УДК 004.421

А.Ю.МИХАЙЛЮК, Т.М.ЗАБОЛОТНЯ

Національний технічний університет України «Київський політехнічний інститут»

КОМБІНОВАНИЙ МЕТОД ВИПРАВЛЕННЯ ОРФОГРАФІЧНИХ ПОМИЛОК У ТЕКСТОВИХ ДАНИХ

В статті подано метод виправлення орфографічних помилок у текстових інформаційних ресурсах, що об'єднує переваги двох широковживаних методів автоматизованої корекції спотворених слів, у межах яких формування варіантів виправлення проводиться шляхом безсловникової генерації гіпотез або через відбір записів із словника. Для зменшення складності алгоритмів роботи відповідних програмних засобів запропоновано на етапі попереднього висунення гіпотез формувати не один, а два набори гіпотез відповідно до кожного з двох зазначених вихідних методів. Обґрунтовано доцільність реалізації описаного метода з використанням технології інтелектуальних програмних агентів, що дозволяє в багатопроцесорних та багатомашинних обчислювальних системах розпаралелювати процес відбору варіантів виправлення.

Вступ

Ефективність комп'ютерних систем, які працюють із текстовими електронними ресурсами, визначається не тільки способом виконання функцій, спрямованих на реалізацію безпосереднього призначення цих програм, а і, значною мірою, застосованими підходами до вирішення задач автоматизації обробки природномовного тексту. Одним з питань, важливість розв'язання якого з часом не зменшується, є виявлення та виправлення орфографічних помилок користувача. Причому основними критеріями вибору алгоритмів для реалізації автокоректорів сьогодні є, перш за все, показники їх точності та швидкодії, тоді як економність щодо залучення апаратних ресурсів стає другорядною.

Метою даної статті є підвищення ефективності (перш за все у сенсі швидкодії) прикладних програмних засобів виправлення орфографічних помилок.

Загальна характеристика етапів процесу автоматизованого виправлення помилок

Аналіз проблеми показав, що у сучасних програмних засобах для реалізації функції корекції помилок використовуються різні методи та їх комбінації [1, 2, 3]. Автори вважають, що в основу роботи прикладного автокоректора доцільно покласти методи, побудовані на словниковому підході, оскільки він краще за інші забезпечує високу точність виправлення орфографічних помилок [4]. Необхідною умовою ефективного використання словника є забезпечення оптимального рівня його наповненості.

Виділяють наступні етапи процесу виправлення спотворених слів [5]:

- висунення гіпотез (найбільш вірогідних кандидатів для виправлення помилки);
- ухвалення однієї з висунутих гіпотез як виправлення, що автоматично вноситься.

Висунення гіпотез може відбуватися двома шляхами.

1. Генерація варіантів виправлення згідно безсловникових методів на основі моделей природної мови, виявлених міжморфемних відношень та фонетичних закономірностей, статистичних даних тощо [6, 7]. Такі методи відомі також як методи розширення вибірки або спел-чекери (spell-checker).

2. Пошук множини варіантів виправлення слова у відповідному словниковому компоненті lingware. У цьому випадку вибір близьких за написанням слів зводиться до задачі пошуку у словнику за схожістю. Ознаки, за якими вибираються імовірні варіанти виправлення помилкового слова, є різноманітними і обираються у залежності від особливостей поставленої задачі та об'єкту обробки (наприклад, критерій альфакоду, критерій довжини слова, критерій першої літери слова тощо)[5, 6].

Висунення гіпотез може проводитися ітераційно, з кожним кроком звужуючи результуючий набір слів, або за допомогою більш точної (але і більш трудомісткої) фільтрації результатів, яка базується на поліграмному методі [4] або морфологічному аналізі [8].

