О. І. КОМАРНИЦЬКА

МОДЕЛІ ТА МЕТОДИ ЛІНГВІСТИЧНОГО АНАЛІЗУ ТЕКСТУ В СИСТЕМАХ ОЦІНЮВАННЯ ЗНАНЬ

У статті обґрунтовано функціональну структуру інтелектуальної системи лінгвістичного аналізу розгорнутої текстової відповіді із застосуванням моделей штучного інтелекту. Розроблено алгоритм семантичного порівняння нечіткої текстової інформації (відповідей на запитання, що подані студентом природною мовою, з варіантами правильних відповідей), в якому формалізовано опис лінгвістичної структури навчального контенту та відповіді. Для формування частотної матриці індексованих слів удосконалено алгоритм нечіткого латентно-семантичного порівняння текстової інформації.

Ключові слова: алгоритм, лексична одиниця, метод, модель, семантика, текст, фрейм, штучний інтелект.

Стрімкий розвиток науки, упровадження інформаційних та високих технологій ведуть до зростання обсягів неструктурованої науково-технічної інформації, що подається переважно природномовними текстами. Зазначений, дуже слабко контрольований процес створює багато складних проблем, які подекуди зводять нанівець переваги, що надають інформаційно-комп'ютерні технології, адже проаналізувати за припустимий час «вручну» надвеликі масиви інформації людина не здатна, а забезпечити повну формалізацію змісту природномовних документів і в такий спосіб адаптувати його до автоматичного опрацювання неможливо навіть теоретично. На думку деяких дослідників [1], для цього необхідно навчити комп'ютери використовувати знання про предметну сферу, зокрема вміти в автоматичному режимі пов'язувати текстові фрагменти з концептами відповідних предметних галузей. Одним із способів концептуалізації текстових документів є семантична розмітка тексту, або семантичне анотування (маркування). Прикладом системи, в якій використовується семантичне маркування, є Semantic Wiki [2].

Слід, однак, визнати, що на сьогодні не існує моделей та засобів, які б достатньою мірою враховували особливості природної мови при інтелектуальному оп-

[1] *Добров Б. В.*, *Лукашевич Н. В.* Онтологии для автоматической обработки текстов: описание понятий и лексических значений.— www.dialog21.ru/dialog2006/-materials/html/Dobrov_files/editdata.mso; *Лесько О. М.*, *Рогушина Ю. В.* Использование онтологий для анализа семантики естественно-языковых текстов // Пробл. програмування.— 2009.— № 3.— С. 59–65.; *Марченко О. О.*, *Дерев'янченко О. В.* Застосування семантико-синтаксичної моделі для поліпшення розпізнавання рукописних текстів // Вісн. Київ. ун-ту.— 1999.— Вип. 4.— С. 200–205.; *Палагін О. В.*, *Світла С. Ю. та ін.* Про один підхід до аналізу та розуміння природномовних об'єктів // Комп'ютерні засоби, мережі та системи.— 2008.— № 7.— С. 128–137.

[2] *Krötzsch M.*, *Schaffert S.*, *Vrandečić D.* Reasoning in Semantic Wikis // G. Antoniou et al. Reasoning Web 2007 : Lecture Notes.— Berlin, 2007.— T. 4636.— S. 310–329.

рацюванні текстової інформації. Це пояснюється труднощами, що виникають при формальному описі системи природної мови, зумовленими її сутністю. Адже особливістю природної мови є її принципова нечіткість [3]. Свідомість людини здатна сприймати нечіткі судження та з контексту робити цілком певні висновки про змісти, актуалізовані в природномовних конструкціях. Але машина здатна сприймати лише те, що чітко задано в описах відповідних моделей. Багатозначність та непрогнозованість контекстної семантики мовних конструкцій не просто знижує якість роботи систем автоматичного опрацювання текстів, а й часто робить їх функціонування неможливим.

