ЗАСТОСУВАННЯ СТАТИСТИЧНОГО МЕТОДУ АВТОМАТИЧНОГО ВИЯВЛЕННЯ КОЛОКАЦІЙ У РОЗРОБЛЕНОМУ КОРПУСІ ТЕКСТІВ

Петрасова С. В., Перевало Я. О.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

У наш час одним із головних інструментів для вирішення лінгвістичних завдань стають масштабні розмічені корпуси (наприклад, British National Corpus, Cobuild Project тощо). Однак, у разі відсутності доступу до лінгвістичних «гігантів» або необхідності вивчати ті тексти, що не увійшли до відомих корпусів, виникає завдання скласти свій власний корпус і проводити дослідження вже на основі нового текстового масиву.

У рамках дослідження проблеми сполучуваності слів (виявлення колокацій) було створено власний лінгвістичний корпус, який містить анотації до англомовних науково-технічних статей за напрямом «Artificial Intelligence». Усі тексти вибрані з ScienceDirect – провідного джерела наукових досліджень, що на сьогоднішній день налічує понад 2 500 наукових журналів, а також 26 000 електронних книг.

Для автоматичного виявлення двослівних колокацій у текстовому середовищі було розроблено обчислювальну програму (модуль корпусного менеджера) та імплементовано статистичну міру MI (Mutual Information). Сутністю розрахунків за мірою MI є порівняння залежних контекстно-пов’язаних частот з незалежними (при випадковій появі слів в контексті). Якщо значення MI більше 1, тоді поєднання слів вважається статистично значимим.

Аналіз створеного спеціалізованого корпусу показав, що найчастіше уживаними колокаціями стали наступні: artificial intelligence, this paper / article / study, neural network(s), machine learning, deep learning та інші.

Слід зазначити, що автоматичний аналіз тексту за допомогою описаного вище статистичного апарату є тільки початковим етапом при виявленні колокацій. У подальшому необхідна експертна оцінка отриманих результатів, у тому числі, із залученням даних із словників (в першу чергу, тлумачних і словників сполучуваності).

Результати дослідження можуть бути використані при створенні пошукових систем, систем машинного перекладу, укладанні словників, тезаурусів та вирішенні інших завдань прикладної лінгвістики.

Літератури:

1 Петрасова С. В. Автоматичне видобування колокацій з корпусу текстів / С. В. Петрасова, М. О. Кузьміна // Вісник НТУ «ХПІ». Серія : Актуальні проблеми розвитку українського суспільства. – Харків : НТУ «ХПІ», 2018. – № 4 (1280). – С. 68–72.