

УДК 004.415:378.147/(042.3)

**ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ МОВНИХ СТРУКТУР
В СИСТЕМАХ ТЕСТУВАННЯ ЗНАНЬ**

Л.М. Бадьоріна

Національний авіаційний університет

e-mail: vada@ukr.net

Для сучасного етапу розвитку вищої освіти характерними є декілька тенденцій, пов'язаних з інтенсифікацією освітніх процесів, їх уніфікацією – наближенням до європейських та світових стандартів, необхідністю ефективного впровадження Болонського процесу з урахуванням національних особливостей, здобутків та традицій української вищої школи. Сказане вимагає розробки нових методів та засобів контролю освітнього процесу – як у цілому, так і окремих його складових, насамперед – методів та засобів контролю індивідуальних знань, набутих студентом під час засвоєння того чи іншого навчального курсу. Цим зумовлюється бум у галузі теорії та практики застосування різних автоматизованих систем навчального характеру, який відбувається зараз у світі, у тому числі і в Україні. Тим часом, робіт присвячених саме розробці моделей, методів та автоматизованих засобів контролю індивідуально набутих знань насправді не так вже й багато, хоча потреба у них надзвичайно гостра. Труднощі у побудові систем такого характеру визначаються тією обставиною, що індивідуальні набуті студентами знання зазвичай виражені в природній формі, що, внаслідок складності формалізації природної мови, створює додаткові проблеми при проектуванні систем. Таким чином зумовлюється актуальність дослідження і розробка концептуальних та алгоритмічних засад систем контролю знань, представлених у вигляді природномовних текстів.

Взагалі застосування природномовних механізмів стає дедалі чіткішою ознакою сучасних інформаційних технологій та систем і, фактично, вже зараз слугує критерієм їх інтелектуальності. Отже на часі є широке впровадження систем автоматизованої обробки природної мови та пов'язана із цим необхідність створення формалізованих моделей і на їх основі програмного забезпечення. Як відзначалося вище, особливої значущості впровадження цих методів та засобів набуває для створення систем комп'ютерного тестування та оцінювання знань студентів. На даний момент актуальним є завдання розробки моделей аналізу відповідей до завдань відкритого типу, що потребує введення з клавіатури визначення (формулювання) того або іншого терміна предметної галузі.

Створення новітніх технологій, ускладнення технічних систем та інше потребує підвищення якості підготовки фахівців для народного господарства України. Одним із засобів досягнення високого рівня цієї підготовки є розробка та впровадження систем автоматизованого контролю знань, здатних обробляти та оцінювати природномовні відповіді тих, кого навчають, подані у довільній формі.

Реалізація такої здатності може бути здійснена шляхом змістовного порівняння двох визначень одного терміну предметної галузі: еталонного, який зберігається в базі знань системи, та введенного ззовні тим, кого навчають, поданого природною мовою з довільною семантичною структурою та кількістю слів. Це потребує застосування моделей представлення знань предметної галузі, розробки методів аналізу природномовних відповідей з урахуванням синонімії понять, алгоритмів оцінювання ступеня змістовної близькості довільної відповіді до еталонного визначення.

Враховуючи вище сказане було запропоновано багатофункціональну лінгвістичну модель, яка дозволяє порівнювати різні текстові визначення предметної галузі з урахуванням синонімії понять. Розроблено алгоритм обробки текстової відповіді того, кого навчають на завдання відкритого типу, який дозволяє установлювати змістовну відповідність між термінами предметної галузі.

Для представлення знань пропонуються наступні лінгвістичні моделі:

$$z_q^1 = x_1 \Delta_1 x_2 \Delta_2 \dots \Delta_{q-1} x_q; \quad (1)$$

$$z_q^2 = (x_1; x_2; \dots; x_q); \quad (2)$$

$$z_q^3 = (x_1; x_{i2}; \dots; x_{in}); \quad (3)$$

$$z_q^4 = \mathbf{R}^3 z_q = (\rho(x_1); \rho(x_{i2}); \dots; \rho(x_{in})); \quad (4)$$

$$z_q^5 = \{\mu z_q; \sigma z_q\}; \quad (5)$$

$$z_q^6 = \mathbf{R}^5 z_q = \{\mathbf{R}^{\mu} z_q; \mathbf{R}^{\sigma} z_q\}; \quad q = \overline{1, p}; \dots; x_i \in V, \quad (6)$$

де x – довільна лексема, яка з урахуванням її словозмінних класів може бути представлена змінною і незмінною компонентою;

$\rho(x)$ – частина лексеми x (квазіоснова);

$\rho(x_i)$, – змінні частини слова у граматичному значенні; $i = \overline{1, n}$;

$V = \{x\}$ – клас елементарних інформаційних одиниць (поняття предметної галузі);

Δ_i – знак пунктуації, пробіл; $i = \overline{1, q}$;

q – довжина ланцюжка;

$\mathbf{Z}$ – сукупності певних ланцюжків елементів інформаційних одиниць;

$\mathbf{R}$ – відображення ланцюжків, які порівнюються;

$\mathbf{X}$ – сукупність ланцюжків визначеної довжини.

Модель (1) збігається з представленням терміна його текстовою формою.

У моделі (2) термін z_q^2 представляється послідовністю термів, що його складають.

У моделі (3) термін z_q^3 представляється неупорядкованою множиною слів, що містяться у множині $(x_1; x_2; \dots; x_q)$. Це означає, що кожне слово з послідовності $(x_1; x_2; \dots; x_q)$ входить у цю модель з кратністю не більше 1.