Сказане цілком стосується й проблематики інформатизації освітньої сфери. Окремою її ділянкою є інтелектуалізація засобів контролю освітнього процесу, одним із аспектів якого є впровадження мовно-інформаційних методів діагностування навчальних досягнень студентів. До них належать методи і засоби автоматизованого контролю знань, здатних обробляти і оцінювати відповіді, подані природною мовою в довільній формі.

Вважається, що перевагами таких систем є можливість повнішого охоплення змісту навчальної дисципліни, мінімізація витрат часу на проведення процедури тестування, можливість автоматизації контролю і оцінювання результатів, інтеграція систем тестування з авторитетними інформаційними масивами з предметних галузей та ін. Проте слід визнати, що існуючі комп'ютерні системи тестування мають чимало недоліків: більшість із них містять запитання, що передбачають короткі відповіді з дуже обмеженим лінгвістичним репертуаром. Часто в системі передбачено лише вибір із запропонованих варіантів; трудомісткою є підготовка тестів, спрямованих на перевірку творчих здібностей і логічного мислення та ін. Деякі із сучасних автоматизованих систем контролю знань містять також і завдання відкритого типу, однак у більшості з них відповідь зараховується як правильна, якщо вона цілком збігається з одним із еталонних варіантів тексту.

Таким чином, у процесі автоматизованого контролю знань студентів виникають суперечності: між ефективністю процедур тестового контролю знань та об'єктивністю їх оцінки; між великими обсягами інформації, що потребують лінгвістичного аналізу при оцінюванні знань та недосконалістю технологій його здійснення. З цього випливає необхідність побудови комплексних лінгвістичних моделей, адаптованих до формального представлення в системах, вільних від перелічених вище вад. Отже, метою статті є аналіз моделей та методів комплексного лінгвістичного аналізу природномовного тексту в системах оцінювання знань студентів, курсантів та слухачів, де враховуються морфологічні, семантичні, синтаксичні та прагматичні його властивості.

У статті запропоновано концепцію побудови інтелектуальної системи оцінювання знань з функціональною структурою (див. рис. 1).

Ця структура інтелектуальної системи оцінювання знань, умінь та навичок студентів вищих навчальних закладів містить такі модулі: базу даних (предмети, модулі, теми, навчальні групи); базу знань (предмети, модулі, теми); лінгвістичну підсистему (аналізатори граматики, орфографії, семантики та прагматики); систему навчання; систему оцінювання.

---

[3] *Аверкин А. Н., Батыршин И. З., Блишун А. Ф. та ін.* Нечеткие множества в моделях управления и искусственного интеллекта.— М., 1986.— 312 с.; *Заде Л.* Понятие лингвистической переменной и его применение к принятию приближенных решений.— М., 1976.— 165 с.; *Рыжов А. П.* Элементы теории нечетких множеств и измерения нечеткости.— М., 1998.— 116 с.

