ЗАСТОСУВАННЯ ТОКЕНІЗАЦІЇ ДАНИХ НА ОСНОВІ BYTE-PAIR ENCODING

Оченашко М.О., Гороховатський В.О.

Кафедра інформатики,
Харківський національний університет радіоелектроніки,
Україна.

E-mail: maksym.ochenashko@nure.ua,
volodymyr.horokhovatskyi@nure.ua

Abstract

Efficient data representation is essential for modern information systems, particularly in natural language processing (NLP) and secure data handling. The analysis examines Byte-Pair Encoding (BPE) as an effective and adaptive method of tokenization that transforms raw text into structured subword units through iterative merging of frequent symbol pairs. The approach is particularly relevant for NLP pipelines and large-scale data management, where compact and semantically consistent representations improve computational efficiency and system scalability. The analysis explores how the frequency-based merging strategy reduces vocabulary size and sequence length without compromising linguistic content. According to the results, BPE can be recommended as a standardized approach for text processing in large language model systems. Its simplicity, adaptability, and efficiency make it a reliable foundation for scalable, secure, and interoperable data pipelines in modern information environments.

Великі мовні моделі (LLM) та сучасні інформаційні системи обробляють величезні обсяги неструктурованого тексту з різних джерел, таких як соціальні мережі, корпоративні записи та онлайн-платформи для спілкування. Оскільки ці системи покладаються на текстові дані для генерації, пошуку або виведення значення, ефективне представлення тексту стало необхідним як для обчислювальної продуктивності, так і для забезпечення точності моделі. Токенізація – процес перетворення безперервного тексту на дискретні одиниці, або токени - є основою всіх конвеєрів обробки тексту LLM. Вона безпосередньо впливає на довжину послідовності, вимоги до пам'яті та здатність моделі “розуміти” невідомі або морфологічно складні слова [1-3].

У токенізації на основі Byte-Pair encoding (BPE) кожне слово розбивається на фрагменти таким чином, що фрагменти слів які зустрічаються часто, стають окремими токенами, а рідкісні слова складаються з менших частин. Це значно зменшує розмір словника, необхідного для охоплення мови, при цьому дозволяючи представляти будь-яке слово без використання загального невідомого токена. Здатність BPE скорочувати довжину послідовності та усувати більшість токенів зробила його стандартом у natural language processing (NLP) конвеєрах. Цей підхід є ключовим для багатьох великих мовних моделей та систем перекладу, які вимагають як компактного представлення вхідних даних, так і гнучкості відкритого словника [1, 2, 4].

BPE працює за допомогою простого жадібного алгоритму, заснованого на частоті символів. Нехай початковий словник V складається із усіх окремих символів, присутніх у тексті. Текст спочатку є послідовністю цих символів. При кожній ітерації найчастіша пара сусідніх символів у тексті (скажімо, лексеми a і b) об'єднується в нову лексему ab. Формально операцію об'єднання можна визначити як:

$$pair^* = arg\ arg\ freq\ (a, b), \tag{1}$$

де $freq(a, b)$ - частота послідовного вживання пари a, b. Потім словник оновлюється, щоб включити новий токен:

$$V \coloneqq V \cup ab, \tag{2}$$

і всі випадки появи a, b в тексті замінюються токеном ab. Цей процес повторюється доти, доки не буде досягнуто заздалегідь визначеного розміру словника або жодна пара не матиме частоту, більшу за 1. Результатом є навчений словник підслів, і будь-який текст може бути токенізований шляхом жадібного зіставлення найдовших можливих токенів з цього словника. Кожне об'єднання зменшує загальну довжину послідовності (оскільки два символи стають одним), тим самим стискаючи представлення тексту.

На рисунку 1 показано процес токенізаці: процес починається з початкового словника окремих символів, потім послідовно підраховує всі сусідні пари символів у збірнику, об'єднує найчастіші пари в новий токен і оновлює словник. Цей цикл повторюється до досягнення цільової кількості токенів, в результаті чого отримуємо словник підслів і стиснуту послідовність токенів.

