

МЕТОДИ И ТЕХНИКИ ЗА ПОДОБРЯВАНЕ НА РЕЗУЛТАТИТЕ ПРИ КОМПЮТЪРНА ОБРАБОТКА НА ТЕКСТ

Снежана Сълова¹

¹ Икономически университет – Варна / Катедра „Информатика“, гр. Варна, България,
ssulova@ue-varna.bg

РЕЗЮМЕ

Компютърната обработка на естествен език е област от изкуствения интелект, която намира приложение в много задачи, свързани с анализа и разбирането на текст. Понастоящем въпреки напредъка на технологиите, правилното разбиране на контекста все още остава не лека задача. Прецизността при обработката и анализа на естествен език е ключов фактор за разбиране и интерпретиране на значението на текстовете. Докладът прави обобщение и систематизиране на съвременните методи и техники, които могат да се използват за подобряване на разбирането на контекста в текстовете. Основният фокус е върху техниките за обогатяване на тренировъчните данни, подобряване на предварителната обработка на текст и моделите за решаване на специфични задачи, свързани с обработката на текст. Дават се препоръки за подобрене, както на конкретни задачи, така и общо за усъвършенстване на процесите на обработка на текст.

КЛЮЧОВИ ДУМИ: обработка на естествен език, предварителната обработка на текст, езикови модели

METHODS AND TECHNIQUES FOR IMPROVING RESULTS IN NATURAL LANGUAGE PROCESSING

Snezhana Sulova¹

¹ University of Economics – Varna / Department of Informatics, Varna, Bulgaria,
ssulova@ue-varna.bg

ABSTRACT

Natural language processing is an area of artificial intelligence that can be used in many tasks related to the analysis and understanding of text. Nowadays, despite the advancement of technology, understanding the context correctly is still a troublesome task. Precision in natural language processing and analysis is a key factor in accurately understanding and interpreting the meaning of texts. The paper summarizes and analyzes modern methods and techniques that can be used to improve the understanding of context in texts. The main focus is on techniques for enriching training data and improving text preprocessing, transformers and other types of models for solving specific tasks related to text processing. Recommendations are given for improvement, both for specific tasks and generally for improving the text processing operations.

KEYWORDS: natural language processing, text preprocessing, language models

ВЪВЕДЕНИЕ

Компютърната обработка на естествен език (Natural language processing – NLP) е област на изкуствения интелект, която може да се разглежда като синтез на лингвистика, компютърни науки и машинно обучение. Целта е компютрите да разбират, интерпретират и генерират човешки език по начин, който е полезен и разбираем (Kochmar, 2022). В съвременния свят,

компютърната обработка на естествен език играе ключова роля в различни области – от чат-ботове и търсачки до сложни системи за анализ на текстови данни в медицината, правото и други сфери. Според Statista (2024) размерът на пазара на NLP се очаква да има годишен темп на растеж 26,01%, което води до пазарен обем от 25,74 милиарда евро до 2030 г.

С напредването на технологиите, появата на нови източници на текстови данни, нуждата от усъвършенстване на методите и техниките за тяхната обработка става все по-значима. Ежедневно чрез уеб, социалните платформи се генерират големи обеми от текстове, които могат да станат полезен източник на данни. За процеса на обработка и трансформация на текстовите данни няма единен алгоритъм, тъй като те са специфични както по структура, така и по начин на формиране. В научните изследвания на базата на проведени проучвания върху различни видове текстови данни се правят предложения за подобрения на методите и техниките за постигане на по-голяма прецизност при обработката. В тази връзка **целта** на доклада е на базата на съществуващите проучвания да изведат и систематизират основните начини, чрез които може да подобрят точността, ефективността и адаптивността на NLP системите. Акцентира се върху използването на по-усъвършенствани модели, обогатяване на данните за обучение, подобряване на разбирането на контекста, и внедряването на нови подходи за обработка на текст.

