АНАЛІЗ БАГАТОЦІЛЬОВИХ МОДЕЛЕЙ NLP ДЛЯ ВИЗНАЧЕННЯ СЕМАНТИЧНОЇ ПОДІБНОСТІ ТЕКСТІВ

Олізаренко С.А.[1], д.т.н., снс; Волков А.Ф.[2];
Галузінський А.Г.[2], Свирідов А.С.[1]
[1]Харківський національний університет радіоелектроніки, м. Харків
[2]Харківський національний університет Повітряних Сил імені Івана Кожедуба, м. Харків

У доповіді наведені результати проведеного аналізу сучасних багатоцільових моделей NLP та надані рекомендації щодо їх використання для визначення семантичної подібності текстів в інформаційних пошукових системах (ІПС).

Семантична подібність текстів – фактор, який визначає найбільший вплив на процес видачі відповідної інформації з використанням ІПС. Перші ІПС для визначення семантичної подібності текстів враховували тільки кількість присутніх в ньому ключових фраз, що в точності відповідають запиту. Причому, при збільшенні числа повторень підвищувалася позиція відповідного тексту. Це приводило до того, що часто зустрічалися тексти, які фактично були семантично мало подібні, тобто не корисні для користувачів.

На даний час при визначенні семантичної подібності текстів ІПС застосовують складні алгоритми і враховують велику кількість чинників. При цьому, одним з найперспективніших підходів вважається підхід з використанням багатоцільових моделей Natural Language Processing (NLP) на основі глибоких нейронних мереж.

Багатоцільові моделі – це моделі які підтримують різноманітні задачі NLP (машинний переклад, системи відповіді на питання, чат-боти, аналіз настроїв і т.д.). Основною концепцією багатоцільових моделей NLP є концепція мовного моделювання з використанням попередньо навчених глибоких нейронних мереж. В рамках дослідженн я проаналізовані основні сучасні глибокі нейромережеві моделі відповідного класу (Universal Language Model Fine-tuning (ULMFiT), Transformer, Transformer-XL, BERT (Bidirectional Encoder Representations From Transformers), OpenAI GPT та ін.).

За результатами аналізу у якості базової багатоцільової моделі NLP для визначення семантичної подібності текстів в ІПС пропонується використовувати багатоцільову глибоку нейромережеву модель BERT. При цьому, виконується попереднє тонке налаштування багатоцільової моделі BERT для визначення семантичної подібності текстів.

Література:

1. Devlin J., Chang M.-W., Lee K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2 [cs.CL] 24 May 2019.

2. S. Olizarenko, V. Argunov. Research into the possibilities of the multilingual BERT model for determining semantic similarities of news content (2019). / https://hipsto.global/BERT-Application-Research-for-HIPSTO-Related-News-Detection.pdf.