УДК 004.89: 004.5

УДОСКОНАЛЕННЯ МЕТОДУ СЕМАНТИЧНОГО АНАЛІЗУ ТЕКСТУ

Яровий Андрій, Кудрявцев Дмитро, Крилик Людмила

Вінницький національний технічний університет

Анотація

Під час проведеного дослідження було розглянуто сучасні методи аналізу текстової інформації. Визначено фактори, що негативно впливають на семантичний аналіз тексту. Запропоновано способи покращення семантичного аналізу тексту. Реалізовано семантичний аналіз тексту на прикладі чат-бота із застосуванням запропонованих рішень та порівняно результат їх роботи.

Abstract

In the given research modern methods of text analyzing were noted. The factors which negatively influent on text semantic analysis were determined. The news ways to improve text semantic analysis were proposed. Text semantic analysis on the example of a chat-bot with the application of the proposed solutions was implemented and the result of their work was compared.

Більшість сучасних рішень автоматичного аналізу тексту застосовують методи та алгоритми, що орієнтовані на морфологічний, фонетичний та синтаксичний рівень аналізу тексту. Кожен з даних рівнів залежить один від одного та в переважній більшості застосовує результати більш нижчого рівня. Найвищим рівнем аналізу тексту є його логіко-семантичний рівень, що визначає тему, ключові слова та дає змогу експерту оцінити важливість інформації за обраними критеріями [1].

Автоматичний аналіз тексту застосовується у соціальних та пошукових мережах, месенджерах, рекламних сервісах та інших сферах діяльності людини. Переважна більшість текстової інформації, що аналізується, формується в автоматичному режимі із використанням ключових слів, що містяться в будь-яких інформаційних текстових джерелах та містять базові морфологічні конструкції з низьким рівнем емоційного забарвлення [2]. Завдяки цьому, існуючі методи морфологічного та семантичного аналізу тексту нехтують можливими відхиленнями в плані визначення тематики інформації та її змісту заради прискорення аналізу. Дане рішення є досить виправданим у разі обробки великих джерел інформації таких як результати пошукових запитів, технічна документація чи сховище даних чат-бота [3]. Розглядаючи метод морфологічного аналізу тексту, було помічено, що внаслідок багаторазового розподілу текстової інформації на морфеми, подальший семантичний аналіз дещо спотворюється із-за таких факторів:

- наявність у синтаксичних конструкцій декількох значень
- застосування літературного стилю
- текстові скорочення

Покращення автоматичного аналізу тексту більшою мірою пов’язане із зменшенням впливу даних факторів, а саме додаванням «словників» та баз знань, що ізолюють текстову інформацію, приналежну до описаних факторів. Але навіть за умови виявлення подібного тексту, його аналіз в більшій мірі спирається на порівнянні із власним джерелом. А у разі його відсутності, аналіз не проводиться взагалі і семантичний аналіз не враховує даний текст. В даному випадку – це рішення потребує більш детального розгляду та передбачає залучення експерта або технологій штучного інтелекту, що включатимуть відповідні засоби для вирішення проблеми.

Для більш детального аналізу даної проблеми було обрано 20 довільних текстів різних типів (літературний текст, записи в соціальних мережах, технічна документація, статті в наукових журналах), кожен обсягом 450-500 слів. В якості засобів для

автоматичного аналізу тексту було застосовано онлайн-сервіс Advego та програмну бібліотеку Dandelion [4]. Зведені результати семантичного аналізу представлені в таблиці 1 (джерела з наукових статей).

