ВИКОРИСТАННЯ МЕТОДУ АВТОМАТИЧНОЇ ЕКСТРАКЦІЇ ВІДНОШЕНЬ СЕМАНТИЧНОЇ БЛИЗЬКОСТІ ДЛЯ РОЗРОБКИ БАЗ ЗНАНЬ

Петрасова С. В.
Національний технічний університет
«Харківський політехнічний інститут»,
м. Харків, вул. Пушкінська 79/2, тел. 707-63-60,
e-mail: svetapetrasova@gmail.com

Сучасний стан розробок баз знань та їх систематизації потребує впровадження універсальних інтелектуальних систем, важливим завданням яких є екстракція інформації з текстів та представлення її у вигляді формальної системи знань.

Найбільш перспективним способом формального вираження знань сьогодні є семантична мережа. Це обумовлено, передусім, наочністю представлення знань та можливістю в явному вигляді виражати семантичні відношення між поняттями [1].

Але розповсюдження даного способу представлення знань стримується як неоднозначністю вираження знань на природній мові, так і трудомісткістю та складністю розробки семантичної мережі.

В такому разі одним з найбільш повних джерел знань для автоматичної побудови бази знань можуть слугувати такі універсальні засоби представлення, накопичення та передачі інформації, як тексти. Серед усіх текстових джерел саме глосарії представляють тексти природної мови з найбільш концентрованим смисловим навантаженням.

В даному дослідженні пропонується використання методу автоматичної екстракції відношень семантично близьких понять для розробки семантичних мереж, який ґрунтується на знаннях глосарія, виражених дефініціями термінів даних об’єктів [2].

Для побудови логічної схеми виявлення семантично близьких термінів вводиться метричний простір лінгвістичних смислових одиниць Θ, який визначається як множина лінгвістичних одиниць лексикону T, на якому граматичні правила задають відношення між одиницями, що виступають обмеженнями для коректних синтаксичних структур [3].

Міру семантичної близькості f формально визначимо співвідношенням через відповідні дефініції глосаріїв d_1 та d_2 як потужності множин, утворених теоретико-множинним перетином та об’єднанням множин термінів дефініцій:

$$f(t',t'') = \frac{2\,|\,d_1 \bigcap d_2\,|}{|\,d_1\,| + |\,d_2\,|},$$

де $d_1 \bigcap d_2$ – спільні терміни дефініцій, а $|d_1|+|d_2|$ – всі терміни дефініцій d_1 и d_2.

В створеному просторі концептів з одним і тим самим сигніфікативним смислом можна виявити такі категорії семантичних відношень, як приналежність до класу, гіперонімія, гіпонімія та меронімія. Для формалізації описаних типів відношень застосовуються шаблони лексичних послідовностей:

$$NN_1 \rightarrow \mathrm{Re}\, l_z \rightarrow NN_2,$$

де NN_1 и NN_2 – зв'язані концепти, представленні ключовими словами та словосполученнями глосаріїв, $\mathrm{Re}\, l$ – лексичні ланцюжки, які виражають відношення z (табл. 1).

Таблиця 1 – Приклади шаблонів лексичних послідовностей

Семантичне відношення, z	**Лексичні ланцюжки,** $\mathrm{Re}\, l$
Приналежність до класу	"є", "вважається"
Гіперонімія	"сукупність", "комплекс", "набір", "сімейство"
Гіпонімія	"наприклад", "тип", "(різно)вид", "екземпляр"
Меронімія	"частина", "елемент"

Таким чином, неоднозначність тлумачення та представлення природної мови є характерною особливістю текстових ресурсів, що не дозволяє однозначно формалізувати виявлення семантичних відношень з текстів. Для вирішення даної проблеми розглянуто метод автоматичної екстракції відношень семантичної близькості, який основується на використанні глосарія як природномовного тексту, що найбільш повно концентрує знання.

Список літератури

1. *Manning C. D.* Introduction to Information Retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. – Cambridge University Press, England, 2009. – 544 p.

2. *Кобозева И. М.* Лингвистическая семантика: Учебное пособие. / И. М. Кобозева. – М. : Эдиториал УРСС, 2000. – 352 с.

3. *Хайрова Н. Ф.* Определение семантической близости на основе когнитивного подхода. / Н. Ф. Хайрова, Н. В. Шаронова, Н. В. Борисова // Бионика интеллекта: науч.-техн. журнал, 2013.