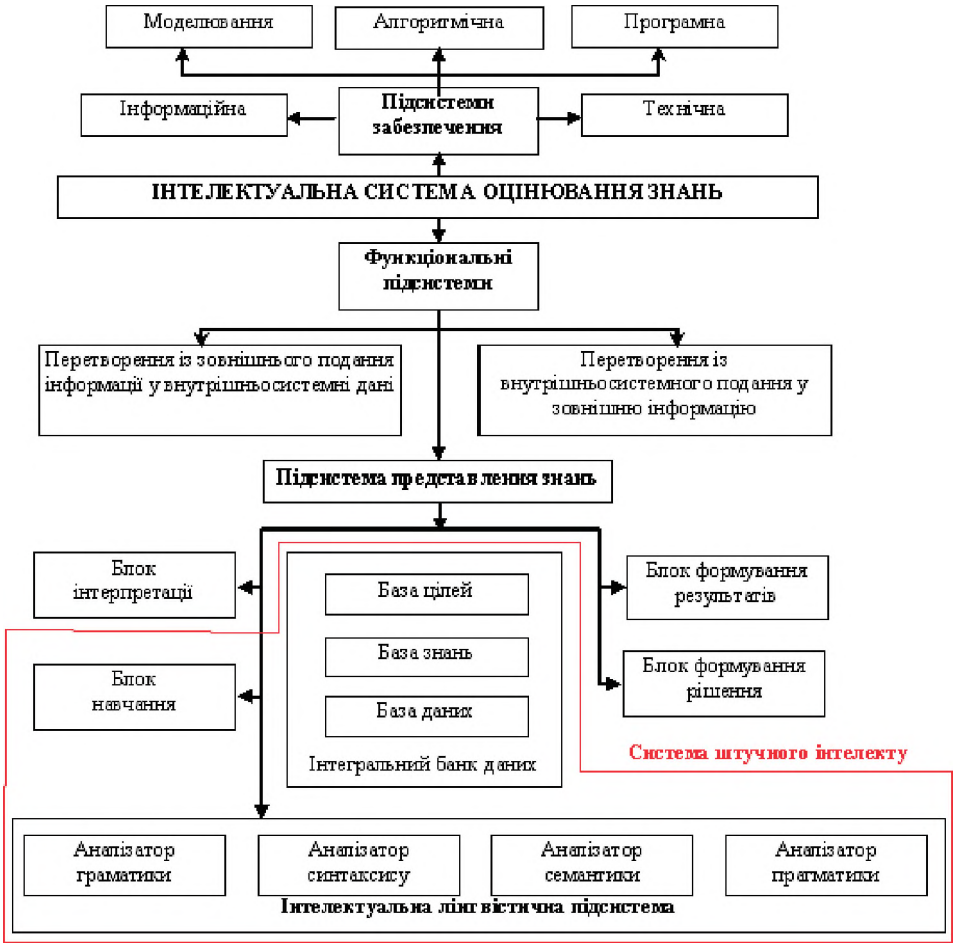

Рис. 1. Функціональна структура інтелектуальної системи оцінювання знань

Для якісної та повноцінної роботи система автоматичного лінгвістичного аналізу повинна мати можливість проаналізувати текст відповіді на запитання з позицій сучасної української морфології, синтаксису, семантики та прагматики, згенерувати текст відповіді в логічне внутрішнє представлення та синтезувати відповідь природною мовою. Структурну схему процесу переведення варіанта відповіді з природної мови у внутрішньосистемне подання наведено на рис. 2.

У процесі роботи морфологічного блоку здійснюється нормалізація словоформ, для кожної лексеми визначається відповідна змістова інформація: лексико-граматичні класи, граматичні, синтаксичні та семантичні характеристики. Слова та абревіатури з помилками замінюються правильними словами, одержаними з бази даних «Словник». Ця послідовність потрапляє далі на вхід блоку синтаксичного аналізу, метою якого є отримання синтаксичної структури фрази, яка записується у вигляді дерева складників або дерева залежностей. У разі використання дерева залежностей для кожного елемента-вершини аналізованого ланцюжка вказується елемент, що ним керує, і тип зв'язку між ними (крім джерела-вершини графа).

Природною мовою однакову за змістом думку можна подати різними лексичними конструкціями. Через це структура текстового подання відповіді може істотно відрізнятися від зразка. Отже, для порівняння за змістом текстової від-

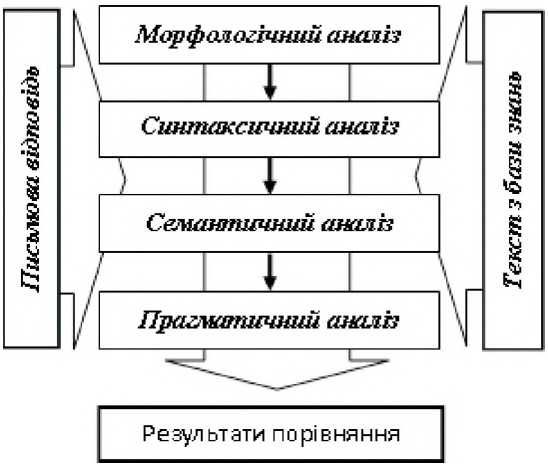

Рис. 2. Схема обробки даних

повіді зі зразком потрібно визначити зміст. Це завдання розв'язується за допомогою семантичного аналізу — виділення з тексту змістової структури (знання), а далі — порівняння семантичного наповнення тексту відповіді та зразка.