Подальший хід процесу корекції помилок визначається тим, який з двох описаних шляхів формування масиву гіпотез був обраний. Для **остаточного ухвалення варіанта виправлення** отримані гіпотези перевіряються на близькість або до еталонного написання слів із словника, або до первинного вигляду помилкового слова (див. рис.1). У першому випадку критерієм вибору гіпотези є її точний збіг із словом словника, у другому випадку – показник схожості гіпотези та первинного написання помилкового слова, котрий може бути обчислений як відстань редагування В.Левенштейна, за допомогою Q-таблиць В.Файна [5] тощо. Пунктирними лініями на рисунку показано, які дані використовуються на кожному етапі процесу корекції.

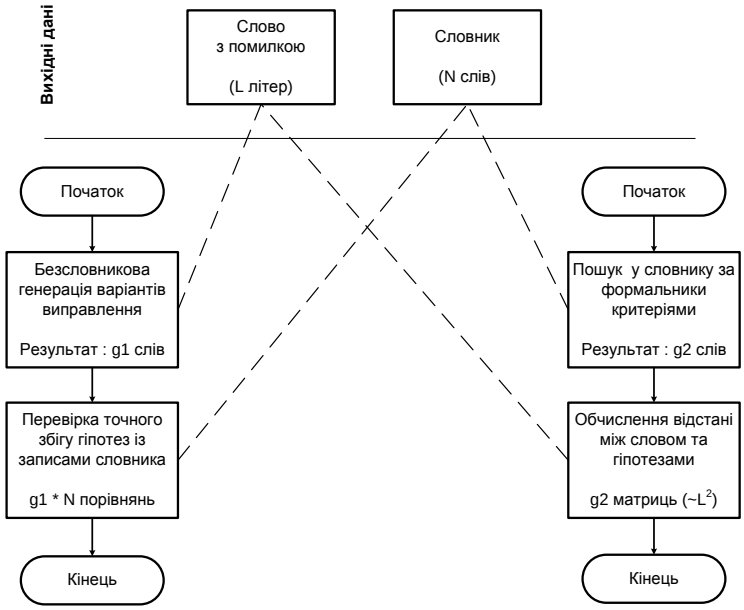

Рис. 1. Узагальнена схема варіантів процесу виправлення помилок у слові

У таблиці 1 наведені результати аналізу переваг та недоліків обох зазначених методів.

Таблиця 1. Порівняльна характеристика методів виправлення помилок

Метод	Переваги	Недоліки
Безсловникова генерація варіантів виправлення помилок	1. На етапі остаточного виправлення слова відбувається перевірка повної ідентичності (рівності) сформованої гіпотези та записів словника, а не міри схожості двох слів.	1. На наявність у словнику перевіряється велика кількість слів (гіпотез), які відсутні у природній у мові. 2. Гіпотези порівнюються з усіма записами словника без попереднього відбору останніх.

		3. Для отримання точних результатів необхідне залучення додаткових лінгвістичних та статистичних даних щодо закономірностей спільного вживання літер та їх сполучень
Пошук варіантів виправлення у словнику за формальними критеріями	1. Відібрані гіпотези є словами природної мови, тому фінальний етап корекції не містить зайвих перевірок схожості відсутніх у мові слів із вихідним словом. 2. Попередній відбір гіпотез із словника дозволяє зменшити набір слів, які на останньому етапі порівнюються із спотвореним словом.	1. На етапі остаточного виправлення слова має місце застосування складної функції визначення схожості слів, яка зазвичай потребує побудови допоміжних структур (наприклад, матриць Левенштейна).

Очевидно, що на певних етапах виправлення, де згідно першого метода виконується формування та обробка надлишкових даних та залучення при цьому додаткових ресурсів, другий метод передбачає більш простий варіант функціонування автокоректора, і навпаки. З огляду на це комбінований метод корекції з використанням переваг обох описаних методів дозволить підвищити швидкодію прикладного програмного забезпечення виправлення орфографічних помилок.