Модель (4) будується з моделі (3) заміною x_i на $\rho(x_i)$ або $\mathbf{X}$ на $\mathbf{X}^R$.

У моделі (5) термін z_q^5 представляється двома компонентами, де z_q^{μ} визначається згідно з моделлю (3), а z_q^{σ} – згідно з моделлю (1). При цьому враховується той факт, що деякі складні терміни можуть містити словосполучення, які для адекватного врахування семантики потребують збереження семантико-синтаксичної та граматичної структури, що має матеріальне вираження в будові конкретних складових-словоформ та роздільників Δ .

Модель (6) вдержується з моделі (5) застосуванням до неї оператора $\mathbf{R}$ (або, що те саме, заміною $\mathbf{X}$ на $\mathbf{X}^R$).

Набір лексикографічних моделей, які специфікують терміни предметної галузі через елементарні терми – елементи відповідної Г-системи, також специфікують дефініції зазначених термів. Зазначені моделі використовуються при побудові тезаурусної системи $\Sigma[\mathbf{Z}]$.

Відношення синонімії між мовними одиницями x та y , яке задається умовою близькості їхніх семантичних станів $|c(x) - c(y)| < \varepsilon$, $x S y$, підлягає кількісній оцінці $\delta = 1 - \varepsilon$, $\delta \in (0, 1)$ за допомогою процедури експертизи словника синонімів, в якому подано лексикографічний опис синонімічних рядів. Тут S – синонімічний ряд.

Відношення синонімії, поширене на складові тезаурусу предметної галузі $\Sigma[\mathbf{Z}]$, визначається на множині дефініцій термінів:

$$C^{\Sigma}(\mathbf{Z}) = \{C^{\Sigma}(z) \mid \forall z \in \Sigma(z)\} = \{\{C^{\Sigma}_1(z); C^{\Sigma}_2(z); \dots; C^{\Sigma}_{l(z)}(z)\} \mid \forall z \in \Sigma(z)\},$$

де $C^{\Sigma}(Z)$ – скінчена множина визначень терміна.

На множині $C^{\Sigma}(z)$ вводиться відношення релевантності REL , яке є кількісною мірою змістової (семантичної) близькості двох ланцюжків $A=z_M$ та $B=z_N$ (довжиною M та N , відповідно):

$$REL: C^{\Sigma}(Z) \times C^{\Sigma}(Z) \rightarrow \Delta,$$

де Δ – підмножина множини невід’ємних чисел. При цьому вважається, що ланцюжок B є релевантним ланцюжкові A , тобто $A REL B$, тоді і тільки тоді, коли значення функції $REL(A, B)$ не менше якогось певного $\delta \in \Delta$: $REL(A, B) \geq \delta$, вибір якого залежить від специфіки предметної галузі та конкретних завдань дослідження та оцінювання.

Кількісне значення релевантності двох висловлень A і B обчислюється за формулою:

$$REL(A, B) = \omega \eta,$$

де: ω – функція числових значень показника синонімії термів, що входять до A та B , тобто функція елементів матриці синонімії $K(x, y)$; η – функція довжин ланцюжків A та B (тобто M та N).

Функція η має такі властивості:

$$1) \eta = \eta_{\max} \text{ тоді і тільки тоді, коли } M = N;$$

Так само очевидно, що функція η є симетричною по змінних M та N , отже вона є симетричною відносно свого максимального значення. З цього випливає друга властивість η :

$$2) \eta(M, N) = \eta(N, M) \text{ і } \eta \text{ симетрична відносно значення } \eta_{\max}.$$

Найпростішою функцією з такими властивостями є функція $\eta = |M - N|$.

Очевидно, що якщо ланцюжки A та B сильно відрізняються за довжиною (тобто за кількістю значущих термів), то вони не можуть бути релевантними, оскільки кожний терм, відсутній в одному ланцюжкові, вносить свій внесок у семантику іншого ланцюжка і, з урахуванням кожного такого терму, різниця в семантиці ланцюжків стає все більшою, а їх взаємна релевантність – меншою.

У процесі досліджень було висунуто визначення про формулювання знання термінології предметної галузі (термінів) з використанням слів-синонімів зі словникового запасу тих, кого навчають, з певної предметної галузі.

Для доведення цієї гіпотези розроблено способи реєстрації синонімів, методика формування міри близькості фактичних і еталонних визначень термінів предметної галузі, запропоновано відповідні алгоритми реалізації і розроблено прикладний програмний засіб.

Результати дослідження були апробовані при проведенні тестування студентів зі знань термінології предметної навчальної дисципліни “Інформатика та програмування” як найбільш близької до професійної діяльності здобувача.

Література

1. Ретинская И.В. Системы и методы поддержки принятия решений по оценке качества и выбору компьютерных средств учебного назначения (Обзор). // Информационные технологии, 1997, № 6. – С. 42-44.
2. Программированное обучение и кибернетические обучающие машины: Сборник статей под ред. Шестакова А.И. – М.: Сов. радио, 1963. – 247 с.
3. Кириличев Б.В., Широков Л.А., Рабинович П.Д. Системный анализ проблемы создания интеллектуальных компьютерных обучающих комплексов. Сборник научных трудов МГИУ. – М.: МГИУ. – 1996. – С. 166-171.
4. Бадьоріна Л.М. Метод нормалізації текстової відповіді в комп'ютерних системах тестування // Проблеми інформатизації і управління – 2006.- №18.- С.12-15.