УДК 004.932

Б.С. Білецький

ВИКОРИСТАННЯ МЕТОДІВ ОБРОБЛЕННЯ ПРИРОДНОЇ МОВИ ДЛЯ ВИЛУЧЕННЯ ВИЗНАЧЕНЬ СЛІВ ІЗ КОНТЕКСТУ

Вінницький національний технічний університет, Україна

Анотація

Це дослідження фокусується на застосуванні методів оброблення природної мови (NLP) для вилучення визначень слів з їхнього природного контексту. У дослідженні використовується комбінація бібліотек і моделей NLP, включаючи SpaCy, bert-extractive-summarizer і NLTK, для аналізу та виявлення речень, які найкраще відображають визначення заданого слова. Використовуючи можливості BERT, найсучаснішої мовної моделі, дослідження досягає точних і контекстуально релевантних визначень слів. Експериментальні результати демонструють ефективність запропонованого підходу для вилучення значень слів з текстових контекстів. Ця робота робить внесок у розвиток методології NLP та її практичне застосування для вдосконалення автоматизованих систем розпізнавання мови та інформаційно-пошукових систем.

Ключові слова: оброблення природної мови, вилучення визначення слова, SpaCy, bert-extractive-summarizer, бібліотека NLTK, BERT, контекстний аналіз, пошук інформації, розуміння мови.

Abstract

This research focuses on the application of natural language processing (NLP) techniques to extract word definitions from their natural context. The research uses a combination of NLP libraries and models, including SpaCy, bert-extractive-summarizer, and NLTK, to analyze and identify sentences that best represent the definition of a given word. Using the capabilities of BERT, a state-of-the-art language model, the study achieves accurate and contextually relevant word definitions. Experimental results demonstrate the effectiveness of the proposed approach for extracting word meanings from textual contexts. This work contributes to the development of NLP methodology and its practical application to improve automated speech recognition systems and information retrieval systems.

Keywords: natural language processing, word definition extraction, SpaCy, bert-extractive-summarizer, NLTK library, BERT, contextual analysis, information retrieval, speech understanding.

Вступ

У сучасну епоху величезного цифрового контенту точне розуміння значення слів у їхньому природному контексті має вирішальне значення для завдань мовної обробки. Це дослідження присвячене проблемі вилучення точних визначень слів за допомогою передових методів обробки природної мови (Natural Language Processing, NLP) [1-9].

Дослідження вивчає можливості популярних бібліотек NLP, таких як SpaCy та NLTK, а також потужної моделі BERT та бібліотеки bert-extractive-summarizer. SpaCy надає лінгвістичні анотації та функціональні можливості, включаючи токенізацію та синтаксичний аналіз залежностей, тоді як bert-extractive-summarizer використовує BERT, найсучаснішу мовну модель, для виявлення ключових речень. Ці технології дозволяють ідентифікувати контекстно-релевантні речення, допомагаючи витягувати визначення слів. Для цього завдання можна було б розглянути альтернативні технології, такі як WordNet. NLTK пропонує інструменти для різних завдань NLP, тоді як WordNet надає визначення і семантичні зв'язки між словами. Попередньо навчені мовні моделі, такі як GPT і RoBERTa, також можуть бути використані для вилучення визначень на основі контексту.

Використовуючи SpaCy, bert-extractive-summarizer та BERT, це дослідження представляє комплексний аналіз вилучення дефініцій слів з природного контексту.

Результати дослідження

Для проведення дослідження було ретельно відібрано колекцію з 12 текстових уривків, які охоплюють різноманітні сфери та мовні стилі. Ці уривки були оброблені за допомогою бібліотеки SpaCy для токенізації речень, що забезпечило детальне представлення тексту. Згодом було використано бібліотеку bert-extractive-summarizer для виявлення речень, найбільш схожих на визначення цільового слова (рис. 1).

target	test_sentence	similarity_score
Інформаційні технології - це	Отже, інформаційні технології (IT) — це сукупність методів і засобів, що використовуються для збору, зберігання, обробки і поширення інформації.	0.454
Інформаційні технології - це	Інформаційні технології призначені для зниження трудомісткості процесів використання інформаційних ресурсів. [джерело?]	0.432
Інформаційні технології - це	На базі цієї техніки з'являється новий вид технологій — інформаційні.	0.423
Інформаційні технології - це	Інформаційні технології — давно звичні для всіх слова, які дуже точно характеризують життя і потреби сучасного суспільства.	0.42
Інформаційні технології - це	Сюди входить її збір, структуризація, оформлення, редагування — ці завдання виконують web-програмісти, web-дизайнери, контент-менеджери, менеджери інтернет-проекту.	0.407
Інформаційні технології - це	Люди, які не пов'язані з IT, швидше за все, скажуть, що це щось складне, незрозуміле і розумне.	0.391
Інформаційні технології - це	І сьогодні вони оточують нас у всіх сферах життя: записна книжка в вашому телефоні – це база даних, бортовий комп'ютер автомобіля – спеціальна обчислювальна система.	0.371
Інформаційні технології - це	Інформація — будь-які відомості або дані, які можуть бути збережені на матеріальних носіях або відображені в електронному вигляді.[джерело?]	0.357
Інформаційні технології - це	Але чи так це насправді?	0.341

Рисунок 1 — Таблиця речень, які зі створеного набору документів мають найвищий рівень схожості з шуканим визначенням.