Комбінований метод виправлення орфографічних помилок

Метод, що пропонується, передбачає проведення етапів, зображених на рис.2. Останній етап тут відчутно спрощений порівняно із аналогічними етапами базових методів: немає порівнянь гіпотез із повним вмістом словника і немає потреби обчислювати складні функції схожості слів. Натомість здійснюється перевірка збігу гіпотез, отриманих двома різними шляхами. Кількість таких порівнянь відчутно менша за число порівнянь гіпотез із словником у методі із безсловниковою генерацією варіантів виправлення.

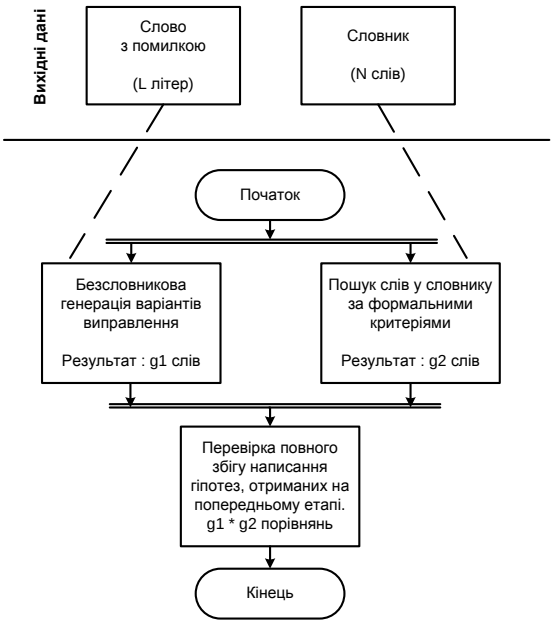

Рис. 2. Узагальнена схема процесу виправлення слів із проведенням паралельного формування двох наборів гіпотез

Для того ж, щоб зробити невідчутним для ефективності роботи програмних засобів (з точки зору швидкості виконання корекції) ускладнення процесу відбору гіпотез, автори вважають за необхідне скористатися тим, що обидві гілки етапу попереднього отримання варіантів виправлення є незалежними одна від одної. Пошук відповідних формальним ознакам словоформ проводиться виключно із залученням словникового ресурсу lingware, а безсловникова генерація гіпотез – на основі первинного написання слова з помилкою. Це дозволяє виконувати складові етапу попереднього формування масиву гіпотез паралельно в багатопроцесорній обчислювальній системі або в розподілених багатомашинних комплексах.

Оцінка ефективності запропонованого метода автоматизованої корекції помилок

Аналіз ефективності провадився, виходячи з наступних припущень:

- відбір варіантів виправлення із словника здійснюється за критерієм довжини слів;
- для генерації гіпотез проводиться повний перебір усіх типів помилок на усіх можливих позиціях помилкового слова;
- поріг припустимої кількості помилок встановлено рівним двом.

Було проведено порівняння складності алгоритмів, які забезпечують реалізацію трьох методів: із безсловниковою генерацією варіантів виправлення, з відбором словникових записів як гіпотез виправлення (використовуючи функцію Левенштейна на етапі остаточного виправлення) та запропонованого у статті комбінованого метода.

1. Метод корекції із безсловниковою генерацією варіантів виправлення на етапі висунення гіпотез. Кількість згенерованих варіантів виправлення однократної помилки для слова довжиною L символів (за умови, що алфавіт даної мови має N літер) налічує $A = 2NL + N + L - 1$ слів [8]. Для помилок більшої, ніж 1, кратності залежність кількості варіантів виправлення від довжини слова, обсягу алфавіту та кількості типів помилок можна приблизно оцінити як:

$$A = 3^m C_L^m N^m \frac{2}{3} \qquad (1)$$

де m – кількість помилок у слові;

L – довжина слова;

N – кількість літер у алфавіті;

3^m - кількість комбінацій типів помилок;

C_L^m - кількість комбінацій позицій у слові, де могли б бути допущені помилки;

$\frac{2}{3} N^m$ - кількість комбінацій літер заданого алфавіту, які вносяться як виправлення помилок у слові.

Множник $\frac{2}{3}$ сюди включено, оскільки одна з трьох операцій корекції (операція видалення літер) слова не передбачає вставки інших літер на місце видалених.

