ЛІНГВО-СТАТИСТИЧНИЙ АНАЛІЗ ПОБУДОВИ ТЕЗАУРУСУ ТЕРМІНОЛОГІЇ ДЛЯ ПРЕДМЕТНОЇ ОБЛАСТІ

Товпинець В.П., Савченко М.В.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Під тезаурусом розуміють словник концептів з певною структурою зберігання даних і набором семантичних відносин, що вказують на спільність або протиставлення значень лексичних одиниць. На противагу словникам, тезаурус – це система представлення даних, яка дозволяє використовувати його не тільки як засіб для відображення інформації в зрозумілому людині вигляді, але і для подальшої роботи з ним, як з джерелом знань для задач, пов'язаних з комп'ютерною лінгвістикою, інформаційним пошуком та системами штучного інтелекту.

Наявність предметно-орієнтованого тезауруса дозволяє значно спростити процес збору, формалізації, зберігання, оцінки та використання знань, а за рахунок семантичної мережі між концептами можливий автоматичний аналіз текстів і ряд інших завдань, що сприяє підвищенню ефективності роботи фахівця або робочої групи обраної предметної області.

Для створення тезауруса предметних областей необхідно об'єднання зусиль цілих груп відповідних фахівців і експертів для обробки великого числа об'ємних джерел інформації: словників, довідників, наукових публікацій та інших текстів. В даний час для цього практикується підхід, заснований на комбінування як ручних, так і автоматичних методів на основі сучасних інформаційних технологій.

В роботі реалізована методологія автоматичного складанням тезауруса на основі термінології предметної області, заснованої на вилученні інформації з визначень термінів, що розглядаються. Семантична близькість термінів оцінюється за допомогою двох метрик близькості і ручної оцінки експерта предметної області.

Для автоматичного вилучення були розглянуті такі відношення: рід-вид, ціле-частина, синонімія. Автоматичне розпізнавання семантичних відносин можливо за допомогою лексико-синтаксичних шаблонів. Метод для вилучення родо-видових відносин полягає у використанні статей Вікіпедії для реалізації методів машинного навчання, заснованих на алгоритмах найближчих і взаємних найближчих сусідів і двох метриках семантичної близькості слів. Результати вилучення використовуються в системі Serelex для пошуку семантично пов'язаних слів.

Для вилучення родо-видових відносин використаний підхід на основі використання визначень тлумачних словників з накладенням ряду правил для розпізнавання відносин між визначеним словом в якості роду і одним з слів дефініції в якості виду.