УДК 004.681
DOI: 10.15587/2313-8416.2018.135550

ЗАСТОСУВАННЯ ЛІНГВІСТИЧНОЇ ТЕХНОЛОГІЇ ПРИ ОЦІНЮВАННЯ ЗНАНЬ

© Л. М. Бадьоріна

В статті запропоновано лінгвістичну технологію, за допомогою якої можливо здійснити когнітивне розпізнання текстових об'єктів та врахувати їх мовні особливості в межах предметної галузі. Обробка тексту спрямована на виявлення в тексті основних компонент знань, відношення між ними з урахуванням мовної специфіки. Для професійної підготовки, зокрема в галузях діяльності, пов'язаних з використанням точної, семантично достовірної термінології, де спотворення формулювань, стандартизованих визначень термінів або недостатнє їх розуміння може призвести до відхилень у виконанні професійної діяльності, помилкам

Ключові слова: *інформаційні технології, природна мова, багатофункціональна модель, лінгвістична багафункціональна модель*

1. Вступ

Одним з напрямків реалізації Національної програми "Освіта. Україна XXI сторіччя" є розробка та впровадження комплексних інформаційно-освітніх середовищ, які поєднують навчальні, науково-методичні інформаційні ресурси, використовуючи сучасні інформаційні технології. Останнім часом у зв'язку з бурхливим розвитком систем автоматизованого навчання актуалізувалася проблема побудови формальних моделей, що описують ті чи інші аспекти зазначеної галузі. Серед них особливо вирізняються моделі та засоби орієнтовані на проведення автоматизованого оцінювання результатів навчального процесу. Слід відзначити, що якщо побудова навчальних контентів та цілісних систем у відзначеній ділянці розроблені достатньо повно, то, власне, автоматизація процесів оцінювання поки що перебуває на етапі розвитку. Це пов'язано, насамперед, з тією обставиною, що результати навчального процесу представляються як відповіді на екзаменаційні та інші питання і через це мають природномовну форму. Отже, технологія оцінювання в такий спосіб набуває характеру автоматичного (автоматизованого) порівняння природно мовних текстів або їх фрагментів. Очевидно, що така технологія апріорі мусить бути мовнозалежною і будуватися для кожної мови окремо. Тим часом, нам невідомі навіть загальносистемні наукові праці, присвячені цьому предмету, що й зумовило необхідність написання даної праці.

2. Загальна структура системи оцінювання відповідей та засоби її моделювання

З огляду на природномовну специфіку предмета нашого дослідження основною теоретичною конструкцією для побудови моделі предметної галузі ми тут обираємо модель лексикографічного середовища (або інтегрованої лексикографічної системи), яку було розроблено [1].

При побудові нашої моделі необхідно сконструювати формальні кореляти мовних конструкцій, які відображають зміст предметної галузі, причому моделювання мусить відбуватися як з боку форми, так і з боку змісту. При цьому ми мусимо враховувати, що мовна система являє собою складну ієрархію різнорівневих комплексів одиниць, об'єктів та відношень.

Першим кроком на шляху побудови моделі, на нашу думку, мусить бути моделювання сукупності лексичних одиниць, що відображають «словник» предметної галузі, яка є об'єктом дослідження, оскільки саме лексична підсистема відіграє центральну роль у мовній системі взагалі. Зазначений словник, як ми вважаємо, повинен містити, насамперед, «клас термів», який складається з граматично специфікованої сукупності лексем предметної галузі. Адекватною моделлю для цього є модель граматичної Л-системи (Г-системи), в структурі якої виділяються такі структурні елементи (рис. 1):

1. Клас елементарних інформаційних одиниць $V=\{x\}$, що відповідає класу всіх слів української мови (у нашому випадку це є класом термів предметної галузі);

2. Клас початкових форм, що для змінюваних частин мови відповідає вихідним (словниковим формам);

3. Клас розкладів слів: $\pi(x)=\rho(x)*\{\omega_i(x)\}$, і відповідно, множина незмінних $\{\rho\}$ та змінних $[F]^k\equiv\{\omega_i(x)\}$ частин для всіх слів (квазіоснов та квазіфлексій, відповідно);

4. Скінченна множина словозмінних (парадигматичних) класів: $\cup\, t_i/\pi_i$;

5. Оператор парадигматизації π, який ставить у відповідність кожному слову x його повну словозмінну парадигму $[x]$;

6. Оператор лематизації λ, який ставить у відповідність будь-якому слову $\xi \in [x]$ його вихідну форму x_0.

