УДК 621.38:[621.38-025.53] DOI: https://doi.org/10.30837/IYF.IIS.2024.026

ДОСЛІДЖЕННЯ МЕТОДІВ АВТОМАТИЧНОГО РЕЗЮМУВАННЯ ТЕКСТІВ НОВИН

Смолярчук С.В.
Науковий керівник – доц. каф. ПЗЕОМ Турута О. П.
Харківський національний університет радіоелектроніки, каф. ШІ,
м. Харків, Україна
e-mail: serhii.smoliarchuk@nure.ua

The study delves into the challenge of summarization textual data in the Ukrainian language within the domain of computational linguistics, focusing on news summarization. Its primary objective is to assess and compare various approaches to language modeling tailored to the Ukrainian language in the context of news articles. The research entails the creation of a corpus of news articles, analysis of existing summarization methods, experimentation with different summarization techniques, and evaluation of the resulting models. Furthermore, the paper explores key linguistic areas such as phonology, morphology, syntax, semantics, and pragmatics, which serve as the theoretical foundation for effectively summarizing news articles computationally.

В сучасному інформаційному середовищі значущим є процес обробки текстової інформації, особливо українською мовою, яка є офіційною для великої аудиторії. Однак, виникає проблема нестачі належного обсягу даних для ефективного навчання моделей обробки природної мови. У цьому контексті, досліджується питання резюмування текстових даних українською мовою з метою покращення якості моделей обробки природної мови.

Опис проблеми: Головною проблемою є нестача належного обсягу даних для ефективного навчання моделей обробки природної мови для української мови. Це обмежує можливості розвитку систем обробки текстів та призводить до необхідності пошуку методів резюмування наявних даних.

Метою цього дослідження є аналіз та порівняння різних методів резюмування текстових даних українською мовою з метою покращення якості моделей обробки природної мови.

Гіпотеза полягає в тому, що застосування різноманітних методів резюмування даних може допомогти збільшити обсяг та різноманітність наявних текстових даних українською мовою, що в свою чергу призведе до покращення якості моделей обробки природної мови.

Ключові знахідки та результати:

Під час дослідження було виявлено, що застосування різних методів резюмування даних дозволяє покращити результати моделей обробки природної мови для української мови. Зокрема, у контексті машинного

навчання і обробки текстів для української мови були використані такі моделі та алгоритми:

Transformer модель: Досліджено та застосовано Transformer модель для вирішення проблеми стандартних підходів до машинного навчання, що використовують архітектури, побудовані на рекурентних нейронних мережах (RNN). Transformer відрізняється тим, що не використовує рекурентні мережі, а замість цього використовує механізм внутрішньої уваги (self-attention). Це дозволяє моделі краще моделювати далекі залежності між словами та забезпечує можливість паралелізації методів, що покращує ефективність навчання.

Методи резюмування даних: Використання методів резюмування даних, таких як заміна синонімів, додавання шуму до текстів, а також генерація нових текстів на основі наявних, сприяє покращенню якості моделей. Ці методи допомагають розширити обсяг та різноманітність наявних текстових даних українською мовою, що в свою чергу призводить до покращення результатів у завданнях машинного перекладу, розпізнавання іменованих сутностей та сентимент-аналізу.

Оцінка та порівняння моделей: Проведено оцінку та порівняння різних моделей, навчених для української мови. Це включає в себе оцінку точності та ефективності в різних мовних завданнях, що дозволяє визначити найбільш оптимальні підходи та методи резюмування даних для даного контексту.

Ці знахідки підкреслюють важливість застосування сучасних моделей машинного навчання, таких як Transformer, разом із методами резюмування даних для покращення якості обробки текстів новин переважно українською мовою. Застосування методів резюмування даних є важливим етапом у покращенні моделей обробки природної мови для новин та їх публікацій. 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 новин.

Список використаних джерел:

1. Американське відділення Асоціації комп'ютерної лінгвістики: Технології людської мови – матеріали конференції. 2019. № Mlm (1). С. 4171-4186.

2. Edmundson H.P. New Methods in Automatic Extracting // Journal of the ACM (JACM). 1969. № 2 (16). P. 264-285.

3. Graham Y. Переоцінка автоматичного узагальнення за допомогою BLEU і 192 відтінків ROUGE // Матеріали конференції – EMNLP 2015: Conference on Empirical Methods in Natural Language Processing. 2015. № 9. P. 128-137.