21. Beowulf . – [електронний ресурс]. – режим доступу : http://northvegr.org/sagas%20annd%20epics/ epic%20poetry/ beowulf/index.html

22. Deirdire. – [електронний ресурс]. – режим доступу : www.dirdire.org.uk

23. Fáfnismál [електронний ресурс]. – режим доступу: http://norse.ulver.com/texts/fafnis.html

24. Hammerrfall: (DG) Destined For Glory // Renegade (2000). – Nuclear Blast, 2000. – [електронний ресурс]. – режим доступу: http:www.darklyrics.com/h/hammerfall.html

25. Havamál. – [електронний ресурс]. – Режим доступу : http://norse.ulver.com/edda/hava.html

26. Jordan R. Winter's Heart.- New York : Tom Doherty Associates, LLC, 2000. - 596 p.

27. Manowar Heart of steel (b) // Kings of metal. - Atlantic, 1998; Courage (b) // Louder Than Hell. - Geffen, 1996. – [електронний ресурс]. – режим доступу : http://www.darklyrics.com/m/manowar.html

28. Rammstein. Mein Herz brennt // Mutter, 2001. – [електронний ресурс]. – режим доступу : http://www.azlyrics.com/lyrics/rammstein/meinherzbrennt.html

29. Völuspá. – [електронний ресурс]. – режим доступу : http://norse.ulver.com/edda/voluspa.html

О. Комарницька
(Київ)

МОДЕЛЮВАННЯ ПРОЦЕДУР ЛІНГВІСТИЧНОГО АНАЛІЗУ ТЕКСТУ В ІНТЕЛЕКТУАЛЬНІЙ СИСТЕМІ ОЦІНЮВАННЯ ЗНАНЬ

У статті розглянуто особливості моделювання та функціонування лінгвістичної підсистеми інтелектуальної системи оцінювання знань. Було створено та застосовано формальну модель семантики для побудови ядра інформаційної технології семантичного аналізу тексту, що дозволяє покращити якість аналізу природномовних відповідей за рахунок детального аналізу багатозначності слів. Така модель передбачає здійснення морфологічного, семантичного, синтаксичного аналізу тексту із застосуванням удосконаленого методу латентно-семантичного аналізу.

***Ключові слова**: алгоритм, аналіз, латентно-семантичний аналіз, лінгвістична підсистема, метод, модель, морфологія, оцінювання, прагматика, природна мова, семантика, текст, штучний інтелект.*

В статье рассмотрены особенности моделирования и функционирования лингвистической подсистемы интеллектуальной системы оценивания знаний. Была создана и получила применение формальная модель семантики для построения ядра информационной технологии семантического анализа текста, которая позволяет улучшить качество анализа естественно-языковых ответов за счет детального анализа многозначности слов. Такая модель предполагает осуществление морфологического, семантического, синтаксического анализа текста с применением усовершенствованного метода латентно-семантического анализа.

***Ключевые слова:** алгоритм, анализ, латентно-семантический анализ, лингвистическая подсистема, метод, модель, морфология, оценивание, прагматика, естественный язык, семантика, текст, искусственный интеллект.*

The article reveals peculiarities of the modeling and functioning of the linguistic subsystem of the intellectual knowledge evaluation system. It has been created and applied a formal semantic model in order to build a core of the information technology of the text semantic analysis, which allows for improvement of the natural language answers analysis due to the detailed analysis of the multiple-meaning of words. This model provides performing of the morphologic, semantic and syntactic text analysis with applying improved latent-semantic analysis. 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. The algorithm provides automatic conversion of student's responses from a natural language into an intersystem form, the formation of lexical units of the text, followed by the implementation of morphologic, syntactic, semantic and pragmatic analysis. 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. The application of the proposed new and improved methods, models and algorithms provides the possibility to detect latent semantic associative dependences in the set of documents; partly withdraw the phenomenon of homonymy, polysemy and synonymy; correct words written by a student with spelling and technical mistakes; consider the order of words in documents and their meaning; logic of the term in the context of the subject area. The significance of the obtained results is represented by the possibility of the automated evaluation of the students' knowledge in real-time, consisting of the tasks, answers to which are given in the form of free-text-scale.

