АВТОМАТИЧНИЙ ПОСТЕГІНГ КОРПУСУ ТЕКСТІВ

Оліфенко І.В., Борисова Н.В.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Необхідність обробки великих обсягів інформації викликає необхідність створення інформаційних систем автоматизованої або автоматичної обробки текстової інформації. Одним з видів таких систем є системи здатні до самонавчання, які, як правило, при первинному навчанні використовують корпуси текстів – набір текстів відібраний за визначеним критерієм та певним чином розмічений. Лінгвістична розмітка корпусів (англ. tagging, annotation) – це процес або результат приписування текстам з корпусу та/або їх компонентам спеціальних міток, що дає можливість ідентифікувати тексти за різними параметрами. Занадто детальна лінгвістична інформація, представлена в розмітці, якою забезпечуються великі корпуси текстів, може бути надлишковою і вимагати витрат невиправдано великої кількості часу та зусиль, в той час як цілі дослідження потребують мінімізувати трудовитрати, обмежившись лише необхідним у даному дослідженні набором міток.

Важливою частиною лінгвістичної розмітки корпусу є так званий постегінг (англ. Part of Speech (POS) tagging) – етап автоматичної обробки тексту, завданням якого є визначення частини мови і граматичних характеристик слів в текстах корпусу з приписуванням їм відповідних тегів.

При обробці текстів німецькою мовою проблеми полягають, по-перше, у недостатній кількості корпусів, а, по-друге, у надлишковій їх розмітці. Для вирішення цих проблем, а також на основі аналізу існуючих корпусів німецької мови та систем тегів розмітки було створено власний корпус, розроблено алгоритм автоматичної розмітки дієслів німецької мови та програму постегінга на мові програмування Python. Ця мова надає розробникам значний арсенал засобів обробки текстової інформації. Розроблений корпус містить тексти новин з трьох найпопулярніших новинних порталів Німеччини. Оскільки у корпусі представлені тексти сучасної німецької мови, то на ньому можна аналізувати сучасні тенденції мови та виявляти закономірності в залежності від мети дослідження. Що стосується розмітки корпусів, то на сьогоднішній день постегінг у німецькій мові реалізовано за допомогою таких інструментів як TreeTagger, OpenNLP Part-of-Speech Tags, Open Xerox, TagAnt, CQP-web та ін. Деякі з них мають також і власну систему розмітки слів, але зазвичай вони схожі та не мають явних відмінностей. Для реалізації власної автоматичної морфологічної розмітки дієслів були використані теги, розміщені на ресурсі Open Xerox. Застосування граматичних правил німецької мови дозволило здійснювати автоматичний постегінг дієслів, надаючи їм лише необхідну морфологічну інформацію. Отримані результати автоматичної розмітки було порівняно з розміткою існуючих корпусів німецької мови. Порівняльний аналіз показав, що розроблена програма постегінга правильно розмічає від 90 до 95% німецьких дієслів. Це свідчить про високу ефективність її роботи.