УДК 004.934:811.521'243

АВТОМАТИЗАЦІЯ ПРОЦЕСУ ВИВЧЕННЯ ЯПОНСЬКОЇ МОВИ НА ОСНОВІ МЕТОДІВ ОБРОБКИ ПРИРОДНОЇ МОВИ

Руднєв К.В.
e-mail: kyrylo.rudniev@nure.ua
Науковий керівник – к.т.н., доц. Афанасьєва І.В.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

This work is devoted to the research of automated Japanese text processing methods to generate optimal flashcards for language learning. The study addresses the complexity of Japanese scriptio continua and agglutinative grammar by comparing various NLP tools, such as MeCab, Juman++, GiNZA, LLMs, and SudachiPy. Based on multi-criteria decision-making, SudachiPy was selected as the optimal tokenizer due to its balance of speed, accuracy, and flexible split modes. Furthermore, the system implements the TF-IDF metric for vocabulary prioritization and integrates with the modern FSRS spaced repetition algorithm. This approach significantly reduces the routine burden of "sentence mining" and enhances the overall efficiency of immersion-based language acquisition.

Сучасний стан розвитку технологій навчання іноземних мов характеризується переходом до імерсивних підходів. Японська мова класифікується як одна з найскладніших для вивчення, що вимагає тисяч годин активного навчання. Сучасна парадигма "Immersion Learning" наголошує на важливості засвоєння матеріалу через масове споживання автентичного контенту, що базується на гіпотезі Стівена Крашена. Однак практична реалізація цього підходу створює значне навантаження на студента. Процес ручного створення флеш-карток для систем інтервального повторення, відомий як "sentence mining", займає від 2 до 5 хвилин на одну картку. При необхідності вивчення великого масиву лексики це перетворюється на сотні годин рутинної роботи.

Стрімкий розвиток цифрових технологій та глобалізація інформаційного простору призвели до безпрецедентної доступності автентичних іншомовних матеріалів. Однак наявність самого лише "сирого" тексту не гарантує його успішного засвоєння. Трансформація неструктурованого масиву даних у дидактично корисний матеріал залишається вузьким місцем у самостійному навчанні.

Крім того, традиційний підхід до читання іноземною мовою зі словником створює значне когнітивне перевантаження. Студент змушений постійно переривати стан "потоку" для пошуку перекладу невідомих слів. Таке перемикання контексту не лише знижує швидкість читання, але й негативно впливає на загальну мотивацію та рівень розуміння прочитаного.

Системи інтервального повторення довели свою високу ефективність у довгостроковому утриманні інформації в пам'яті, проте їхній головний недолік полягає в проблемі "холодного старту". Алгоритми SRS чудово оптимізують графік повторення вже існуючих карток, але етап їх первинного створення залишається повністю на плечах користувача.

Метою даного дослідження є проведення комплексного порівняльного аналізу сучасних методів NLP для японської мови та наукове обґрунтування вибору оптимального технологічного стека для створення системи автоматичної генерації флеш-карток. Отже, актуальним є створення автоматизованої системи, яка б інтелектуально вилучала слова з тексту та пріоритизувала їх [1].

Ключовою технічною проблемою, що стоїть на заваді простої автоматизації, є специфіка японської писемності. На відміну від більшості західних мов, японська мова використовує scriptio continua - безперервне написання без пробілів. Для комп'ютерного алгоритму без контекстного розуміння розбиття тексту на слова є складним завданням. Додаткову складність вносить аглютинативна природа японської граматики. Дієслова та прикметники приєднують до незмінного кореня ланцюжки суфіксів, які модифікують значення. Наприклад, слово «ikitanakatta» складається з кореня та трьох суфіксів, і вимагає складного процесу лематизації для зведення до словникової форми «iku». Також існує проблема гранулярності при розбитті складних слів. Термін «外国人参政権» (виборче право іноземців) класичні аналізатори часто розбивають на базові ієрогліфи, що руйнує семантичний зв'язок. Для навчання найбільш ефективним є збереження терміну як єдиної лексичної одиниці.