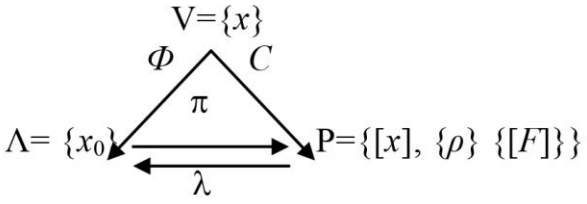

Рис. 1. Схематична структура Г-системи

Визначимо на Г оператор: $\boldsymbol{R} = \pi_\rho^\circ \Phi$, де π_ρ – є обмеження π на ρ. Тоді для будь-якого $\xi \in [x]$ справедливо:

$$\boldsymbol{R}\xi = \rho(\xi). \quad (1)$$

Оператор $\boldsymbol{R}$ буде використано при побудові системи аналізу відповідей. [2, 3].

3. Мета та завдання дослідження

Мета дослідження – аналіз, створення, обґрунтування та практична реалізація лінгвістичної технології, зокрема активація та використання здобутих знань.

Для досягнення мети були поставлені наступні задачі:

1. Дослідити інформаційні технології інтелектуальної обробки знань.

2. Дослідити когнітивне розпізнавання текстових об'єктів з певної предметної галузі.

3. Формалізувати граматичні структури природної мови з метою коректної обробки висловів.

4. Представлення знань предметної області

Знання – орієнтований підхід до автоматизації оцінювання знань тих, хто навчається, за текстовими відповідями передбачає наявність засобів приведення еталонних відповідей, представлених природною мовою, і відповідей тих, хто навчається, до формалізованого подання у вигляді моделі знань про предметну галузь. Кожна модель відповіді тих, хто навчається порівнюється на відповідність з еталонною моделлю. Предметна галузь містить різноманітні логіко-семантичні відношення між поняттями, по кожному з яких необхідно встановлювати ступінь відповідності. Введемо наступні обмеження на оцінювання відповідності відповідей еталонному зразку:

– розглядаємо в якості відповідей лише визначення (тлумачення) термінів і понять з певної навчальної дисципліни;

– на множині відношень, які задані для термінів та понять з відповідної навчальної дисципліни, виділимо тільки родо-видові відношення та відношення синонімії [4, 5].

Визначення поняття – в самому широкому розумінні є логічна операція, в процесі якої розкривається зміст поняття. В основі правил еталонних визначень термінів і понять покладено 7 правил, що вивчаються формальною логікою:

– поняття визначається через родові і видові відмінності;

– визначення повинно мати такий же вимір, що і поняття, тобто обсяг значення поняття, що визначається, і поняття, за допомогою якого здійснюється визначення, мають відповідати одне одному;

– видовою відмінністю має бути ознака або група ознак, що притаманні лише даному поняттю, і відсутні в інших поняттях, що відносяться до того ж родового поняття; визначення не повинно містити кола, тобто поняття, що визначається, не повинно визначатися через поняття, яке стає зрозумілим лише через поняття, що визначається; визначення не повинно бути тільки заперечним, оскільки заперечення вказує на відсутні ознаки і не дає суттєвих ознак, що характеризують дане поняття;

– визначення не повинно бути суперечливим з точки зору формальної логіки;

– визначення повинно бути ясним, чітким і не містити подвійного тлумачення.

Нехай S – множина всіх еталонних визначень понять і термінів з відповідної навчальної дисципліни, представлених у вигляді природно мовного тексту й укладених за вище визначеними правилами. Множина S є кінцевою й неупорядкованою:

$$S = \{s_i : 1 \le i \le n\}, \quad (2)$$

де s_i – визначення терміна; n – ціле число.