На останньому етапі маємо $SL * A$ перевірок (SL – обсяг словника). При значній наповненості словника для виправлення однієї помилки необхідно виконати мільйони порівнянь, що є неприйнятним.

2. Метод корекції із відбором варіантів виправлення із словника на етапі висунення гіпотез. Кількість вибраних із словника за формальними ознаками слів, і відповідно час здійснення даної операції, знаходяться у експоненціальній залежності від довжини слова. Це твердження базується на припущенні, що розподіл слів різної довжини у словнику близький до нормального:

$$f(L) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(L-a)^2}{2\sigma^2}} \tag{2}$$

де L – довжина слова;

a – середня довжина слів;

σ - стандартне відхилення довжини.

Для оцінки складності процедури вибору гіпотез виправлення спотвореного слова потрібно урахувати також і кількість помилок, яку необхідно виправити. Варіантами виправлення можуть бути усі слова, довжина яких відрізняється від довжини заданого слова за абсолютною величиною не більше, ніж на кількість помилок. Тому число гіпотез, вибраних із словника, можна оцінити за формулою:

$$B = \sum_{i=L-m}^{i=L+m} f(i) * SL \tag{3}$$

Для обраних слів проводиться побудова матриці розмірністю $\sim L^2$. Тут слід зазначити, що квадратична залежність кількості операцій від довжини слова, яке обробляється, є досить неточною. Згідно метода оцінки трудомісткості алгоритмів, запропонованому у [9], для побудови кожної матриці Левенштейна потрібно виконання $28L^2 + 4L + 1$ операцій. Оскільки довжина слова L рідко перевищує 15-18 літер, у межах даної задачі коефіцієнт «28», який стоїть при L^2, збільшує складність алгоритму, принаймні, на один порядок. Тому ним неможна знехтувати. Отже, кількість операцій, які необхідно здійснити на останньому етапі даного алгоритму обчислюється як:

$$B = (28L^2 + 4L + 1) \sum_{i=L-m}^{i=L+m} f(i) * SL \tag{4}$$

3. Комбінований метод корекції. В результаті застосування метода із безсловниковою генерацією гіпотез ми отримуємо $A = 3^m C_L^m N^m \frac{2}{3}$ слів; методом з відбором гіпотез із словника за критерієм довжини слова - $B = \sum_{i=L-m}^{i=L+m} f(i) * SL$ слів. На етапі остаточного визначення варіантів виправлення запропонованого метода корекції потрібно виконувати попарне порівняння гіпотез ($A * B$ порівнянь), отриманих обома розглянутими (див. пп.1,2) методами. Нижче наведено графічне подання залежності кількості операцій на цьому етапі, які виконуються кожним з алгоритмів, від довжини слова, яке коригується. *(Приклад розрахунку виконано за умови, що N = 2500000 словоформ; L = 10 літер; m=1 помилка.)* Як бачимо, запропонований метод потребує виконання значно меншої кількості операцій, ніж інші два.

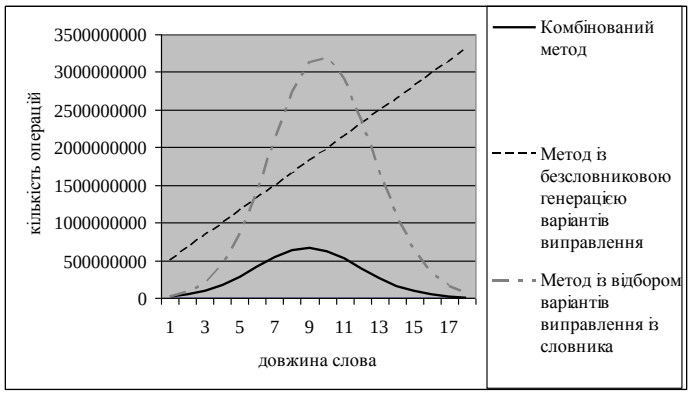

Рис. 3. Графік залежності кількості операцій на етапі остаточного вибору гіпотез від довжини слова