Існуючі інструменти управління станами навчання мають суттєві обмеження. Браузерні розширення на кшталт Yomitan залишають процес повністю ручним та синхронним. Плагіни типу MorphMan базуються на застарілих версіях аналізаторів і часто роблять помилки в лематизації. Комерційні платформи зі штучним інтелектом є закритими системами, які не дозволяють експортувати дані без втрат та погано працюють зі специфічною термінологією.

Для розв'язання задачі морфологічного аналізу було досліджено п'ять альтернатив: MeCab, SudachiPy, GiNZA [2], Juman++ та Великі мовні моделі (LLM) [3].

MeCab є надзвичайно швидким (до 60 000 речень на секунду), але схильний до надмірної сегментації.

Juman++ забезпечує найвищу точність завдяки рекурентним нейромережам, але має критично низьку швидкість роботи (близько 200 речень/сек).

Великі мовні моделі здатні працювати без словників, але часто генерують "галюцинації", вигадуючи невірні читання або значення.

За допомогою методу лінійної адитивної згортки було проведено математичне моделювання вибору. Критеріями оцінювання виступили:

точність (K_1), якість лематизації (K_2), гнучкість гранулярності (K_3), швидкість обробки (K_4), вартість (K_5) та автономність (K_6). Найвищий бал отримав Juman++ ($U = 0.795$) Проте, враховуючи необхідність обробки великих обсягів тексту в реальному часі, для практичної реалізації системи було обрано SudachiPy ($U = 0.773$). Цей аналізатор пропонує унікальну функцію – три режими розбиття тексту (Split Modes), де режим "Mode C" дозволяє зберігати складні терміни як єдине ціле.

Просте виділення всіх невідомих слів з тексту є неефективним. Запропоновано гібридний підхід: використання частотних списків корпусу BCCWJ для відсіювання занадто рідкісної лексики, та метрики TF-IDF (Term Frequency - Inverse Document Frequency) для виявлення ключової термінології конкретного тексту. Математична модель пріоритизації має вигляд:

$$P_w = \alpha \cdot \frac{1}{loglog\left(Rank_{BCCWJ}\right)} + \beta \cdot TF \cdot IDF,$$

де α та β – вагові коефіцієнти балансу між загальною та специфічною лексикою.

Для ефективного засвоєння відібраних слів використовується новітній алгоритм інтервальних повторень FSRS, заснований на моделі пам'яті DSR (Складність, Стабільність, Згадуваність) [4]. Формула прогнозу ймовірності згадування (R) у момент часу t:

$$R(t) = (1 + F \cdot t \cdot S^{-1})^{-1},$$

де F – фактор забудькуватості користувача. На відміну від застарілих алгоритмів, FSRS дозволяє системі автоматично призначати частотним словам довші початкові інтервали, а складним – коротші, оптимізуючи загальне навантаження.

Проведене дослідження доводить, що використання аналізатора SudachiPy у поєднанні з метрикою TF-IDF та алгоритмом пам'яті FSRS дозволяє створити високоефективну систему для автоматизації генерації навчального контенту. Такий підхід значно зменшує рутинне навантаження на студентів, максимізуючи педагогічну ефективність.

Список використаних джерел:

1. Horbenko V., Onyshchenko K., Afanasieva I., Kameniev R. Analysis of existing methods and models of in-depth learning in the problems of natural language processing. БИОНИКА ИНТЕЛЛЕКТА. 2021. Vol.2. С 33–38.
2. Free Spaced Repetition Scheduler. URL: https://github.com/open-spaced-repetition/fsrs4anki/wiki/The-Algorithm (дата звернення: 09.03.2026).
3. Labs M. GiNZA – Japanese NLP Library. URL: https://megagonlabs.github.io/ginza/ (дата звернення: 09.03.2026).
4. Rusli A., Shishido M. An Experimental Evaluation of Japanese Tokenizers for Sentiment-Based Text Classification. URL: https://arxiv.org/abs/2412.17361 (дата звернення: 09.03.2026).