Таблиця 1 – Зведені результати СА для джерел з наукових статей

Тип тексту	**Статті в наукових журналах**				
Частота появи слів, що зустрічаються не більше 2-ох разів	11.4%	29.3%	31.2%	32.4%	21.7%
Час виконання аналізу, с	0.20	0.28	0.46	0.25	0.32
Частота появи ключових слів	75.89%	80.09%	81.6%	71.2%	74.58%
Літературність тексту (частота повторень, (понад 5 разів), що не входять до ключових слів)	4.8%	1.8%	5.6%	9.4%	10.2%
Кількість скорочень (знайдених / перевірених):	2/8	1/15	9/12	8/20	3/17
Кількість ключових слів:	370	389	383	342	378

За результатом семантичного аналізу, доцільно зробити висновок щодо низької здатності до ідентифікації текстових скорочень та наявності надлишкової інформації. Для вирішення даної проблеми пропонується декілька рішень, а саме:

- Застосувати словник для скорочень
- Додати фільтрацію слів із низькою частотою появи у тексті
- Виконувати семантичний аналіз не лише усього тексту, але і його рівних за розміром частин

Для перевірки дієвості даних рішень, було створено чат-бот для задачі семантичного аналізу тексту. Оскільки кожне з рішень матиме власний вплив на результат перевірки, кожне рішення було реалізовано окремо та додана можливість їх поєднання. Отримані результати на прикладі джерел із статей в наукових журналах представлено в таблицях 2-4.

Таблиця 2 – Порівняння результатів при додаванні словника для скорочень

Тип тексту	**Статті в наукових журналах**				
Кількість скорочень без словника (знайдених / перевірених):	2/8	1/15	9/12	8/20	3/17
Кількість скорочень зі словником (знайдених / перевірених):	6/8	12/15	10/12	17/20	11/17

При додаванні словника скорочень, кількість зафіксованих скорочень значно зросла, що дає змогу не використовувати їх при визначенні теми тексту та не включати до списку ключових слів. Це в свою чергу підвищує точність визначення теми текстової інформації.

Таблиця 3 – Порівняння результатів при фільтрації слів із низькою частотою появи

Тип тексту	**Статті в наукових журналах**				
Частота появи ключових слів	68.42%	73.05%	66.94%	68.04%	73.89%
Кількість ключових слів без фільтрування:	370	389	383	342	378
Кількість ключових слів з фільтруванням:	325	347	318	330	351

Додавання фільтрації для слів із низькою частотою появи спричинило зменшення кількості ключових слів на 2-15%, що також позитивно впливає на визначення теми текстової інформації.

Таблиця 4 – Порівняння результатів при семантичному аналізі окремих частин тексту (по 100 слів у частині)

Тип тексту	**Статті в наукових журналах**				
Частота появи ключових слів (без фільтрування)	75.89%	80.09%	81.6%	71.2%	74.58%
Частота появи ключових слів (з фільтруванням)	45.89%	57.47%	54.73%	36%	30.31%
Час виконання аналізу (без аналізу окремих частин), с	0.20	0.28	0.46	0.25	0.32
Час виконання аналізу (з аналізом окремих частин), с	0.31	0.61	0.89	0.77	0.54
Кількість ключових слів без фільтрування:	370	389	383	342	378
Кількість ключових слів з фільтруванням:	218	273	260	171	144

Додатковий семантичний аналіз окремих частин тексту виявився найкращим серед представлених рішень та зменшив кількість ключових слів на 20-45%, що суттєво підвищить точність визначення тематики тексту. В результаті проведеного дослідження, розглянуто задачу підвищення точності визначення тематики текстової інформації за допомогою семантичного аналізу. Серед запропонованих рішень найкращим виявилось розбиття тексту на окремі частини та порівняння їх результатів, але внаслідок чого зріс час обробки. Застосування даного рішення доцільно в сфері інтелектуального аналізу даних, а саме чат-ботах. В подальшому дослідженні буде проведено поєднання даних рішень між собою та аналіз їх застосування.

Список використаних джерел

1. Інтелектуальна обробка текстів: [навчальний посібник] / В. Ю.Тарануха. – Київ: електронна публікація на сайті факультету, 2014. – 80 с. 2016.

2. Pereira J. F. F. Social media text processing and semantic analysis for smart cities //arXiv preprint arXiv:1709.03406. – 2017.

3. Andrii Yarovyi, Dmytro Kudriavtsev, Serhii Baraban, Volodymyr Ozeranskyi, Liudmyla Krylyk, Andrzej Smolarz, and Gayni Karnakova "Information technology in creating intelligent chatbots", Proc. SPIE 11176, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019, 1117627 (6 November 2019); https://doi.org/10.1117/12.2537415.