ПРОГАМНИЙ КОМПОНЕНТ ДЛЯ ВИЯВЛЕННЯ ТА ВИПРАВЛЕННЯ ОРФОГРАФІЧНИХ ПОМИЛОК В ТЕКСТІ

Кириченко В.В.
Науковий керівник – к.т.н., доцент Шевченко О.Л.
Харківський національний університет радіоелектроніки
(61166, Харків, просп. Науки, 14, каф. ПІ, тел. (057) 702-14-46)
e-mail: volodymyr.kyrychenko@nure.ua

This work is devoted to the review and analysis of the problem of detecting and correcting spelling errors in the documents` text of various kinds and the possibility of automating this process. The article also includes examples of systems that require a software solution, as well as some elements of the approach to solve this problem using finite state transducers and operations on them. The main purpose of this work is to identify the importance of this problem for users, provide a possible solution and describe the difference between the proposed solution and existing systems.

На сьогоднішній день, в часи стрімкої комп'ютеризації, підприємства та установи різного рівня, фахівці різних галузей та звичайні користувачі ПК широко використовують можливості електронної документації для своїх потреб. Але незалежно від певних потреб конкретних груп користувачів, всі вони можуть зіткнутися з проблемою наявності орфографічних помилок в набраних текстах та їх виправлення.

Ситуацію покращує той факт, що в процесі набору тексту документа текстовий процесор зможе допомогти користувачеві в разі виявлення словоформи, відсутньої в словнику програми. Але доволі часто людині доводиться мати справу з вже набраними текстами, в яких можуть бути допущені помилки. І таких текстів може бути доволі багато, а кожен з них може містити сотні сторінок, що унеможливлює ручну перевірку кожного документа.

Також не треба забувати про програмні системи, функції яких включають в себе аналіз електронних документів. Прикладами таких систем є системи перевірки робіт на плагіат; системи, відповідальні за класифікацію і тональний аналіз текстів. Такі системи доволі чутливі до наявності в документах помилок, наприклад, системи перевірки на плагіат можуть не розпізнати плагіат за наявності великої кількості орфографічних помилок в тексті. Через це проблема автоматизованого виявлення та виправлення помилок є досить актуальною.

Зараз на ринку присутні багато різних текстових процесорів, які мають вбудовану функцію перевірки орфографії, такі як Microsoft Office Word та LibreOffice Writer. Подібні можливості надають і різні онлайн-сервіси: Grammarly, OnlineCorrector, LanguageTool та інші. Але вони не надають можливості автоматичного виправлення помилок без втручання користувача, а лише підказують найвірогідніші варіанти для поданого контексту. Це хоч і зменшує трудомісткість роботи людини, яка має

перевіряти тексти, але все одно потребує від неї пильного контролю над процесом заміни помилкових слів на правильні.

Задача виявлення орфографічної помилки на даний час може вважатися вирішеною, так як вже складені мовні словники, які містять сотні тисяч слів та словоформ, які використовуються у певній мові. А сам процес виявлення помилки виглядає як порівняння вхідного токена зі словами словника у певному порядку.

Головний інтерес в рамках даної теми представляє задача виправлення помилки. В якості рішення поданої проблеми можна запропонувати модель, побудовану на основі кінцевих перетворювачів (finite-state transducers) та операцій над ними. Кінцевий перетворювач – один з видів кінцевих автоматів, у якого кожна дуга має два символи: вхідний і вихідний. Ця властивість дозволяє не тільки допускати або відкидати рядки, але також і перетворювати вхідну послідовність символів на вихідну. В якості міри подібності слів буде використано відстань Левенштейна, яка обчислюється як мінімальна кількість операцій вставки, видалення і заміни, необхідних для перетворення одної послідовності в іншу. Для підбору найбільш вірогідного слова будуть використані N-грами як ефективний засіб для передбачення на основі ймовірнісних моделей. Для отримання кінцевого результату потрібно виконати композицію чотирьох перетворювачів: $R = S \circ E \circ L \circ M$, де S – вихідний рядок, E – модель помилок, L – лексикон, M – мовна модель, $\circ$ – операція композиції. Модель помилок фактично є перетворювачем з переходами, які представляють операції редагування (вставка, видалення, заміна). Лексикон – це перетворювач, що з'єднує букви в слова для передачі в мовну модель. В якості мовної моделі виступає перетворювач, побудований на основі N-грам. В отриманому перетворювачеві R потрібно знайти найкоротший шлях, який і буде шуканим варіантом виправлення вихідного рядка.

Наразі на ринку присутні програмні бібліотеки, такі як OpenFST, за допомогою яких можна розв'язувати задачі виправлення помилок. Але вони надають лише базові інструменти для створення та проведення операцій над кінцевими перетворювачами. Тобто, користувачу, крім детального ознайомлення з бібліотекою, потрібно самому будувати вихідні об'єкти та реалізовувати певну логіку для отримання бажаного результату, що не дуже зручно, і для деяких користувачів може стати серйозною перешкодою.

Кінцевим результатом розробки буде створення відкритого API для виявлення та виправлення орфографічних помилок в тексті, яке дозволить користувачам автоматизувати процес виправлення помилок в документах.

Список використаних джерел:

1. Lothaire M. Applied Combinatorics on Words – Cambridge University Press, 2004, – С. 199-226.
2. OpenFst Library. URL:http://www.openfst.org/twiki/bin/view/FST/WebHome