УДК 004.934:811.161.2 DOI: https://doi.org/10.30837/IYF.IIS.2024.052

ДОСЛІДЖЕННЯ ПІДХОДІВ ДО СУМАРИЗАЦІЇ ДЛЯ УКРАЇНСЬКИХ ТЕКСТІВ

Кардаш Д. М.
Науковий керівник –доцент Турута О. П.
Харківський національний університет радіоелектроніки, каф. ШІ,
м. Харків, Україна
e-mail: dmytro.kardash@nure.ua

The problem of textual data summarization in the Ukrainian language in the context of computational linguistics is studied. The aim of the study is to evaluate and compare different approaches to language modeling for the Ukrainian language. The work includes creating a data corpus, analyzing existing methods, conducting experiments with different language units, and evaluating the resulting models. In particular, the paper discusses the main areas of linguistics, such as phonology, morphology, syntax, semantics, and pragmatics, which form the theoretical basis for computer text processing.

У сучасному світі зростає значення обробки текстової інформації, особливо українською мовою, яка є офіційною мовою великої кількості людей. Проте, однією з основних проблем у цьому контексті є нестача належної кількості даних для ефективного навчання моделей обробки природної мови. У зв'язку з цим, метою цього дослідження є аналіз підходів до сумаризації текстових даних українською мовою з метою покращення ефективності моделей обробки природної мови.

Гіпотеза цього дослідження полягає в тому, що різноманітні методи сумаризації тексту можуть сприяти збільшенню якісної обробки великої кількості текстових даних українською мовою. Основна задача полягає у вивченні та порівнянні різних підходів до автоматичної генерації кратких, інформативних анотацій текстів українською мовою.

Умови дослідження передбачають створення корпусу текстових даних, складених з новинних статей або наукових публікацій на українській мові. Далі, проводиться аналіз існуючих методів сумаризації тексту та їхній експериментальний перевірка на вибраному корпусі. Зокрема, оцінюється якість та точність отриманих кратких анотацій за допомогою об'єктивних метрик, таких як ROUGE (Recall-Oriented Understudy for Gisting Evaluation).

Ключові знахідки, результати дослідження:
Складність сумаризації тексту полягає у врахуванні ключових інформаційних елементів тексту, відсіюванні надлишкової чи неважливої інформації та генерації змістовних та лаконічних анотацій. Особлива увага приділяється збереженню смислової повноти та структури оригінального тексту в анотаціях. Деякі складнощі виникають у визначенні критеріїв

важливості різних частин тексту та їхнього подальшого використання для створення коротких анотацій. Також важливо враховувати специфіку української мови, такі як мовні вирази, фразеологізми та відмінність у синтаксичних конструкціях, що може ускладнити процес сумаризації.

Необхідно проаналізувати Transformer модель, для забезпечення вирішення проблеми стандартних підходів, використовуючи архітектуру, повністю побудовану на механізмі внутрішньої уваги (self-attention).

Оскільки ця модель не використовує рекурентні мережі, які можуть запам'ятати, як послідовності слів використовуються в моделі, виникає потреба присвоїти кожному слову відносне положення, оскільки послідовність залежить від порядку її елементів. Ці позиції додаються до вбудованого подання (n-вимірного вектора) кожного слова. Функцію внутрішньої уваги при цьому можна описати як відображення входу і набору пар ключ–значення на вихід, де запит, ключі, значення і вихідні дані є векторами. Вихідні дані обчислюються як зважена сума значень, де вагу, присвоєну кожному значенню, обчислює функція сумісності запиту з відповідним ключем. Загалом було визначено основні переваги моделі: покращене моделювання далеких залежностей слів; на відміну від минулих підходів на основі RNN, можлива паралелізація методів.

А також її основні недоліки в ході дослідження: складність перенавчання seq2seq моделей, оскільки архітектура складається з двох частин; механізм уваги працює з рядками фіксованої довжини, як наслідок вимагає розбиття, що в деяких випадках веде до втрати контекстної інформації.

Застосування методів сумаризації даних є важливим етапом у покращенні моделей обробки природної мови для української мови. Результати дослідження свідчать про те, що збільшення обсягу та різноманітності навчальних даних сприяє покращенню якості моделей, що використовуються в різних мовних завданнях. Це може мати велике значення для подальшого розвитку прикладних систем обробки текстів українською мовою.

Список використаних джерел:

1. Айсіна Р. М. Огляд засобів візуалізації тематичних моделей колекцій текстових документів // Машинне навчання та аналіз даних (http://jmlda.org). – 2015. – Т. 1, № 11. – С. 1584-1618.

2. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning [Електронний ресурс] / [O. Turuta, E. Erdem, M. Kuyu та ін.]. – 2022. – Режим доступу до ресурсу: https://scholar.google.com.ua/citations?view_op=view_citation&hl=ru&user=_0Gh01QAAAAJ&citation_for_view=_0Gh01QAAAAJ:p2g8aNsByqUC.