Використовуючи модель BERT, було згенеровано контекстні пропозиції, які вдосконалили процес вилучення дефініції слова. Результати продемонстрували, що запропонований підхід дає точні та релевантні визначення слів, враховуючи семантичні та контекстуальні аспекти тексту. Поєднання SpaCy, bert-extractive-summarizer і BERT значно підвищило точність і повноту вилучення дефініцій слів порівняно з традиційними методами.

Продовжуючи дослідження в цій галузі, можна продовжувати вдосконалювати і розширювати можливості систем NLP, що в кінцевому підсумку дозволить більш точне і нюансоване розуміння мови. Знання, отримані в результаті цього дослідження, закладають основу для майбутніх досліджень, сприяючи інноваціям і прогресу в галузі обробки природної мови.

Важливо зазначити, що хоча сучасні технології обробки природної мови (NLP) досягли значного прогресу, вони не можуть працювати оптимально для всіх мов, включаючи українську. Це обмеження підкреслює необхідність розробки та впровадження національних технологій, спеціально адаптованих для української мови. Українська мова має унікальні лінгвістичні характеристики та нюанси, які вимагають спеціальної уваги та лінгвістичних ресурсів для ефективного опрацювання та аналізу тексту.

Висновки

Оглянуто сучасні підходи до визначення слів в їх природного контексті. Запропоновано застосування технологій SpaCy, bert-extractive-summarizer та BERT виявилося ефективним підходом для виявлення семантичних нюансів та контекстних варіацій значень слів. Результати експерименту демонструють важливість контекстно-орієнтованого аналізу для точної ідентифікації дефініцій слів, прокладаючи шлях до покращення розуміння мови та інформаційно-пошукових систем. Результати цього дослідження роблять внесок у сферу NLP, демонструючи потенціал сучасних моделей і бібліотек у вилученні точних і змістовних визначень слів. Подальші дослідження можуть бути зосереджені на розширенні набору даних, вивченні додаткових методів NLP та оцінці запропонованого підходу на різних мовах і доменах для подальшого підвищення його застосовності та узагальненості.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Livinska, H. V., and Oleksandr Makarevych. "Feasibility of improving BERT for linguistic prediction on Ukrainian corpus." CEUR Workshop Proceedings. 2020.
2. Berko, Andrii, et al. "The text classification based on Big Data analysis for keyword definition using stemming." 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT). Vol. 1. IEEE, 2021.
3. Cheilytko, Nataliia, and Ruprecht von Waldenfels. "Exploring Word Sense Distribution in Ukrainian with a Semantic Vector Space Model." Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP). 2023.
4. Zaiev, Andrii, and Oleksii Turuta. "APPLICATION OF GATED UNITS TO BERT-BASED MODELS." Збірник наукових праць ΛΌГОΣ (2020): 36-38.
5. Ткаченко, Олександра Олексіївна, and Олена Володимирівна Олійник. "МОЖЛИВОСТІ ТА ТРУДНОЩІ ВИКОРИСТАННЯ ОБРОБКИ ПРИРОДНОЇ МОВИ." Практичні та теоретичні питання розвитку науки та освіти (частина І): матеріали ІІ Міжнародної науково-практичної конференції м. Львів, 19-20 грудня 2020 року.–Львів: Львівський науковий форум, 2020.–74 с.: 73.
6. Супрун, О. П. Інтелектуальна технологія обробки природної мови. MS thesis. Сумський державний університет, 2021.
7. Мокін В. Б. Інформаційна інтелектуальна технологія автоматизованої обробки текстової природно-мовної інформації / В. Б. Мокін, М. А. Гораш, Є. М. Крижановський // Матеріали L науково-технічної конференції підрозділів ВНТУ, Вінниця, 10-12 березня 2021 р. – Електрон. текст. дані. – 2021. – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2021/paper/view/12924
8. Мокін В. Б. Інформаційна інтелектуальна технологія автоматизованої геоприв'язки екологічної текстової природно-мовної інформації / В. Б. Мокін, М. А. Гораш, Є. М. Крижановський, Т. Є. Вуж // Наукові праці ВНТУ [Електронний ресурс]. – 2020. – № 4. – Режим доступу: https://praci.vntu.edu.ua/index.php/praci/article/view/624
9. В.Б. Мокін, М.А. Гораш, Д. Пасічнюк, О. Радецький. Концепція інтелектуальної NLP технології для геоприв'язки та класифікації відкритої текстової інформації про масиви вод // Матеріали XV міжнародної конференції "Контроль і управління в складних системах (КУСС-2020)", м. Вінниця, 8-10 жовтня 2020 р. - Вінниця, 2020. – Режим доступу: https://ir.lib.vntu.edu.ua/handle/123456789/30607

Білецький Богдан Сергійович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: bohdanbeletskyi@gmail.com.

Biletskyi Bohdan S. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: bohdanbeletskyi@gmail.com.