синонімічно близьких понять і термінів (тобто формулювання, що відповідає інфологічній моделі досліджуваного терміна у свідомості того, кого навчають).

Терміни і визначення будь-якої предметної області являють собою підмножину загальної мови. Завдання полягає в тому, щоб вибрати таку підмножину, у якій зберігалася б властивість загальної мови: будь-яке поняття предметної області повинне виражатися через інші елементи цієї ж

підмножини. Іншими словами, термінологічна підмножина мови предметної області успадковує всі ознаки його множини, що її породила – загальну національну мову.

Мова як сукупність лінгвістичних одиниць (слів) являє собою повну замкнуту множину елементів. Мається на увазі, що будь-яке поняття мови можна виразити через інші елементи цієї ж множини.

Використовуючи діаграми Ейлера-Венна, можна умовно представити мову у виді деякої замкнутої частини простору, кожна крапка якого являє собою визначену лінгвістичну одиницю (слово, привід, знак і т.п.).

Загальна кількість слів і слівформ у будь-якій природній мові обмежені в силу загальних об'єктивних принципів побудови самої мови. Тому можна стверджувати, що з математичної точки зору кожна така множина, так само як і її підмножини, є кінцевою множиною, тобто має кінцеве число складових його елементів.

Частина інформації (як природної мови, так і мов предметних областей) представлена у виді текстових документів (неструктуровані і слабоструктуровані дані) та у виді реляційних баз даних (структуровані дані). У реляційних базах даних (БД) зберігається вже до деякої міри оброблена інформація, у тому числі і та, що стосується до розряду текстової [2].

2. Оцінювання довільної відповіді з використанням синонімії

Нехай S – множина усіх термів національної мови. Оскільки кількість термів у мові не є нескінченною, множина S є кінцевою і чисельною, а також неупорядкованою:

$$S = \{s_i : 1 \leq i \leq n\},$$

де s_i – терм; n – ціле число.

Представимо сукупність термів будь-якої предметної області як множину T . Дана множина є частиною множини термів мови S і має всі її властивості:

$$T \subset S ; T = \{t_i : 1 \leq i \leq m\},$$

де m – ціле число; $m \leq n$.

Використаємо такі умовні позначення:

$p(T)$ – сукупність всіх можливих підмножин множини T ;

e – елемент підмножин сукупності $p(T)$, який найбільше повно відповідає описуваному об'єктові або ситуації.

Будемо вважати, що один терм не може бути використаний для опису декількох об'єктів або ситуацій, тобто $e_i \neq e_j$.

Із сукупності $p(T)$ можна виділити такі підмножини D_e , елементи яких характеризуються семантичною близькістю, що відповідає синонімічному зв'язку слів один з одним:

$$D_e \subseteq p(T).$$

Елемент e , щодо якого утвориться множина (синонімічний ряд) D_e , назвемо базовим термом, інші елементи множини D_e (слова, об'єднані синонімічним зв'язком з базовим термом) назвемо залежними термами. У рамках кожної множини D_e введемо відношення між його елементами, що позначає наявність синонімічного зв'язку між базовим і залежними термами. Позначимо цей зв'язок функцією $\theta(e, t)$. На кожній підмножині D_e вона задає частковий порядок, такий що:

$$t \leq s \leftrightarrow \theta(e, t) \leq \theta(e, s); \quad \forall t \in D_e; \quad \forall s \in D_e.$$

Цій частковий порядок може бути заданий у числовому виді відображенням пари термів на множину дійсних чисел:

$$\theta: T \times T \rightarrow R.$$

Тоді для двох довільних термів e і t можна вказати кількісний показник їхньої синонімії:

$$k = \theta(e, t); \quad e, t \in D_e,$$

значення якого знаходиться в інтервалі $[-1, 1]$.

Властивості та обмеження функції $\theta(e, t)$ такі:

- 1) якщо $k = \theta(e, t) = 1$, то терм t є повним синонімом терму e ;
- 2) якщо $\theta(e, t) \geq \theta(e, s)$, то $t < s$; $e, t, s \in D_e$;
- 3) якщо $\theta(e, t) - \theta(t, s) > 0$, то $t < s$; $e, t, s \in D_e$;
- 4) якщо $\theta(e, t) > 0$ і $\theta(t, s) = 1$, то $\theta(e, s) = \theta(e, t)$;
- 5) якщо $\theta(t, s) = 0$, то $\theta(t, s) = 0$.

Деякі залежні терми з множини D_{e_i} можуть бути базовими й у множинах D_{e_j} . Це означає наявність непрямого синонімічного зв'язку між базовим термом множини D_{e_i} і залежними термами множини D_{e_j} .