1. ОСНОВНИ КОНЦЕПЦИИ В КОМПЮТЪРНАТА ОБРАБОТКА НА ТЕКСТ

Компютърната обработка на естествен език води своето начало от 50-те години на миналия век. По същество това е сравнително изследвана област, която обаче непрекъснато се развива и днес се базира на използването на модели с дълбоко машинно обучение, които работят с огромни обеми от данни, самообучават се и използват мощни изчислителни ресурси. Целта на NLP е компютърът да може по-добре да „разбира“ съдържанието на текстовите документи, да открива настроенятия в контекста, да извлича информация и идеи от текстови документи и да категоризира самите документи.

Процесът на автоматизирана обработка на текст е еволюирал в своето развитие. Началото се свързва с предложената от британския математик Алън Тюринг през 1950 г. концепция – тестът на Тюринг за проверка на интелигентността на машините (Turing, 1950). Първите методи за обработка на текст се базират на ръчно създавани правила и речници и се фокусират върху прости задачи като машинния превод и анализа на синтаксиса. След това навлизат статистически методи, N-грамните модели, вероятностните модели на Марков (Hidden Markov models). Еволюцията продължава с модели, базирани на машинно и дълбоко машинно обучение и понастоящем се използват генеративни предварително обучени трансформаторни модели, като Generative Pre-trained Transformer – GPT и Bidirectional Encoder Representations from Transformers – BERT.

Обработката на естествен език намира практическо приложение при:

- интелигентно семантично търсене, в търсачките, системите за превод когато е необходимо да се идентифицира смисъла на текста;
- анализ на доклади, текстови документи и извличане на информация, свързана с развитие, пазарни ситуации, тенденции;
- резюмиране на текст, намиране на главната идея на дадена статия, откриване на актуални теми, ключови думи;
- анализ на потребителски мнения, настроения и персонализиране на услуги;

- чат бот системи и виртуални асистенти за разговор с потребител и автоматичен отговор на въпроси;
- преводачи и работа с многоезични модели за обработка на текст на множество езици;
- откриване на спам, фалшиви новини, чрез анализ и категоризация на новини и др.

Изборът на подход за осъществяване на компютърна обработка на естествен език е важен за прецизността на анализите, базирани на текстови данни, същевременно обаче този процес се свързва и с редица предизвикателства. Проблеми възникват при текстовия анализ и работа с близки езици, с текстове, в които има неформална реч, чуждици и жаргони. Съществуват голям брой изследвания, които представят специално разработени за даден език изчислителни методи или такива, които са адаптирани за предварителна обработка, нормализация, анализиране на подобни езици, езикови разновидности и диалекти (Zampieri, Nakov, Scherrer, 2020). Сложността на граматиката и езиковите структури в много от езиците също създават проблеми при автоматичната обработка. Друго ограничение на NLP е че повечето специализирани ресурси, речници, езикови колекции и системи са достъпни само за определени езици, най-често за английски, френски, испански, немски и китайски език. Развитието на уеб услугите води до генериране на нови видове текстови данни в уеб пространството и става предпоставка за търсенето на подходящи, специализирани начини за тяхното извличане и обработка както и прилагането на иновативни методи за подобряване на резултатите.

2. МЕТОДИ И ТЕХНИКИ ЗА ПОДОБРЯВАНЕ НА ПРЕЦИЗНОСТТА ПРИ ОБРАБОТКАТА НА ТЕКСТ

За да се идентифицират ключовите понятия и фрази при съвременните научни изследвания, посветени на методите за обработка на текст са открити и анализирани публикации в световно известната база с научна литература Скопус. Идентифицирани са научните разработки за последните 5 г. в област Computer science, отговарящи на заявка, включваща ключовите думи „Natural language processing“, methods, techniques, improving. Обработени са резюметата на намерените 652 публикации и са извлечени най-често използваните думи и фрази в тях. Честотата на използваните термини е схематично представена на фигура 1.

