

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ІНВЕРСНОГО КОНТЕКСТНО-АСОЦІАТИВНОГО МЕТОДУ АВТОМАТИЗОВАНОЇ ОРФОКОРЕКЦІЇ

Т.М.Заболотня

НТУУ «КПІ», м. Київ, Україна, tatiana104@yandex.ru

Запропоновано алгоритм, який реалізує інверсний контекстно-асоціативний метод автоматизованого виправлення орфографічних помилок. Експериментально досліджено ефективність використання інверсного контекстно-асоціативного методу за критеріями швидкодії та точності роботи відповідного програмного забезпечення. Визначені рекомендовані значення параметрів алгоритму виправлення помилок, встановлення яких забезпечує ефективну роботу програмного орфокоректора.

ЛІНГВІСТИЧНЕ ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ, АВТОМАТИЗОВАНА ОРФОКОРЕКЦІЯ ПРИРОДНОМОВНИХ ТЕКСТОВИХ ОБ'ЄКТІВ

Вступ

Від моменту виникнення систем автоматичної обробки текстів (АОТ) і до теперішнього часу їх важливою задачею є перевірка орфографії в текстових даних. З огляду на постійне збільшення обсягу електронної документації, забезпечення ефективної комп'ютеризованої обробки природномовних об'єктів неможливе без підвищення продуктивності програмних орфокоректорів, особливо за критеріями точності та швидкості обробки вхідних даних [1].

Дослідження останніх вітчизняних та зарубіжних робіт у галузі автоматичної обробки природномовних текстів показало, що підвищенню ефективності автоматизованої орфокорекції значною мірою сприяє модифікація існуючих алгоритмів виправлення помилок у напрямку залучення контекстної інформації та лексико-семантичних словникових ресурсів для її обробки. Проведення аналізу асоціативних зв'язків між словами уможливило коректне звуження множини гіпотез, які мають бути оброблені під час визначення варіантів виправлення, що позитивно впливає на ефективність функціонування систем АОТ в цілому. Але на даний момент в існуючих програмних засобах орфокорекції методика врахування контексту спотвореного слова за допомогою семантичного інструментарію є недостатньо розробленою, а отже, дослідження даного питання видається актуальним.

1. Постановка задачі

Метою даної роботи є експериментальне дослідження ефективності інверсного контекстно-асоціативного методу автоматизованої орфокорекції, сформульованого та теоретично обґрунтованого у [2].

У відповідності до поставленої мети **задачами дослідження** є:

- розробка алгоритму визначення варіантів виправлення орфографічних помилок, що реалізує інверсний контекстно-асоціативний метод орфокорекції;

- дослідження ефективності роботи побудованого програмного коректора за критеріями точності та швидкості обробки текстових даних.

2. Основні теоретичні відомості щодо забезпечення автоматизованої контекстно-асоціативної орфокорекції текстових даних

2.1. Схема автоматизованого виправлення орфографічних помилок

Загальноприйнята схема автоматизованої корекції спотвореного слова [3, 4] передбачає реалізацію:

- етапу висунення гіпотез (вірогідних варіантів виправлення помилки) і

- етапу перевірки гіпотез та ухвалення однієї (декількох) з них як виправлення, що пропонується програмою до внесення.

На першому етапі послідовно виконуються підбір первинної множини варіантів виправлення із словника та попередня фільтрація її вмісту; на другому етапі – перевірка гіпотез на подібність до спотвореного слова за певними критеріями. Методи висунення та перевірки гіпотез виправлення по своїй суті передбачають фільтрацію заданої множини слів, адже в результаті застосування кожного з них відбувається звуження поточної множини варіантів корекції спотвореного слова.

З огляду на вищенаведене будемо вважати процес визначення варіантів виправлення таким, що складається із застосування композиції функцій фільтрації до множини слів W_{dict} , яка міститься у словнику [2]. Фільтром множини W_x називатимемо функцію $filter: W_x \rightarrow W_y$, за допомогою якої з елементів W_x формується множина слів W_y , що відповідають певному критерію схожості зі спотвореним словом ($W_y \subseteq W_x$) [2].

$$filter: W_x \rightarrow W_y, W_y \subseteq W_x, \quad (1)$$

де W_x, W_y - множини природномовних слів.

Позначимо послідовність фільтрів як

$$FILTERS = f_m \circ f_{m-1} \circ \dots \circ f_i \circ \dots \circ f_2 \circ f_1: W_{dict} \rightarrow W_{retr}, m > 1, \quad (2)$$

де $f_i: W_{i-1} \rightarrow W_i$ ($i = 2, \dots, m$) – фільтр множини слів, отриманої у результаті виконання f_{i-1} (для f_1 -

множини W_{dict}); W_{retr} - множина слів, визначених коректором як можливі варіанти виправлення за ознаками їх близькості до спотвореного слова.

Під *точністю* машинної орфографічної корекції спотвореного слова будемо розуміти відношення числа запропонованих орфокоректором вірних варіантів написання слова (це одиниця або нуль) до загальної кількості підібраних слів.

$$PRECISION = \frac{|W_{corr} \cap W_{retr}|}{|W_{retr}|}, \quad (3)$$

де W_{corr} - множина вірних варіантів корекції спотвореного слова у словнику [2].