Сукупність відповідей тих, хто навчається, представлених також природною мовою, визначимо як множину T . Дана множина є підмножиною множини S та має всі її властивості:

$$T \subset S;\ T = \{t_i : 1 \le i \le m\}, \quad (3)$$

де m – ціле число; $m \le n$.

Кожна відповідь з множини T може містити терміни і поняття, які пов'язані родо - видовими відношеннями, або відношеннями синонімії з поняттями відповідної еталонної відповіді множини S. Взаємозв'язок термінів і понять в заданій предметній галузі (навчальній дисципліні) представимо у вигляді тезаурусу. Тезаурус – словник, що відбиває семантичні відношення між поняттями в певній предметній галузі і призначений для пошуку заданого слова за його смисловими зв'язками в з іншими словами [6].

Структура тезаурусу, як правило, включає наступні відношення:

поняття: = <рід-вид> <частина-ціле> <синоніми> <антоніми> < асоціації>.

Відношення рід-вид дозволяє включити у пошукове поле більш абстрактні або конкретні поняття. Відношення частина-ціле включає у пошукове поле частини цілого об'єкту. Відношення синонімії й антонімії дозволяє здійснювати пошук синонімів й антонімів. Відношення асоціації різноманітні та індивідуальні за своєю природою і вказують на контекстну залежність пошукового поняття.

Відповідь того, хто навчається, визначається певною структурою понять і термінів. З урахуванням визначених обмежень кожне поняття в тлумачній частині може описуватися через синоніми [7, 8].

Елемент e, відносно якого утворюється множина (тобто синонімічний ряд) D_e, назвемо базовим термом, інші елементи множини D_e (слова-синоніми) назвемо залежними термами. Необхідно встановити відповідність між термами еталонного визначення і термами відповіді, спираючись на поняття синонімічної відповідності термів, яке підставляється з тезаурусу, можна обчислити показник релевантості і еталонного визначення і відповіді того, хто навчається. Таким чином, еталонне визначення

слід розглядати як сукупність базових термів, а відповідь як сукупність термів *t*, для кожного з яких необхідно знайти відповідний базовий терм *e*. [3].

Якщо *A* - множина термів еталонного визначення, *B* – множина термів відповіді, то формалізоване подання еталонного визначення і відповіді буде мати наступний вигляд:

A= {e_1, e_2, ..., e_i, 1≤ и≤ N}, де N - кількість термів еталонного визначення.

B= {t_1, t_2, ...,t_i, 1≤ <и≤ <M}, де M — кількість термів відповіді.

В результаті ми можемо отримати одне з наступних співвідношень між множинами *A* і *B*.

1. $A = B$ – відповідь того, хто навчається, повністю збігається з еталонною відповіддю.

2. $A \subset B$ – відповідь, того, хто навчається, містить всі терми з еталонної відповіді і додаткові терми.

3. $B \subset A$ – відповідь того, хто навчається, частково відповідає еталонній відповіді, в ній відсутні деякі базові терми.

4. $A \cap B = \varnothing$ – відповідь того, хто навчається повністю не відповідає еталонній відповіді.

5. $A \cap B \neq \varnothing$ – еталонна відповідь і поточна відповідь спільні терми.[2]

Продемонструємо вище викладене на наступному прикладі. Нехай ми маємо еталонне визначення:

«Програма – опис ***алгоритму*** розв'язання задачі, заданий на ***мові обчислювальної машини***.» [9].

В еталонному визначенні жирним курсивом виділені ключові базові терми, які є відповідають умовам необхідності і достатності правильної відповіді для тих, хто навчається. Інші поняття є додатковими. Вони також можуть мати синонімічні ряди, але не враховуються під час кількісного оцінювання відповіді того, хто навчається. Тобто, для правильної відповіді визначається два необхідних і достатніх поняття, які за правилами побудови тлумачної частини терміну «програма» формують його унікальні відмінні ознаки. Для цих базових термів з тезаурусу можна побудувати наступний синонімічний ряд:

Алгоритм := {сукупність правил; послідовність операцій; сукупність дій};

Мова обчислювальної машини:= {мова програмування; штучна мова; машинна мова; формальна мова, мова ЕОМ}.