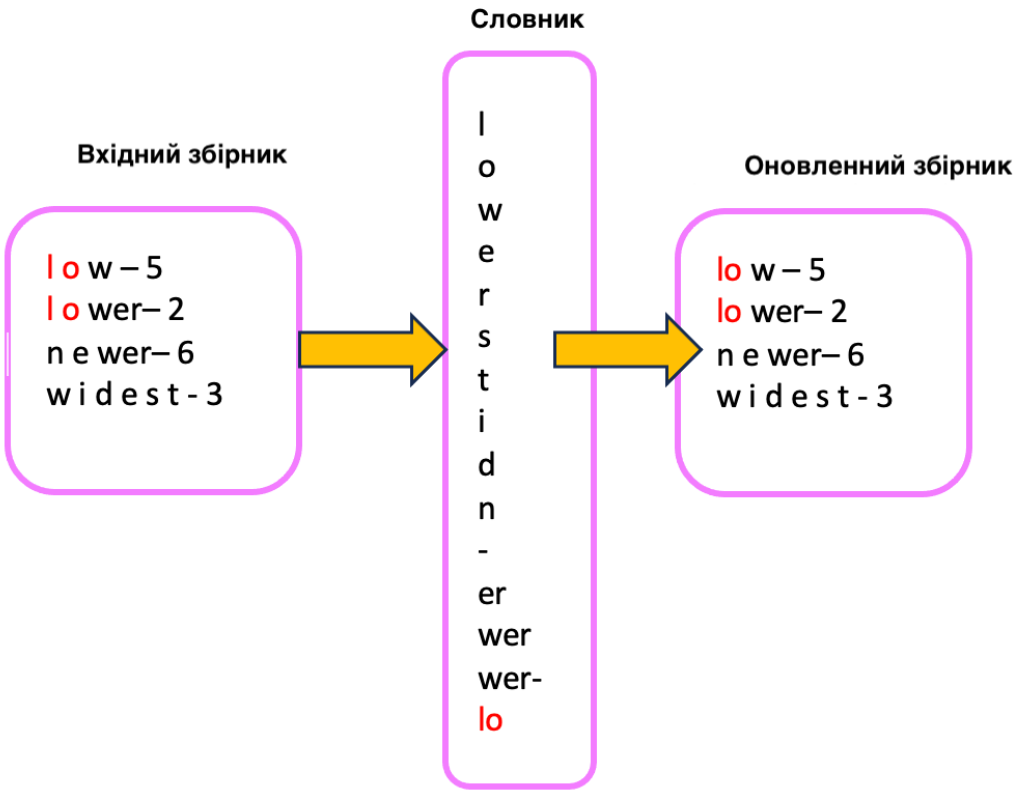

Рис. 1. Процес токенізації

Жадібне об'єднання на основі критерію частоти гарантує, що загальні підрядки (такі як «ing», «tion» або «pre») стають окремими токенами. Часті слова в результаті стають одним або декількома токенами, тоді як рідкісні або морфологічно складні слова розбиваються на частини, які самі по собі є загальними. Наприклад, якщо «unbelievable» є рідкісним, але «un», «believe» і «able» є загальними, BPE може представити «unbelievable» як un + believable або навіть un + believ + able. Це мінімізує кількість невідомих токенів і забезпечує стиснення довжини послідовності тексту порівняно з кодуванням символів. Варто зазначити, що оскільки BPE базується виключно на частоті, він не враховує лінгвістичне значення – деякі об'єднання можуть розділятися по межі морфем, а інші – ні.