При виправленні двократних помилок відчутне зростання кількості гіпотез, згенерованих без залучення словника, спричиняє падіння ефективності запропонованого метода (див. (1), (3)). І хоча комбінований метод за швидкодією виявляється кращим за метод із безсловниковою генерацією гіпотез, він програє методу, що передбачає відбір варіантів виправлення із словника. Автори статті вбачають вихід із даного становища у звуженні кількості гіпотез, згенерованих без словника, наприклад, із використанням поліграмного або морфологічного контролю. Залучення більш точних способів формування гіпотез дозволить уникнути формування та обробки надлишкових даних на обох етапах методу. Якщо, згідно із статистичними фактами [8], поліграмний контроль скоротить число гіпотетичних варіантів виправлення у п’ять разів, а морфологічний – до декількох одиниць, відповідно зменшиться і тривалість роботи програмних засобів виправлення помилок.

Автокоректор загалом відноситься до класу сервісного програмного забезпечення, котре, як правило, входить до складу більш масштабних систем. У тому числі він може бути вбудований і до багатопроцесорних або багатомашинних систем автоматизованої обробки текстових даних. Тому **для розпаралелювання етапів відбору гіпотез** та для полегшення можливої інтеграції автокоректора з іншими програмними засобами пропонується **використання технології інтелектуальних агентів.**

Автори вважають за доцільне ввести до складу системи автоматизованої корекції помилок спеціальний компонент (агент), який буде відповідати за взаємодію основної програми та БД словника (див. рис. 5). Така зміна структурної організації програмних засобів дозволить основному модулю перекладати функцію відбору потрібних записів із словника на цей додатковий елемент. Для виконання даної задачі достатньо використання реактивного агента: він працюватиме на рівні стимульно – реактивних зв'язків, керуючись простими правилами «ситуація-дія» [10, 11].

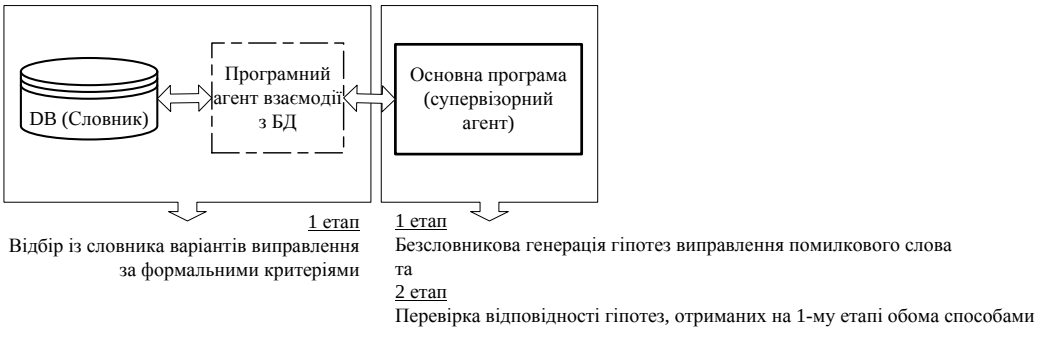

Рис.5. Узагальнена схема взаємодії програмних засобів з БД за допомогою інтелектуального агента

Синхронізація процесів формування гіпотез основним модулем та програмним агентом має відбутися перед початком етапу остаточного визначення варіантів виправлення. У випадку відсутності у агента результатів відбору записів із словника основний модуль затримає роботу до моменту надходження необхідних даних.

З огляду на те, що програмні агенти функціонують тільки у середовищі собі подібних, основний модуль, який є носієм узагальненого алгоритму роботи програми, потрібно реалізувати у вигляді інтелектуального агента, оскільки він повинен мати більш детальне уявлення про інші складові системи, до якої включено даний автокоректор.

