УДК 004.8

М.В. Шклярук, О.К. Карнаухов
Тернопільський національний технічний університет імені Івана Пулюя, Україна

АВТОМАТИЗОВАНЕ ОПРАЦЮВАННЯ ТЕКСТІВ УКРАЇНСЬКОЮ МОВОЮ

M. V. Shklyaruk, O.K. Karnaukhov

AUTOMATED PROCESSING TEXTS IN UKRAINIAN

Стрімкий розвиток інформаційних технологій та збільшення інформаційного потоку потребує створення засобів автоматизованого опрацювання інформації. Одним з ключових питань в цьому напрямі є машинне опрацювання інформаційних об'єктів, які подаються неформально та є близькими до мови спілкування людей.

Математична лінгвістика (комп'ютерна лінгвістика) – математична дисципліна, що розробляє формальний апарат для опису природних і штучних мов. Розділами комп'ютерної лінгвістики є розпізнавання та синтез мови, синтаксичний аналіз та генерація, автоматичне реферування.

В розробці систем автоматизованого опрацювання текстів найважливішим є уміння побудувати лінгвістичний алгоритм аналізу мовного явища в тексті, а саме – задати комп'ютеру певні формальні ознаки мовних одиниць та їх сполук. Доцільно проводити одноразове повне опрацювання тексту, яке дає практичні результати і виходи в роботу різних задач комп'ютерної лінгвістики. Розв'язання даних завдань неможливе без існування баз даних та баз знань.

Основою лінгвістичної бази даних є автоматичний словник, який зберігає всю необхідну інформацію для реалізації алгоритмів. Тип автоматичного словника для конкретної бази даних зазвичай визначається приналежністю мови словника до аналітичного або синтетичного типу. Українська мова належить до синтетичних мов, тому для опрацювання текстів доцільно використовувати автоматичний словник основ. У загальному випадку словникова стаття – уся інформація про дану лінгвістичну одиницю – складається з номера основи, самої основи, ланцюжка граматичних, семантико-синтаксичних та семантичних характеристик основи. Обов'язковою складовою автоматичного словника є словник зворотів, які не можуть розглядатись послівно, інколи також словник семантико-синтаксичних фреймів.

Прикладом лексико-семантичної бази знань української мови, що створена на базі СУБД, є проект UWN. Він використовує такі структурні елементи як синсети (набори синонімів, що описують єдине поняття) та набори семантичних і лексичних зв'язків. Проект є однією з перших спроб створити універсальну україномовну онтологію – специфічну базу знань, що містить інформацію трьох типів: об'єкти, властивості, дії. Зважаючи на те, що він є важливою розробкою в галузі комп'ютерної лінгвістики, UWN є також вкладом в таку галузь як штучний інтелект, адже комп'ютерна лінгвістика є напрямом саме цієї області науки.

Отже, створення засобів опрацювання словесної інформації українською мовою є важливим завданням науковців України і можливістю проводити дослідження в напрямі штучного інтелекту, який є достатньо молодою, але перспективною галуззю науки.

Література

1. Комп'ютерна лінгвістика (автоматичне опрацювання тексту) : підручник. Н.П.Дарчук. К.: Видавничо-поліграфічний центр «Київський університет», 2008. – 351 с.

2. Українська Лінгвістична Лабораторія [Електронний ресурс] – Режим доступу: URL: http://www.lingvoworks.org.ua/index.php. – Назва з екрану.