BPE токенізатор це де-факто стандарт у NLP-конвеєрах. Він використовується в системах машинного перекладу, де було продемонстровано, що воно покращує переклад рідкісних слів, представляючи їх як підслова [1, 5]. Підхід також є невід'ємною частиною навчання великих мовних моделей (LLM), таких як GPT-2, GPT-3 та інших, які використовують BPE або подібні схеми підслів для етапу токенізації. Компактність токенів BPE означає, що моделі можуть мати керовані розміри словника (наприклад, від 30 до 50 тисяч токенів) замість мільйонів слів, що зменшує обсяг пам'яті та дозволяє ефективно навчати вектори вкладання. Крім того, BPE є незалежним від мови на рівні байтів, що дозволяє застосовувати уніфікований підхід до багатомовного тексту. Наприклад, один і той самий

алгоритм BPE може бути застосований до мов із спільними наборами символів, а спільні словники можуть бути вивчені для багатомовних збірників, щоб дозволити міжмовний обмін підсловами. В системах пошуку та отримання інформації токенізація допомагає нормалізувати текст (наприклад, обробляючи «connection» і «connections» із загальним кореневим токеном, таким як «connection»), що може поліпшити відповідність і зменшити розмір індексу. Детермінований характер є корисним для забезпечення узгодженості індексації та запитів у пошукових системах [1, 6–8]. Окрім навчання моделей NLP, токенізація відіграє важливу роль у процесах обробки великих даних [8, 9]. Стислі представлення токенів можна зберігати замість необробленого тексту, щоб заощадити простір і пришвидшити аналіз.

На основі проведеного аналізу було встановлено, що апарат Byte-Pair Encoding є гнучкою та ефективною технікою для оптимізації представлення даних в інформаційних системах, досягаючи балансу між ефективністю стиснення та гнучкістю словника. BPE значно скорочує довжину послідовностей (покращуючи швидкість обробки та зменшуючи використання пам'яті), уникаючи при цьому проблеми виходу за межі словника, властивої токенізації на рівні слів. Стабільність, адаптивність, та детермінізм цього процесу забезпечують послідовну обробку різноманітних лінгвістичних даних та безперебійну інтеграцію між різними компонентами системи. Дослідниками встановлено, що включення BPE в інформаційні конвеєри підвищує оперативну масштабованість особливо у об’ємних середовищах даних, які вимагають надійного та безпечного управління цифровим контентом [7–9].

Література

1. Sennrich, R., Haddow, B., & Birch, Neural Machine Translation of Rare Words with Subword Units, 2016. URL: https://arxiv.org/pdf/1508.07909 (дата звернення 09.10.2025).

2. Radford, A, (2019), Language Models are Unsupervised Multitask Learners. OpenAI Technical Report.

3. Gorokhovatsky, V. (2014), Structural Analysis and Intellectual Data Processing in Computer Vision, SMIT, Kharkiv, 316 p.

4. Daradkeh Y.I., Gorokhovatskyi V., Tvoroshenko I., and Zeghid M. (2024) Improving the effectiveness of image classification structural methods by compressing the description according to the information content criterion, Computers, Materials & Continua, vol. 80, no. 2, pp. 3085-3106.

5. Gorokhovatskyi, O., Peredrii, O., Gorokhovatskyi, V., Vlasenko, N. (2023) Explanation of CNN Image Classifiers with Hiding Parts. In: J. Benois-Pineau, R. Bourqui, D. Petkovic, G. Quenot (eds), Explainable Deep Learning Artificial Intelligence, pp. 125-146, Academic Press, 346 p.

6. Gorokhovatskyi V., Gadetska S., Stiahlyk N. (2020) Image structural classification technologies based on statistical analysis of descriptions in the form of bit descriptor set. In CEUR Workshop Proceedings: Computer Modeling and Intelligent Systems (CMIS-2020), 2608, 1027-1039.

7. Gorokhovatskyi, V., Chmutov, Y., Tvoroshenko, I., & Kobylin, O. (2025). Reducing computational costs by compressing the structural description in image classification methods. Advanced Information Systems, 9(1), 5–12.

8. Gorokhovatskyi V., Tvoroshenko I., Yakovleva O., and Hudáková M. (2025) Image description compression in classification structural methods, IEEE Access, vol. 13, pp. 43631-43641, doi: 10.1109/ACCESS.2025.3548910.

9. Gorokhovatskyi V., Tvoroshenko I. (2024) An effective method for transforming an image description into a compact vector for classification. Information Technology and Implementation (Satellite): Conference Proceedings, November 21, 2024, Kyiv, Ukraine / V. Snytyuk (Editor). – Kyiv: Publishing House «Caravela», 25-28.