***Key words**: algorithm, analysis, latent-semantic analysis, linguistic subsystem, method, model, morphology, evaluation, pragmatics, natural language, semantics, text, artificial intelligence.*

У рамках Державної програми “Інформаційні та комунікаційні технології в освіті і науці” на 2006–2013 роки та проекту “Створення та впровадження програмних засобів пілотної системи поточного і підсумкового контролю знань студентів у вищих навчальних закладах” було проведено низку науково-дослідних робіт, спрямованих на створення нової лінгвістично-інформаційної технології контролю та оцінювання знань. Необхідність виконання цих робіт зумовлена тією обставиною, що в умовах інформаційного суспільства виникла нагальна суспільна потреба модернізації засобів освіти і серед них – засобів контролю та оцінювання знань.

Автоматизація систем контролю знань вимагає розв'язання низки наукових проблем, які виникають на тлі цього питання, зокрема, створення нових та удосконалення існуючих систем лінгвістичного аналізу природномовної відповіді, поданої в довільній формі та застосування у таких системах моделей штучного інтелекту. Взагалі, моделювання інтелектуальної діяльності людини з оброблення текстової інформації є дуже складною задачею. Її успішна автоматизація призвела б до значного підвищення ефективності самих комп'ютерів, даючи можливість людині спілкуватися з комп'ютером природною мовою.

Моделювання природної мови здійснюється на різних рівнях. Найбільш складними для моделювання є рівні, на яких ведеться робота із семантикою окремих одиниць і тексту в цілому. Нетривіальні зв'язки між структурою тексту та його значенням не дають можливості побудови навіть простих моделей опрацювання текстової інформації без урахування значень елементів, що складають текст. Велика кількість сучасних досліджень в сфері штучного інтелекту спрямована на розробку моделей семантики, які дозволять зробити якісний стрибок у семантичній інтерпретації текстів і поліпшити результати практичної роботи систем обробки текстової інформації.

Наведені положення обумовлюють актуальність досліджень, пов’язаних з подальшим розвитком та вивченням формалізації семантики природної мови. Дослідження, проведені в цій роботі, спрямовані на створення формальної моделі семантики та її застосування для побудови ядра інформаційної технології семантичного аналізу тексту, яка дозволить покращити якість аналізу текстів природної мови за рахунок детального аналізу багатозначності слів.

У різних комп'ютеризованих системах виникає необхідність обробки інформації, представленої природною мовою. В системах, що включають людину як свою органічну ланку, основною формою передання інформації є документи, що містять значну кількість текстової інформації. Комп'ютерне моделювання процесів обробки текстів дозволить автоматизувати багато видів інтелектуальної діяльності людини, розширити його можливості. Ефективність інтелектуальних систем оцінювання знань визначається їх здатністю обробляти інформацію неформалізовану або слабко формалізовану.

У системах тестового контролю для семантичного порівняння текстів відповіді та зразка застосовуються різні методи, одним з найефективніших з яких, на нашу думку, є метод латентно-семантичного аналізу (ЛСА) [3, 11]. Принцип дії цього методу полягає у визначенні ступеня подібності за змістом текстів на підставі оцінки кореляції між різними текстовими одиницями. Проте, метод ЛСА для цілей цієї роботи необхідно вдосконалити, оскільки він не враховує суттєву лінгвістичну інформацію (порядок слів у реченні, ключові слова, помилки, логіку та морфологію).

Роботу із удосконалення методу ЛСА у відзначених аспектах виконано в рамках науково-дослідної роботи [5], де розроблено інтелектуальну систему оцінювання знань, умінь та навиків студентів вищих навчальних закладів (ІСОЗ). Функціональна структура ІСОЗ містить такі модулі: базу даних (предмети, модулі, теми, навчальні групи), базу знань (предмети, модулі, темі), лінгвістичну підсистему – аналізатори морфології, синтаксису, семантики і прагматики, навчання; оцінювання.

Робота лінгвістичної системи базується на алгоритмі порівняння за змістом розгорнутих відповідей студентів, представлених в електронному вигляді, з варіантами правильних відповідей, представлених в XML-форматі. Цей алгоритм надалі використовується при здійсненні латентно-семантичного аналізу і забезпечує автоматизоване формування індексу лексичних одиниць тексту з подальшим здійсненням морфологічного, синтаксичного, семантичного та прагматичного аналізу, за результатами яких порівнюються представлені відповіді з варіантами правильних відповідей з бази знань (рис. 1).

