

УДК 004.91(004.62)

*Посторонка В.М., Довженко Д.В., Дегтяр Ю.В.
Національний авіаційний університет*

ТЕХНОЛОГІЇ ПЕРЕВІРКИ ПРАВОПИСУ НАДРУКОВАНОГО ТЕКСТУ

Перевірка правопису надрукованого тексту виконується з метою дослідження тексту на помилки, адже для усіх людей властиво помилятися. Виходячи з цього, постає задача для інженерів ПЗ створити ефективну систему з набором алгоритмів пошуку помилок з пропозиціями варіантів виправлення.

Базовим алгоритмом перевірки тексту є звірювання слів тексту з словником уже відомих слів застосунку. Цей словник містить лише слова та додаткову інформацію про граматичні і лексичні атрибути. Додатковим кроком може виконуватись перевірка морфології.

Наразі найпопулярнішим застосунком для перевірки правопису тексту є LanguageTool. LanguageTool є ПЗ з відкритим кодом для стилевої, орфографічної та граматичної перевірки тексту. Його можна використовувати як самостійний застосунок, в популярних текстових процесорах (Google Docs, MS Word, OpenOffice), так і на веб-сторінках.

Language Tool розбиває увесь уведений текст на речення, аналізує їх в такому вигляді. Далі ці речення розбиваються на лексеми, і виконується аналіз. Потім кожному слову задаються теги, які містять дані про використане слово. В останню чергу, виконується перевірка по правилам, зазначеним у спеціальному xml файлі, і в ще одному окремому файлі вводять правила, які комбінації лексем точно не можуть бути. Застосунок використовує базу знань, яка наповнюється спеціалістами та користувачами Інтернету.

LanguageTool використовує продукційну модель баз знань, тобто застосунок сприймає написані правила в xml як продукції і використовує винайдені правила для визначення місця помилки. Застосунок можна налаштувати для будь-якої мови.

Базовий алгоритм перевірки правопису є достатньо малоефективним, альтернативою буде застосунок з базою знань, яку можна постійно доповнювати, як LanguageTool. Такий застосунок повинен постійно перевірятись філологічними експертами, але його викори-стання буде набагато ефективніше, ніж його базовий аналог.