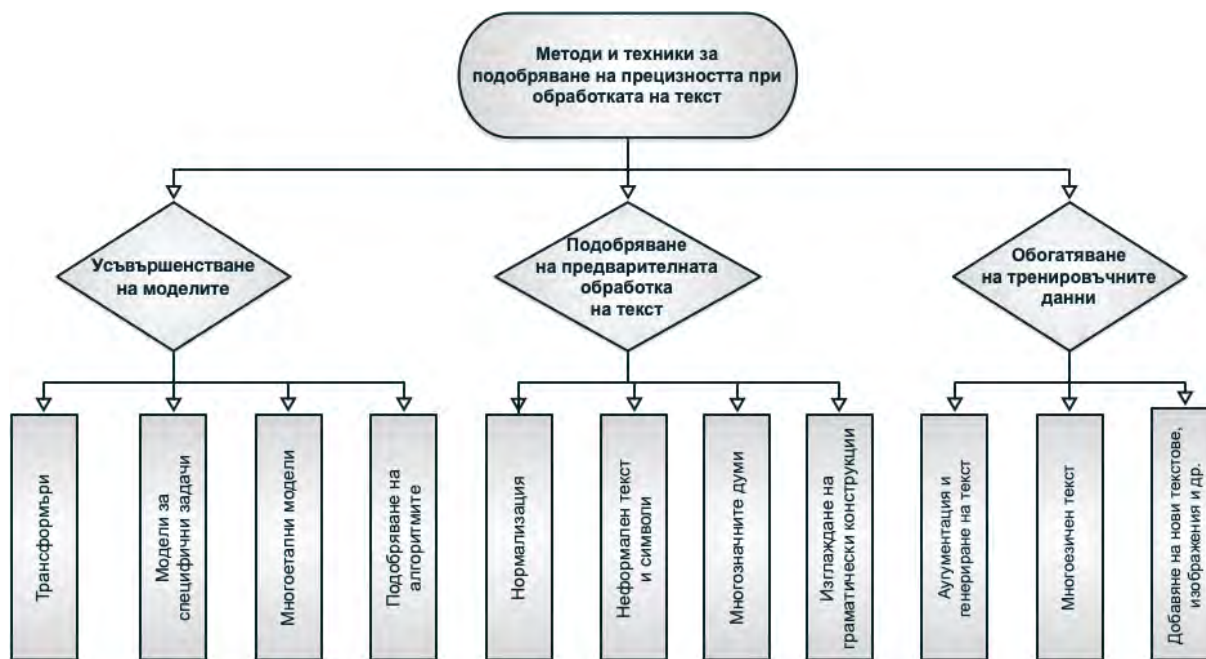


Фигура 1. Wordcloud диаграма на идентифицираните термини

Източник: Собствена разработка

На базата на проведеното по-горе проучване, както и въз основа на предходни изследвания на автора по темата за компютърната обработка на текст, може да се обобщи, че основни групи методи и техники, които се използват за подобряване на прецизността при обработката на текст се базират на (вж. фиг. 2):

- приложение на усъвършенствани модели;
- подобрена предварителна обработка на текст;
- обогатяване на тренировъчните данни.



Фигура 2. Методи и техники за подобряване на прецизността при обработката на текст

Източник: Собствена разработка

В научните изследвания в последните години голяма част от предлаганите подходи се базират на различни усъвършенствани, предварително обучени модели. Такива са моделите, базирани на трансформаторни модели като BERT и GPT. Те позволяват на NLP системите да разберат по-добре контекста, тъй като разполагат със self-attention механизъм (Vaswani et al., 2017), който помага на моделите да идентифицират по-добре връзките между думите и да обръщат внимание на различните части на текста, за да интерпретират правилно неговия смисъл. След като един трансформаторен модел е предварително обучен върху голям набор от данни, той може да бъде фино настроен върху специфичен набор от данни, за да изпълни определена задача. Така вместо да се обучават моделите от нулата се използват предварително обучени трансформаторни модели върху специфични задачи, свързани със специализирани текстове. Например, ако е необходимо да се създаде чатбот за обслужване на клиенти, може фино да се настрои модела върху голям набор от диалози между клиенти и служители. За някои специфични задачи в медицината или друга област може да се наложи да се използва специална токенизация, която да отчита особеностите на текста и да разпознава използваните термини и съкращения.

