КІБЕРНЕТИКА

УДК 004.89

О.Г. СКУРЖАНСЬКИЙ
Київський національний університет імені Тараса Шевченка, Київ, Україна,
e-mail: *oleksandr.skurzhanskyi@gmail.com.*

О.О. МАРЧЕНКО
Київський національний університет імені Тараса Шевченка, Київ, Україна,
e-mail: *rozenkrans17@gmail.com.*

А.В. АНІСІМОВ
Київський національний університет імені Тараса Шевченка, Київ, Україна,
e-mail: *avatatan@gmail.com.*

СПЕЦІАЛІЗОВАНЕ ПОПЕРЕДНЄ НАВЧАННЯ НЕЙРОМЕРЕЖЕВИХ МОДЕЛЕЙ НА СИНТЕТИЧНИХ ДАНИХ ДЛЯ ПОКРАЩЕННЯ ГЕНЕРАЦІЇ ПЕРЕФРАЗУВАННЯ

Анотація. Генерація перефразувань є фундаментальною проблемою в галузі обробки природних мов. Завдяки значному успіху технології перенесення навчання підхід «попереднє навчання → точне налаштування» став стандартним. Однак популярні універсальні методики попереднього навчання зазвичай потребують величезних наборів даних та значних обчислювальних потужностей, а доступні навчені моделі обмежені фіксованою архітектурою та розміром. Запропоновано простий та ефективний підхід до попереднього навчання спеціально для генерації перефразувань, який помітно підвищує якість генерації перефразувань та забезпечує суттєве покращення моделей загального призначення. Використано як наявні публічні дані, так і нові, згенеровані великими мовними моделями. Досліджено, як ця процедура попереднього навчання впливає на нейронні мережі різної архітектури, та доведено, що вона працює ефективно для всіх архітектур.

Ключові слова: штучний інтелект, машинне навчання, нейронні мережі, генерація перефразування, попереднє навчання, точне налаштування.

ВСТУП

Задача генерації перефразувань є однією з найпопулярніших та найскладніших у галузі обробки природних мов. По-перше, ця задача є спеціальним випадком генерації тексту. Є багато моделей генерації тексту, які можна застосувати до задачі генерації перефразувань. По-друге, остання по суті є аналогом машинного перекладу. Єдина відмінність полягає у тому, що речення має бути «перекладене» на ту саму мову, але з використанням інших слів. Тому до цієї задачі можна безпосередньо застосовувати не лише моделі машинного перекладу, а й метрики якості машинного перекладу, які часто підходять для оцінювання систем перефразування.

Особливістю задачі генерації перефразувань, на відміну від інших задач обробки природних мов, є велика кількість публікацій, які не використовують анотовані дані, а оперують лише звичайними текстовими корпусами. Зауважимо, що вхід та вихід для цієї задачі є взаємозамінними: якщо з речення $x_1, x_2, \ldots, x_n$ можна з високою ймовірністю отримати речення $y_1, y_2, \ldots, y_k$, то