У розробленій лінгвістичній підсистемі вперше запропоновано алгоритм семантичного порівняння нечіткої текстової інформації (відповідей на запитання, що подані студентом природною мовою в довільній формі, із варіантами правильних відповідей), в якому формалізовано

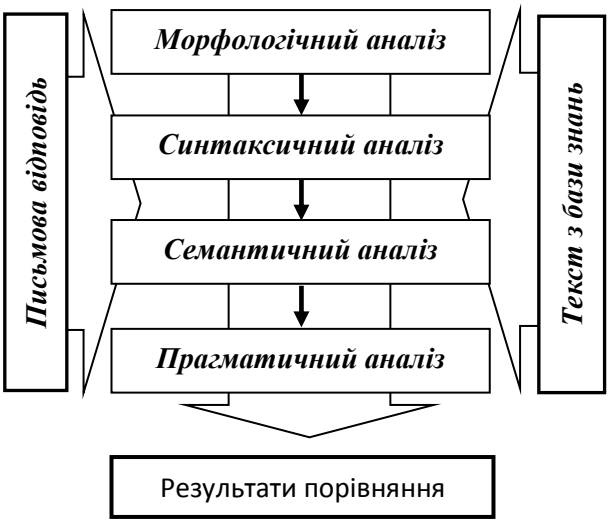

Рис. 1 – Схема обробки даних

опис лінгвістичної структури навчального контенту та відповіді. Розроблений алгоритм передбачає автоматичну конвертацію відповіді студента природною мовою до внутрішньо-системного вигляду, екстракцію лексичних одиниць тексту з подальшим здійсненням морфологічного, синтаксичного, семантичного та прагматичного аналізу. Застосування розробленого алгоритму дозволяє усувати помилки, що можуть бути у вихідному тексті (неправильні закінчення, нестандартні скорочення тощо), визначати належність вихідного тексту до певної предметної сфери, формувати загальну оцінку відповіді на питання за комплексним показником, у якому враховується присутність у відповіді слів, які є у зразку (у тому числі за умови нечіткості), відповідність структур зразка і відповіді (порядку слів). Також у лінгвістичній підсистемі удосконалено алгоритм методу латентно-семантичного аналізу [6], що передбачає на етапі формування частотної матриці індексованих слів застосування алгоритму нечіткого семантичного порівняння текстової інформації, внаслідок чого індексовані слова (терми) замінюються лексичними одиницями із баз даних, що містять перелік слів в усіх відмінках, перелік скорочень та значень абревіатур словосполучень, перелік ключових слів, котрі використовуються для опису процесів і явищ предметної сфери. Процедуру стемінга замінено лематизацією [7] на основі результатів автоматичного морфологічного аналізу текстів задля забезпечення більш високої якості роботи алгоритму. Застосовано алгоритми нечіткого пошуку, а саме удосконаленого варіанту метрики Левенштейна для виправлення некоректних слів. Запропоноване суттєво розширює прикладне та наукове значення удосконаленого методу латентно-семантичного аналізу.

Для якісної та повноцінної роботи система автоматичного лінгвістичного аналізу повинна мати можливість проаналізувати текст відповіді на запитання з позицій морфології, синтаксису, семантики та прагматики. Далі система повинна згенерувати логічне внутрішнє представлення та просинтезувати свою відповідь природною мовою.

Передумовою для здійснення усіх наступних етапів лінгвістичного аналізу тексту є процедура *графематичного аналізу*, що забезпечує виділення синтаксичних та структурних одиниць із вхідного тексту: абзаців, речень, окремих слів та розділових знаків.

Після здійснення графематичного аналізу інформація надходить до блоку *морфологічного аналізу*, завданням якого є нормалізація словоформ, отримання граматичної інформації і синтез словоформ. Морфологічний аналіз здійснюється шляхом розбиття всіх лексем на два класи: основозмінні та флективні класи слів. Змінювані слова відносять до певного класу за ознакою приналежності до однієї із синтаксичних груп та типом словозміни [2].

Вхідними даними процедури морфологічного розпізнавання є графемна структура тексту, отримана на попередньому етапі та еталонні моделі, які включають: словозмінну модель та словотвірну модель вхідної мови. Об'єктом розпізнавання є закономірності взаємодії морфем в межах мовної лексеми. Еталонні моделі складаються зі словника службових слів, словника морфем (квазізакінчень, суфіксів, префіксів тощо). Для одного вхідного слова може бути встановлено декілька основ і морфологічних параметрів. Ці дані перевіряються на відповідність інформації, що міститься в базі даних. Після такої перевірки слова вважаються правильними і надходять на вихід блоку морфологічного аналізу. Слова та абревіатури, що містять помилки, замінюються правильними словами, одержаними із бази даних “Словник”.