Встановимо відповідність між термами еталонного визначення і термами відповіді, спираючись на поняття синонімічної відповідності термів, і на основі цієї відповідності можна обчислити показник релевантності еталонного визначення і довільної відповіді. Висловлення розглядаються як сукупність термів. Таким чином, еталонне визначення варто розглядати як сукупність базових термів, а відповідь – як сукупність термів t , для кожного з яких потрібно знайти відповідний базовий терм e . Пошук відповідності базового терму і терму відповіді припускає визначення функції $e = \varphi(t)$ і обчислення величини синонімічної відповідності $k = \theta(e, t)$. Таким чином, пари $\langle e, k \rangle$ дозволять характеризувати терм t стосовно терму – еталону e . Це означає нормалізацію термів відповіді до базових термів [3].

Нехай A – множина термів еталонного визначення, B – множина термів відповіді.

Тоді опис еталонного визначення і відповіді має вигляд

$A = \{e_1, e_2, \dots, e_i, | \leq i \leq N\}$, де N – кількість термів еталонного визначення.

$B = \{t_1, t_2, \dots, t_i, | \leq i \leq M\}$, де M – кількість термів відповіді.

Для обчислення відповідності між термами еталонного визначення і відповіді необхідно характеризувати терми t стосовно терм-еталонів e . З цією метою проводиться нормалізація термів відповіді до термів еталонного визначення.

Метою нормалізації є приведення термів, що утримуються у відповіді до термів, які входять в еталонне визначення. Результатом нормалізації є одне зі співвідношень між множинами A і B .

1. $A = B$ – відповідь того, кого навчають, цілком збігається з еталонним визначенням.
2. $A \subset B$ – відповідь того, кого навчають, містить усі терми з еталонного визначення і надлишкові терми.
3. $B \subset A$ – відповідь того, кого навчають, частково відповідає еталонному визначенню. У відповіді відсутні деякі базові терми.
4. $A \cap B \neq \emptyset$ – у відповіді того, кого навчають, та еталонному визначенні є однакові терми.
5. $A \cap B = \emptyset$ – відповідь того, кого навчають, не відповідає еталонному визначенню.

При формуванні питання щодо визначення деякого терміна E розглянутої предметної області той, кого навчають, повинний сформулювати визначення за аналогією.

Встановивши відповідність між термами еталонного визначення і термами відповіді, спираючись на поняття синонімічної відповідності термів і на основі цієї відповідності обчислити показник релевантності еталонного визначення і відповіді. Для цього потрібно:

1. Серед пересічних множин вибрати множину $\{a_i, b_j\}$, яка характеризується найвищим числовим показником $k_h = \theta(a_i, b_j)$.

2. Видалити з подальшого розгляду всі інші множини, в яких присутні обрані елементи a_i і b_j .

3. Серед інших множин продовжувати вибір і видалення множин з максимальним показником $k_a = \theta(a_b, b_a)$ за цим правилом доти, поки всі пересічні множини не будуть вичерпані.

У випадку, якщо одночасно існують кілька множин з однаковим максимальним числовим показником k , то варто вибрати тільки одне з них, а над іншими знову провести операцію нормування.

В результаті всі пересічні множини повинні бути вилучені. Встановлено однозначну відповідність між значимими термами еталонного і фактичного визначень, тобто нормалізація довершена.

Нехай еталонна відповідь така: „Програма - послідовність операцій над даними, необхідна для обробки інформації і реалізації визначеного алгоритму”.

Після обробки еталонного і фактичного визначень (відкидання стоп-слів, виключення з розгляду загальноживаних слів, що не відносяться до розглянутої предметної області, виділення термів) одержимо таку послідовність термів:

для еталонного визначення

данн, керуван, компонент, систем, обробк, інформаці, цілий, реаліз, визначен, алгоритм;

для фактичного визначення

сукупн, данн, послідовн, операці, обробк, інформаци, реаліз, визначен, алгоритм.

Складемо дві множини термів: A – множина термів, що входять в еталонне визначення,

B – множина термів, що входять у відповідь:

a_1 = данн	a_6 = інформац	b_1 = сукупн	b_5 = обробк
a_2 = керуван	a_7 = цілий	b_2 = данн	b_6 = інформаци
a_3 = компонент	a_8 = реаліз	b_3 = послідовн	b_7 = реаліз
a_4 = систем	a_9 = визначен	b_4 = операці	b_8 = визначен
a_5 = обробк	a_{10} = алгоритм		b_9 = алгоритм

З таблиць кількісних показників синонімії термів предметної області «Інформатика», складених експертами, була побудована таблиця 1, у якій зібрані зведення про синонімію термів, використаних в еталонному і фактичному визначеннях.

Таблиця 1. Показники синонімії термів еталонного визначення і відповіді

Еталонне визначення										
	Данн	Управл	Компонент	Систем	Оброк	Інформації	Цілий	Реалізац	Визначен	Алгоритм
Сукупн				0,8						
Данн	1		0,5			0,7				
Послідовн				0,2						
Операц			0,6							
Обробка		0,8			1			0,5		
Інформац	0,7					1				
Реалізац					0,3			1		
Визначен									1	
Алгоритм				0,6						1

На основі даних табл.1 отримана сукупність множин.