Програмні засоби виправлення помилок, як правило, входять до складу більш масштабних систем роботи з електронними документами, які користуються не тільки лексикографічним словником. Тому для інтегрування будь-яких інших словникових ресурсів до таких систем можуть бути використані агенти подібним до описаного чином.

Висновки

Для створення програмного забезпечення автоматизованого виправлення орфографічних помилок у текстових даних обраний словниковий підхід як такий, що забезпечує високу точність роботи системи.

Запропоновано комбінований метод пошуку варіантів виправлення, який поєднує у собі позитивні риси двох відомих широковживаних методів корекції помилок. У межах комбінованого метода на етапі висунення гіпотез передбачено формування двох наборів варіантів виправлення: шляхом безсловникової генерації гіпотез та шляхом вибору записів із словника за формальними критеріями.

Взаємна незалежність процесів висунення гіпотез обома базовими методами дозволяє розпаралелити їх виконання в багатопроцесорній (багатомашинній) обчислювальній системі. При цьому досягається підвищення швидкодії відповідного програмного забезпечення.

Проведена оцінка складності алгоритмів підтвердила ефективність застосування комбінованого метода для задачі виправлення однократної помилки.

Новий метод запропоновано реалізувати на основі технології програмних агентів. Агент стає посередником між основною програмою виправлення помилок та базою даних і дозволяє виконувати безсловникове формування гіпотез та пошук варіантів виправлення слова у словнику одночасно. Крім того, залучення агентів може бути використане і для полегшення можливої інтеграції автокоректора з іншими програмними засобами.

Література

1. Лавошникова Э.К. О компьютерной коррекции «популярных» ошибок в текстах на русском языке. НТИ. 2003, №9. - с.28-34

2. Е.М.Ронин, В.И.Рублинецкий, В.А.Чикина Программа создания частотного словаря слов и выражений для русского языка. Проблемы бионики, 1999, №51, .34-43

3. Долгополов А.С. Программа автоматической коррекции текстов // НТИ, сер.2. 1986. - №4. – с.26-29

4. Большаков И. А. Проблемы автоматической коррекции текстов на флективных языках // Итоги науки и техники. Теория вероятностей. Математическая статистика. Техническая кибернетика. Т.28. — М.: ВИНИТИ,

1988. — С.111–139.

5. Машинное понимание текстов с ошибками/В.С.Файн, Л.И.Рубанов. – М.:Наука,1991. – 151с.

6. Андреевски А., Дебили Ф., Флур К. Об одном важном свойстве лексики естесственных языков и его использование при автоматическом исправлении опечаток // Прикладные и экспериментальные лингвистические процессоры, новосибирск: ВЦ СО АН СССР, 1982, с.98-109

7. Кондратюк Д. Корекція орфографічних помилок в українському тексті // Проблеми українізації комп'ютерів: Матеріали 2-ї міжнар. конф. (Львів, 29 вересня- 1 жовтня 1992 р.) / Інститут кібернетики ім. В.М.Глушкова / Р.П. Базилевич (відп.за вип.), М.М. Онопрієнко (відп.за вип.). — К., 1992. — с.51 – 55

8. Белоногов Г.Г., Дуганова И.С. и др. Экспериментальная система автоматизированного обнаружения и исправления орфографических ошибок в текстах//НТИ. Сер.2. – 1984. - №3. – с.20-22

9. Макконелл Дж. Основы современных алгоритмов/ Пер. с англ. под ред. С.К.Ландо; Доп. М.В.Ульянова. — 2-е изд., доп.. — М.: Техносфера, 2004. — 366 с.: ил.. — (Мир программирования)

10. В.П.Тарасенко, А.Ю.Михайлюк, Т.М.Заболотня Спеціалізовані інтелектуальні агенти як засіб інтеграції гетерогенного програмного забезпечення // Інформаційні технології та комп"ютерна інженерія, №3(7), 2006 - с. 96-101

11. Городецкий В.И., Грушинский М.С., Хабалов А.В. Многоагентные системы (обзор) // Новости искусственного интеллекта, №2, 1998. с.64 - 117.