УДК 004.891

В.Н. ВРУБЛЕВСЬКИЙ
Київський національний університет імені Тараса Шевченка, Київ, Україна,
e-mail: *vitalii.vrublevskyi@gmail.com.*

О.О. МАРЧЕНКО
Київський національний університет імені Тараса Шевченка, Київ, Україна,
e-mail: *omarchenko@univ.kiev.ua.*

РОЗРОБЛЕННЯ ТА ДОСЛІДЖЕННЯ МОДЕЛІ ПРЕДСТАВЛЕННЯ СЕМАНТИКИ РЕЧЕНЬ

Анотація. Наведено огляд ефективної та простої моделі представлення семантики речень у контексті задачі ідентифікації парафразів. Дерево залежностей обрано як основну структуру для представлення зв'язків між словами у реченні. Для представлення семантики слова використано попередньо навчені моделі представлення слів. На основі цих двох ключових складових розроблено декілька ознак, які допомагають точно визначити парафрази. Проведені експерименти довели, що модель є ефективною. Результати її застосування є відносно близькими до результатів найсучасніших моделей.

Ключові слова: оброблення природної мови, ідентифікація парафразів, семантична подібність, дерево залежностей, векторне представлення слів.

ВСТУП

Нині побудова моделей представлення семантики слів, речень та текстів природної мови справедливо посідає центральне місце в галузі комп'ютерної лінгвістики та штучного інтелекту загалом. Широку популярність мають такі моделі, як BERT, RoBERTa, AlBERT, GPT-2, GPT-3. Їх створили та обчислили найбільші світові IT-компанії з використанням надпотужних ресурсів своїх дата-центрів. Ці багатовимірні векторні моделі демонструють найкращі результати рівня state-of-the-art під час розв'язання переважної більшості задач комп'ютерної лінгвістики, залишаючи конкурентів далеко позаду.

Починаючи з моделей представлення семантики слів, дослідники намагалися закодувати у векторі слова інформацію про його контекст, наприклад, k сусідніх слова зліва та k сусідніх слова справа, як у моделях CBOW (continuous bag of words) та Skip-gram [1]. Пізніше, під час моделювання семантики речень у тексті за моделлю Skip-thought [2], текст представляли як послідовність речень подібно до представлення за моделями, зазначеними вище. Це уявлення є занадто спрощеним. У ньому не взято до уваги розмаїття різних нелінійних складних зв'язків між словами всередині речення, а також зовнішніх зв'язків між реченнями в межах тексту. Структура речення не є лінійною послідовністю слів, а має радше структуру дерев з елементами рекурсії. Ігнорування цих реалій мови може не спричиняти значних негативних наслідків, коли дослідники працюють з аналітичними мовами (наприклад, з англійською), де є чітко фіксований порядок слів у реченні. У випадку синтетичних мов, де порядок слів у реченні може вільно змінюватися, нехтування справжньою синтаксичною структурою може призвести до значного зниження ефективності моделі.

Головним завданням цієї роботи є дослідження ефективності використання синтаксичної структури речення як ключової ознаки під час побудови моделей представлення семантики речень природної мови. Для експериментального дослідження ефективності побудованих моделей семантики речень взято класичну задачу комп'ютерної лінгвістики — визначення парафразування.