Позначимо через А1 множину, що визначає синонімічний ряд для поняття «алгоритм», і через А2 – синонімічний ряд для поняття « мова обчислювальної машини.

Тоді формалізоване подання еталонної відповіді буде мати наступний вигляд:

Програма := опис{представлен} А1 ⊂ {алгоритм; сукупн+правил; послідовн+операцій; сукупн+дій} розв'язання{вирішення; обчислення}

задач

∧

задан{представлен;опис} на А2 ⊂ {мов+обчислювальн+машин; мов+програмування; машинн+мов; формальн+мов; мов+ЕОМ}.

В даному прикладі поняття, через які відбувається тлумачення, представлені у вигляді пошукових образів, через знак «+» поєднуються слова, які складають термін для заданої навчальної дисципліни, логічна операція ∧ вказує на обов'язкову присутність двох базових термів. Інші відношення в силу введених раніше обмежень пропущені. Дане представлення є основою для порівняння з поточними відповідями тих, хто навчається.

Після необхідних перетворень формалізоване подання відповіді того, хто навчається буде мати наступний вигляд:

Програма := *A1*{послідовн+операций}
над дан
необхідн
обробк/обробок
інформаці
реалізаці
***A1*{ алгоритм}.**

З наведеного прикладу видно, що поняття тлумачної частини терміну «програма» збігаються тільки з множиною ***A1*** еталонного зразку. Причому у відповіді знайдено 2 еквіваленти, оскільки з формули (1) витікає, що вона приймає значення **1**, якщо знайдено хоча б один відповідник, тому згортання всіх знайдених відповідників з однієї множини дає значення **1**, тобто $f(a_1,b)=1$, $f(a_2,b)=0$.

Кількісна оцінка обчислюється за формулою (4).

$$K = \frac{1}{2} = 0.5$$

Таким чином, якщо привести інтервал [0,1] до десятибальної шкали оцінювання, то дана відповідь буде мати оцінку 5.

5. Релевантність термінів та їх дефініцій

Синонімія окремих термів, які в лінгвістичному сенсі є елементами лексичної системи, на складові тезаурусу предметної галузі Σ[**Z**]. У цьому завданні є два аспекти – формальний і змістовий [1].

З формальної точки зору завдання полягає у встановленні семантичної близькості, аналогічної до властивості синонімії, але не на множині окремих термів, а на множині ланцюжків вигляду $x_1\Delta_1 x_2\Delta_2 \ldots \Delta_{q-1}x_q$, $q=1, 2, \ldots$, де x довільне слово, знак пробіл – знак пунктуації, за умови, що елементи $x_1, x_2, \ldots, x_q$, потрапляють до області визначення функції $K(x\ y)$. Змістовий аспект передбачає встановлення відношення семантичної близькості, аналогічної до властивості синонімії, на множині дефініцій термінів:

$$C^{\Sigma}(\mathbf{Z})=\{C^{\Sigma}(z)\,|_{\forall z\in\Sigma(z)}\}= \{\{C^{\Sigma}{}_1(z); C^{\Sigma}{}_2(z); \ldots; C^{\Sigma}{}_{l(z)}(z)\}\,|_{\forall z\in\Sigma(z)}.\}. \quad (5)$$

Оскільки поняття синонімії в лінгвістиці визначається лише для лексичної системи, то для встановлення змістової (семантичної) близькості елементів з $C^{\Sigma}(z)$ введемо назву *відношення релевантності*, яке позначатимемо символом ***REL***.

З цією метою визначимо кількісну міру релевантності двох ланцюжків $A=z_M$ та $B=z_N$ (довжиною *M* та *N*, відповідно), яку позначатимемо як:

$$\boldsymbol{REL}(A,B). \quad (6)$$

Таким чином, визначається відображення $\boldsymbol{REL}$: $C^{\Sigma}(\mathbf{Z}) \times C^{\Sigma}(\mathbf{Z}) \rightarrow \Delta$, де Δ – певна підмножина множини невід'ємних чисел. При цьому вважатимемо, що ланцюжок B є релевантним ланцюжкові A, тобто A $\boldsymbol{REL}$ B , тоді і тільки тоді, коли значення функції $\boldsymbol{REL}(A,B)$ не менше якогось певного $\delta \in \Delta$: $\boldsymbol{REL}(A,B) \geq \delta$, вибір якого залежить від специфіки предметної галузі [10] та конкретних завдань дослідження та оцінювання .