Відповідно до формули (3), для того, щоб досягти високого показника точності роботи орфокоректора, необхідно, по-перше, забезпечити постійне входження вірного слова до сформованого масиву варіантів виправлення ($|W_{corr} \cap W_{retr}|=1$), а по-друге, - зменшити загальну кількість слів, які пропонуються програмою як найбільш вірогідні кандидати виправлення помилки (W_{retr}).

2.2. Місце семантичної складової у схемі виправлення орфографічних помилок

Згідно з класичною послідовністю обробки текстових даних (морфологічний, синтаксичний та семантичний аналіз) семантичні фільтри набору гіпотез мають стояти наприкінці композиції *FILTERS* (див.(2)). Відповідно ж до сучасних тенденцій щодо зміни загальноприйнятого порядку етапів обробки текстів [3, 5] можливим є підвищення ефективності програмного орфокоректора у випадку перенесення перевірки гіпотез за семантичними критеріями ближче до початку *FILTERS*.

Теоретичні дослідження впливу зміни місця семантичної складової у схемі орфокорекції на точність та швидкодію програмного коректора показали такі результати [2]:

- зміна розташування семантичного фільтру f_{cont} у схемі орфокорекції не впливає на точність роботи коректора (*PRECISION*);

- застосування семантичного фільтру f_{cont} під час висунення гіпотез виправлення забезпечує більш швидке отримання результату роботи орфокоректора, ніж його використання для остаточної перевірки множини гіпотез за умови виконання нерівності $t_{f_{cont}}(W_{dict}) \ll t_{f_{cont}}(W_m)$, де W_m - результат фільтрації множини W_{dict} із використанням композиції функцій $f_m \circ \dots \circ f_2 \circ f_1 : W_{dict} \rightarrow W_m$ (для $i=1$ роль W_m виконує безпосередньо W_{dict}).

У даному випадку f_{cont} - це функція, яка застосовується для відбору із вихідного набору слів тих

словоформ, що узгоджені з контекстним оточенням спотвореного слова за змістом.

Факт підвищення швидкості пошуку варіантів виправлення при збереженні точності роботи орфокоректора у разі перенесення семантичної складової схеми корекції на її початковий етап [2] і ліг в основу інверсного контекстно-асоціативного методу автоматизованої орфокорекції.

2.3. Інверсний контекстно-асоціативний метод та узагальнений алгоритм автоматизованого виправлення орфографічних помилок

Для забезпечення орфокоректора даними про семантично-асоціативні зв'язки між словами природної мови використаний онтологічний словниковий ресурс у формі орієнтованого графу G , вершинами якого є лексеми природної мови, поєднані лексико-семантичними відношеннями [5, 6].

З огляду на те, що із збільшенням дистанції між вершинами графу словника сила семантичного зв'язку між ними швидко зменшується [5], слова, відстань від яких до контексту за структурою графу G перевищує певний поріг *maxdist*, вважатимемо нескінченно віддаленими від нього і не включатимемо їх до множини гіпотез.

Отже, згідно інверсного контекстно-асоціативного методу процедура автоматизованої орфокорекції являє собою послідовне виконання таких дій:

- 1) встановлення радіусу пошуку r рівним мінімально припустимому значенню;

- 2) висунення гіпотез виправлення за ознакою семантичної близькості до контекстного оточення спотвореного слова;

- 3) перевірка гіпотез виправлення на подібність до спотвореного слова за формальними ознаками;

- 4) збільшення радіусу пошуку гіпотез виправлення та перехід до п.2 даного методу у випадку, якщо, по-перше, $r < \text{maxdist}$, а по-друге, якщо на заданій відстані r від вершин графу словника, котрі відповідають словам контексту, не знайдено жодного слова, яке задовольнило б усім критеріям схожості зі спотвореним словом; в іншому разі - закінчення пошуку варіантів виправлення [2].

Важливою особливістю методу є його ітераційний характер. Він дозволяє зменшити кількість дій щодо обробки слів під час орфокорекції, і тим самим підвищити швидкість її виконання.

На основі інверсного контекстно-асоціативного методу організації роботи орфокоректора процес корекції спотворених слів пропонується упорядкувати таким чином, як це описано нижче (див. рис.1).

Крок 1. Визначення елементів контекстного оточення, які будуть враховуватися автокоректором для виправлення спотвореного слова: від того, які слова будуть відібрані з контексту на цьому кроці, залежить

хід подальшої корекції помилок та ефективність застосування методу в цілому (див. блоки 2, 3, 6 рис.1).