Одним з надійних методів порівняння за змістом текстів є метод латентного семантичного аналізу (ЛСА), який дозволяє на підставі оцінки кореляції між словами і текстами зробити висновок про ступінь близькості змісту цих слів чи групи слів. Однак для методу ЛСА існують певні обмеження: у ньому не враховується порядок слів і, як наслідок, нівелюються синтаксичні відношення, логіка та морфологія.

З огляду на сказане розроблено метод нечіткого семантичного порівняння за змістом розгорнутих відповідей студентів, поданих в електронному вигляді, з варіантами правильних відповідей. Розроблений алгоритм (див. рис. 3) передбачає автоматизоване виділення лексичних одиниць тексту з подальшим здійсненням морфологічного, синтаксичного, семантичного та прагматичного аналізу.

Для порівняння нечітких лексичних одиниць використано так звану метрику Левенштейна, що дозволяє встановлювати ступінь відповідності еталонного тексту з бази даних предметної галузі тексту відповіді.

Застосування розробленого алгоритму дозволяє усунути можливі помилки у вихідному тексті (неправильні закінчення, нестандартні скорочення тощо), визначити належність вихідного тексту до певної предметної галузі, сформувати загальну оцінку відповіді на питання тестових завдань на основі комплексного показника.

У розробленій системі оцінювання знань удосконалено алгоритми аналізу рядків. Кожний текстовий рядок – це вектор в N-вимірному просторі, де N — кількість символів у рядку. Для нечіткого порівняння текстової інформації у відповідях студентів під час тестування було удосконалено алгоритм, у якому зразок і відповідь розбиваються на окремі слова. Після цього проводиться нечіткий пошук збігу слів у зразку і відповіді, для чого застосовується метрика Левенштейна.

Удосконалення алгоритму методу ЛСА полягає в тому, що на етапі формування частотної матриці індексованих слів (терм) застосовано алгоритм нечіткого семантичного порівняння текстової інформації. У результаті його роботи індексовані слова (терми) замінюються лексичними одиницями з баз даних. Процедуру стемінгу було замінено на лематизацію текстових одиниць, тобто