Известно е, че при обработката и анализа на текст предварителната обработка е важен етап. Тя е от значение за превръщане на текста във вид удобен за числово представяне, което да се разбере от моделите. Конвенционалните практики за предварителна обработка на текст

обикновено са достатъчни, но има случаи, при които предварителната обработка на текст трябва да бъде персонализирана спрямо текста, за да даде по-добри резултати при анализа (Chai, 2023). Нормализацията на текста чрез коригиране на грешки при въвеждане, премахване на излишни символи и уеднаквяване на форматите за дати, валути може да подпомогне обработката. При работа с текстове от социалните медии, където се използва неформален език за общуване, съдържащ жаргон, емотикони, съкращения и неправилна граматика е добре да се прилагат специални модели и техники за работа с неформален език. Екип от италиански учени правят проучване и доказват, че изборът на правилната техника за предварителна обработка на текстовете може значително да подобри точността при класификационните алгоритми и дори в някои случаи подходящата стратегия за предварителна обработка дава по-добри резултати от трансформаторните модели (Siino, Tinnirello, La Cascia, 2024).

Обогатяването на тренировъчните данни е важен елемент за подобряване на точността и надеждността на NLP моделите. Позволява на моделите да се научат да разпознават по-широк спектър от езикови конструкции и нюанси в текстовете, което е предпоставка за по-точни прогнози. Използването на различни техники, като аугментация на данните чрез заместване на думи със синоними, ротация на думи и фрази в изреченията, генериране на нови данни. В някои от случаите тренирането на моделите върху многоезични корпуси и включването на допълнителни данни като нови текстове, изображения или аудио също може да подобри разбирането на текстовете като даде нова и важна информация.

ЗАКЛЮЧЕНИЕ

Обработката на естествения език е динамична област, която непрекъснато се развива. Моделите за обработка на текст стават все по-ефективни, но и по-сложни. Подобряването на автоматизирана обработка на текста може значително повиши качеството на услугите, базирани на NLP ето защо е необходимо избор на правилен метод за анализ и разработване на модели, които са по-устойчиви на грешки в данните и при необходимост комбиниране на лингвистични, социални и културни знания за по-добро разбиране на текста.

БЛАГОДАРНОСТИ

Това изследване е направено по проект ПНИ-KC24-04-DCAAITM „Развитие на компетентности за прилагане на средства и методи на изкуствен интелект“.

REFERENCES / ИЗПОЛЗВАНА ЛИТЕРАТУРА

1. CHAI, C. P. (2023) Comparison of text preprocessing methods. Natural Language Engineering, 29(3), pp. 509-553.
2. KOCHMAR, E. (2022) Getting started with natural language processing. New York: Simon and Schuster.
3. SIINO, M., TINNIRELLO, I., LA CASCIA, M. (2024) Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Information Systems, Vol. 121, 102342, DOI: <https://doi.org/10.1016/j.is.2023.102342>.
4. STATISTA (2024) Language translation NLP – Global. Market Forecast [Online] Available from: <https://www.statista.com/outlook/tmo/artificial-intelligence/natural-language-processing/language-translation-nlp/worldwide?currency=EUR/> [Accessed 5/10/2024].

5. TURING, A. (1950) Computing machinery and intelligence. IEEE Intelligent Systems. Vol. 2.
6. VASWANI, A. et al. (2017) Attention is all you need. Curran Associates Inc. pp. 6000–6010.
7. ZAMPIERI, M., NAKOV, P., SCHERRER, Y. (2020). Natural language processing for similar languages, varieties, and dialects: A survey. Natural Language Engineering, 26(6), pp. 595-612.