Зазначена формула насправді враховує певні ефекти семантичної близькості мовно-інформаційних об'єктів так що її можна застосовувати як інструмент до аналізу ситуацій, що виникають при порівнянні еталонних (поданих у нормативних джерелах, зокрема підручниках) формулювань понять та дефініцій предметної галузі з фактичними їх формулюваннями, що є об'єктами оцінювання, якщо і перші і другі представлені ланцюжками вигляду A і B.

У сукупності отримані результати формують основу для створення інтелектуальних інформаційних технологій , знання-орієнтованих систем, які передбачають якісну підготовку та отримання знань.

Результати дослідження дозволили застосувати лінгвістичну технологію для представлення знань на основі термінів предметної галузі, що дало можливість оцінювати релевантність текстів, поданих у вигляді природного тексту з довільної кількості слів.

6. Висновки

Лінгвістична технологія на основі методу обробки знань, які містяться в навчальних текстах дозволить перевести на новий рівень програмне і прикладне забезпечення. Отримано узагальнені результати проведеного дослідження, а саме:

1. Досліджено інформаційні технології інтелектуальної обробки знань. На лінгвістичному етапі розпізнавання тексту враховано морфологічний, синтаксичний і семантичний аналіз розпізнавання, вилучення знань про предметну галузь, які містяться у тексті.

2. Досліджено когнітивне розпізнавання текстових об'єктів з певної предметної галузі, зокрема розпізнавання багатозначних текстових одиниць, що дозволило вирішити задачу вибору на множині понять. Основою методу вибору на заданій множині понять є алгоритм співставлення канонічного (тобто представленого в базі знань системи) значення поняття з його контекстним значенням.

3. На прикладах показано формалізацію граматичних структур природної мови з метою коректної обробки висловів.

Література

1. Широков В. А. Інформаційна теорія та системотехнічні засади комп'ютерної лексикографії: автореф. дис. … д-ра техн. наук. Київ, 1999. 32 с.

2. Badyorina L. M. Synonymy of therms and terms and its presentation in the informative system // Problemy systemnoho pidkhodu v ekonomitsi. Kyiv: NAU, 2012. P. 206–212.

3. Badyorina L. M. Method of grammatical structure formalization of natural language // Visnyk NAU. 2013. Issue 1. P. 44–47.

4. Пещак М. М. Стан і перспективи комп'ютерної лексикографії в Україні // Мовознавство. 1996. № 4-5. С. 8–11.

5. Пиотровский Р. Г. Лингвистический автомат и его речемыслительное обоснование. Минск, 1999. 195 с.

6. Бадьоріна Л. М., Замаруєва І. В. Метод кількісного оцінювання відповідей в системах тестування знань // Системний аналіз та інформаційні технології. 2011. № 2. С. 41–46.

7. Пиотровский Р. Г. Моделирование фонологических систем и методы их сравнения. Москва-Ленинград, 1966. 300 с.

8. Забезпечення процесів діяльності з визначеним рівнем надійності в ІТС спеціального призначення / Теленик С. Ф. та ін. // Збірник наукових праць ВІТІ НТУУ „КПІ". 2007. № 3 С. 134–138.

9. Шенк Р. Обработка концептуальной информации. Москва, 1980. 360 с.

10. Павлов О. А., Халус О. А. Модифікований алгоритм розв'язання задачі мінімізації сумарного запізнення виконання завдань: міжнар. наук.-пр. Інтерн.-конф. // Перспективні інновації в науці, освіті, виробництві та транспорті. 2007.

Дата надходження рукопису 10.05.2018

Бадьоріна Любов Миколаївна, доктор технічних наук, старший викладач, кафедра комп'ютерних наук, Київський національний університет культури і мистецтв, вул. Євгена Коновальця, 36, м. Київ, Україна, 01601

E-mail: vada@ukr.net