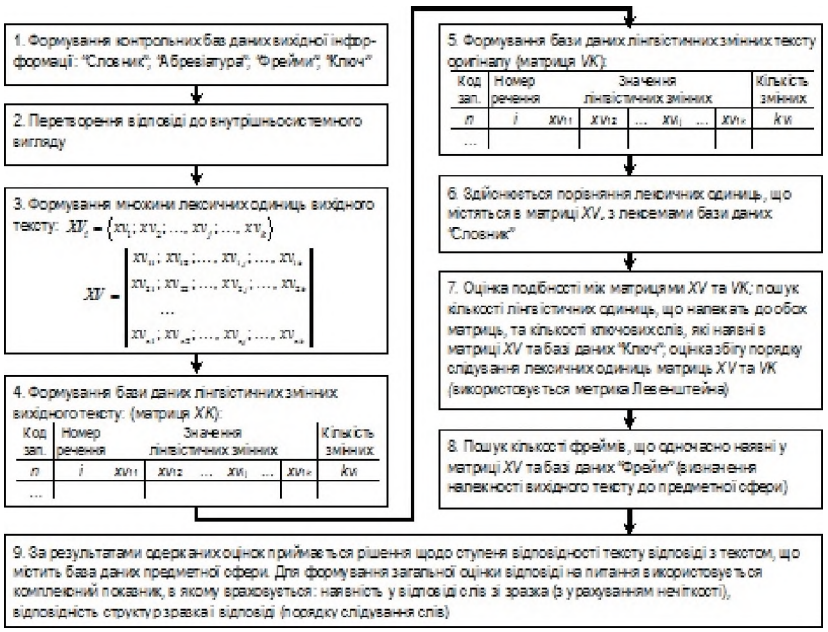

Рис. 3. Алгоритм методу нечіткого семантичного порівняння за змістом розгорнутих відповідей

процедуру зведення формальних варіантів слова в тексті до його певного усталеного інваріанта — леми, або канонічної (вихідної, словникової) форми слова. Вихідним для дії автоматичного лематизатора є текст, усім словам якого присвоєно коди граматичних класів та граматичних підкласів.

Цей підхід дозволив виявляти латентні асоціативно-семантичні залежності у множині документів; частково усувати омонімію, полісемію та синонімію; виправляти слова, що написані студентом з орфографічними та технічними помилками; ураховувати синтаксичні відношення, логіку побудови терм у контексті предметної сфери тощо. Це значно розширює наукове та прикладне значення вдосконаленого методу латентно-семантичного аналізу.

Для порівняння текстової інформації за змістом на етапах семантичного та прагматичного аналізу розроблено моделі штучного інтелекту. За результатами семантичного аналізу будується семантична мережа – структура для подання знань у вигляді вузлів, пов'язаних дугами (зв'язками). Під час прагматичного аналізу визначається належність відповіді до визначеної предметної галузі. Семантичний і прагматичний аналіз запропоновано проводити на основі використання нейромережі. На відміну від відомих методів семантичного й прагматичного аналізу, розроблені алгоритми на основі моделей штучного інтелекту дають можливість з більшою достовірністю автоматизовано проводити перевірку відповідей, поданих у довільній текстовій формі природною мовою. Незалежно від побудови речень, додаткових суджень, несуттєвих якісних характеристик, які можуть бути у відповіді та зразку, з них виділяється основний «зміст» у формі семантичної мережі. Порівняння двох семантичних мереж (тексту відповіді та зразка) дозволяє достовірно оцінити ступінь їх тотожності, що підвищує об'єктивність оцінки.

Розроблені моделі й алгоритми істотно підвищують ефективність і достовірність роботи системи тестування, яка може використовуватися для поточного, модульного, рейтингового та підсумкового контролю. На відміну від систем

оцінювання з використанням тестів, така система дозволяє оцінювати природномовні відповіді студентів, подані в довільній формі.

На основі описаних моделей і алгоритмів розроблено відповідні комп'ютерні інструментальні засоби, які забезпечують автоматизовану оцінку знань студентів у реальному часі.

Результати практичного застосування розроблених моделей, методів та засобів лінгвістичного аналізу тексту в Інституті інтелектуальної власності Національного університету «Одеська юридична академія», Хмельницькому кооперативному, торговельно-економічному інституті, Національній академії Державної прикордонної служби України ім. Богдана Хмельницького продемонстрували достатню ефективність при перевірці розгорнутих відповідей на питання відкритого типу.

O. I. KOMARNYTSKA

MODELS AND METHODS OF TEXT LINGUISTIC ANALYSIS IN KNOWLEDGE EVALUATION SYSTEMS

A functional structure of an intellectual system of linguistic analysis of a deployed text response utilizing models of artificial intelligence has been developed in this article. An algorithm of fuzzy semantic comparison of textual information - answers to questions submitted by a student in natural language, with options of correct answers, which formalizes description of linguistic structure of the study content and answers has been elaborated. In order to form a frequency matrix of the indexed words there has been improved the algorithm of fuzzy latent-semantic comparison of textual information.

Keywords: algorithm, lexical unit, method, model, semantics, text, frame, Artificial Intelligence.