Ця послідовність потрапляє далі на вхід блоку *синтаксичного аналізу*, метою якого є отримання синтаксичної структури фрази, яка записується у вигляді дерева складових або дерева залежностей. У разі використання дерева залежностей для кожного елементу-вершини аналізованого ланцюжка вказується елемент, що ним керує, і тип зв'язку між ними (окрім джерела-вершини графа).

Природною мовою ту ж саму думку можна виразити різними фразами. Через це структура текстового подання відповіді може суттєво відрізнятись від зразка. Отже, для

порівняння за змістом текстової відповіді зі зразком потрібно виділити зміст. Вирішити це завдання можна за допомогою *семантичного аналізу* – виділення з довільного тексту природною мовою змістовної структури (знання) із застосуванням удосконаленого методу латентно-семантичного аналізу.

Передбачено, що порівняння поданої відповіді і зразка проводитиметься за декількома ознаками: за кількістю слів, кількістю ключових слів і фреймів, порядком слів, значимістю слів і фреймів. Під час *прагматичного аналізу* визначається належність відповіді до визначеної предметної сфери.

Застосування запропонованих нових та удосконалених методів, моделей та алгоритмів аналізу тексту надає можливість виявляти латентні асоціативно-семантичні залежності у множині документів; частково знімати явища омонімії, полісемії та синонімії; виправляти слова, написані студентом із помилками; враховувати синтаксичні відношення; логіку побудови терм у контексті предметної сфери тощо.

Описані моделі і методи обробки природномовної відповіді в ІСОЗ дозволяють проводити автоматизовану оцінку знань студентів у реальному часі із використанням таких завдань: завдання, що передбачають коротку вільну відповідь у вигляді числа або одного слова; завдання, що передбачають точну відповідь у вигляді правила, визначення, теореми тощо; завдання, що передбачають логічну відповідь (вибір із множини, впорядкування за ознакою, доведення тощо); завдання відкритого типу, відповідь на які повинні бути подані в довільній формі природною мовою – математичних викладень або тексту.

Предметом подальших досліджень є удосконалення моделей та методів синтаксичного аналізу тексту інтелектуальної системи оцінювання знань.

Список літератури

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / [Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А. и др.]. – М. : МИЭМ, 2011. – 272 с.

2. Дарчук Н. П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту) : підручник / Наталія Петрівна Дарчук. – К. : Видавничо-поліграфічний центр "Київський університет", 2008. – 351 с.

3. Заболеева-Зотова, А. В.Латентный семантический анализ : новые решения в Internet / А. В. Заболеева-Зотова, А. Ю. Пастухов, П. В. Сердюков, Н. А. Козлова, С. А. Чернов // Информационные технологии. – 2001, № 6. – С. 67–82.

4. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений / Л. Заде; пер с англ. – М. : Мир, 1976. – 165 с.

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

6. Комарницкая О. И. Совершенствование алгоритма латентно-семантического анализа нечеткой текстовой информации / Современный научный вестник. № 29 (225). Серия : Филологические науки. – Белгород : Руснаучкнига. – 2014. – С. 58–62.

7. Олексієнко Л., Дарчук Н. Лематизація парадигм іменників української мови // Проблеми українізації комп'ютерів. – К., 1993. – С.62–65.

8. Марченко О. О. Алгоритми семантичного аналізу природномовних текстів : Дис. канд. фіз.- мат. наук : 01.05.01 / КНУ ім. Тараса Шевченка. – К., 2005. – 150 с.

9. Поспелов Г.С, Поспелов Д.А. Искусственный интеллект и прикладные системы. – М. : Знание, 1985. – 43 с.

10. Штангей С.В. Моделі і інформаційні технології контролю знань в системі дистанційного навчання. – Рукопис. Дис. к. техн. н. : 05.13.06 – інформаційні технології. – Харківський національний університет радіоелектроніки, Харків, 2009. – С. 165.

11. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.A. 1990.Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41. – P. 391-407.

Т. Коропатніцька
(Чернівці)

СЕМАНТИЧНЕ ГРАДУЮВАННЯ ПОРІВНЯННЯ У ДИСКУРСІ

У роботі розглядаються шляхи включення категорії градуювання як мовної універсалії до якісно-кількісної характеристики прототипічних засобів вербалізації порівняння в німецькій мові.