Одержимо сукупність функцій $k_i = \theta(a_i, b_j)$, які означають числові показники синонімії термів еталонного і фактичного визначень:

$$\begin{aligned}
 k_1 &= \theta(a_1, b_2) = \text{синонім}\{\text{данн, данн}\} = 1; & k_9 &= \theta(a_4, b_9) = \text{синонім}\{\text{систем, алгоритм}\} = 0,6; \\
 k_2 &= \theta(a_1, b_6) = \text{синонім}\{\text{данн, інформац}\} = 0,7; & k_{10} &= \theta(a_5, b_5) = \text{синонім}\{\text{обробк, обробк}\} = 1; \\
 k_3 &= \theta(a_2, b_4) = \text{синонім}\{\text{керуван, операці}\} = 0,2; & k_{11} &= \theta(a_5, b_7) = \text{синонім}\{\text{обробк, реалізац}\} = 0,3; \\
 k_4 &= \theta(a_2, b_5) = \text{синонім}\{\text{керуван, обробк}\} = 0,8; & k_{12} &= \theta(a_6, b_2) = \text{синонім}\{\text{інформаци, данн}\} = 0,7; \\
 k_5 &= \theta(a_3, b_2) = \text{синонім}\{\text{компонентів, данн}\} = 0,5; & k_{13} &= \theta(a_6, b_6) = \text{синонім}\{\text{інформаци, інформаці}\} = 1; \\
 k_6 &= \theta(a_3, b_2) = \text{синонім}\{\text{компонентів, операці}\} = 0,6; & k_{14} &= \theta(a_6, b_6) = \text{синонім}\{\text{реалізац, обробк}\} = 0,5; \\
 k_7 &= \theta(a_4, b_1) = \text{синонім}\{\text{систем, сукупн}\} = 0,8; & k_{15} &= \theta(a_8, b_7) = \text{синонім}\{\text{реалізац, реалізац}\} = 1;
 \end{aligned}$$

$$k_8 = \theta(a_4, b_3) = \text{синонім}\{\text{систем, послідовн}\} = 0,2; \quad k_{16} = \theta(a_9, b_8) = \text{синонім}\{\text{визначен, визначен}\} = 1;$$

$$k_{17} = \theta(a_{10}, b_9) = \text{синонім}\{\text{алгоритм, алгоритм}\} = 1.$$

З описаних даних можна бачити, що деякі множини $\{a_i, b_j\}$ перетинаються в точці a_i або b_j .

Наприклад, множина $\{a_1, b_2\} = \{\text{данн, данн}\}$ перетинається з множиною

$\{a_1, b_6\} = \{\text{данн, інформаці}\}$ у точці $a_1 = \text{данн}$:

$$\{a_1, b_2\} \cap \{a_1, b_6\} = \{a_1\}.$$

Це означає, що терму «данн» еталонного визначення відповідають два терма фактичні визначення: «данн» і «інформац». За умовою потрібно вибрати один з двох термів. Для цього порівняємо їхні функції $k_h = \theta(a_i, b_j)$.

$$k_1 = \theta(a_1, b_2) = 1,$$

$$k_2 = \theta(a_1, b_6) = 0,7,$$

$$\theta(a_1, b_2) > \theta(a_1, b_6).$$

Отже, множина $\{a_1, b_6\}$ з подальшого розгляду виключається і не приймає участі у процесі формування оцінки знань.

Те саме можна сказати про множини $\{a_4, b_3\}$ і $\{a_4, b_9\}$. Вони не приймають участі у процесі формування оцінки знань, тому що пересічна з ними множина $\{a_4, b_1\}$ має найбільше значення функції $k = 0,8$.

3. Висновки

Розроблена теоретико-множинна модель уявлення та контролю знань на основі синонімії між термінами предметної області, яка дозволяє вирішити задачу обчислення релевантності двох речень на природній мові, представлених у виді вільного тексту з довільної кількості слів. За допомогою використання чисельного показника к синонімії термінів для обчислення показників синонімічної релевантності еталонної відповіді та відповіді можливо аналізувати та контролювати знання.

СПИСОК ЛІТЕРАТУРИ

1. Харламов А. Автоматический структурный анализ текстов // Открытые системы. – 2002. – № 10. – С. 35–39.
2. Цаленко М.Ш. Моделирование семантики в базах данных. – М.: Наука, 1989. – 288 с.
3. Шаров Д.А. Система анализа формулировок // Материалы Международной конференции "Компьютерная лингвистика и интеллектуальные технологии". – 2004. – <http://www.dialog-21.ru/Archive/2004/Sharov.pdf>.