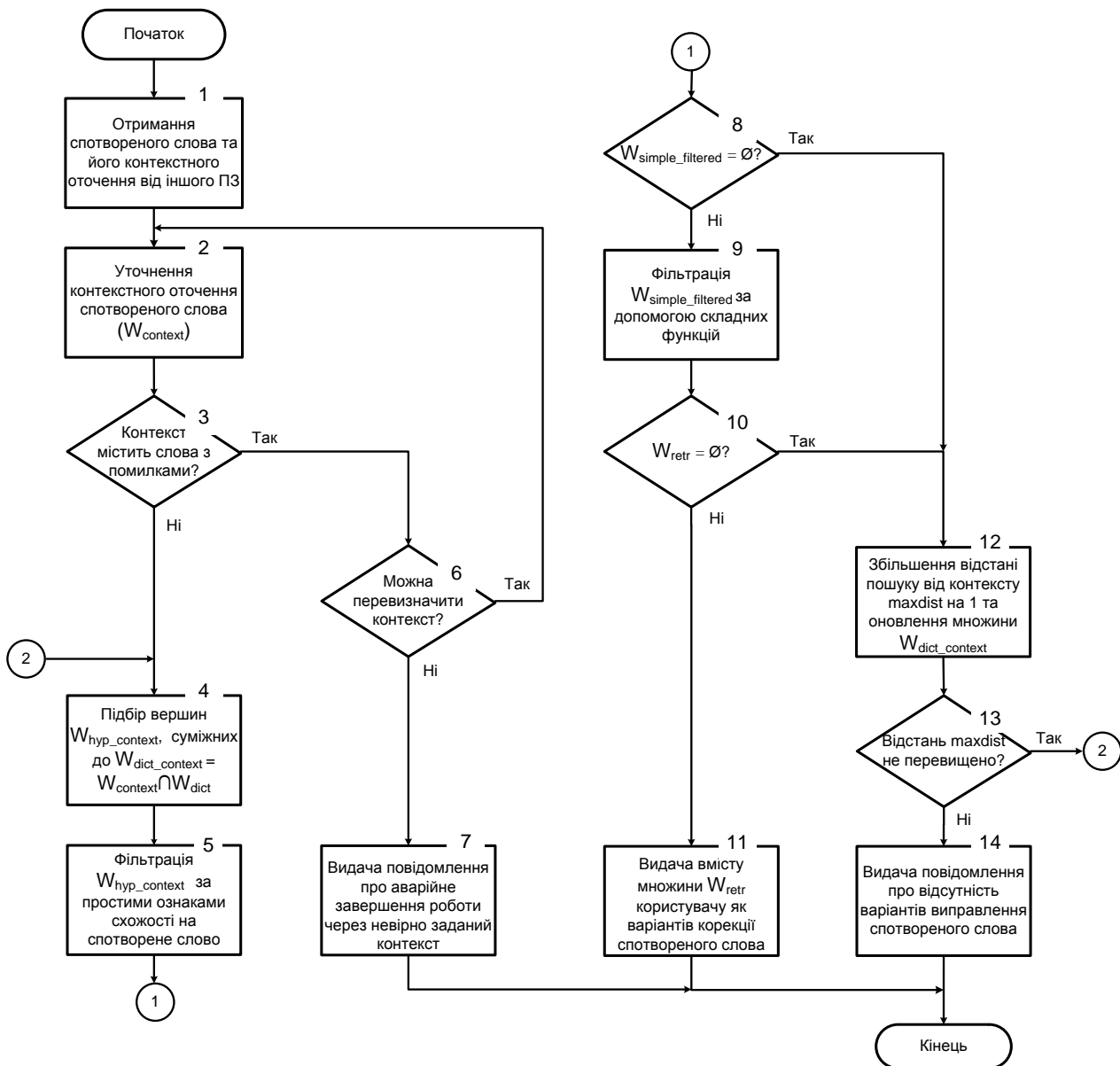


Рис. 1. Схема узагальненого алгоритму визначення варіантів виправлення

Множина слів контексту $W_{context}$ може містити слова $W_{dict_context} \subset W_{dict}$ та слова $W_{error_context}$, яких у словнику немає (будемо вважати їх такими, що містять помилку).

Якщо $W_{error_context} \neq \emptyset$ (див. блок 3 рис.1), подальший хід корекції помилок зводиться до перевірки можливості перевизначення контексту (див. блоки 2, 6 рис.1). Для систем реального часу будь-яка додаткова обробка слів контекстного оточення, які містять помилку, є неприпустимою. У такому разі умова $W_{error_context} \neq \emptyset$ є критерієм завершення процесу корекції конкретного слова.

Крок 2. Проведення підбору слів $W_{hyp_context}$, які є суміжними за структурою словника до елементів множини $W_{dict_context}$ (див. блок 4 рис.1).

$$W_{hyp_context} \subset W_{dict} \setminus W_{dict_context} \quad (4)$$

$$\forall w \in W_{hyp_context} \exists v \in W_{dict_context} : (v, w) \in E$$

Відібрані таким чином слова є близькими за змістом до слів контекстного оточення спотвореного слова $error_word$. Але крім семантичної складової, необхідно враховувати і формальний бік схожості слів. Тому наступним кроком є

Крок 3. Фільтрація $W_{hyp_context}$ за простими ознаками схожості на спотворене слово (див. блок 5 рис.1). Відбір слів відбувається за такими критеріями, які дозволяють швидко визначити, які слова точно не є

варіантами виправлення спотвореного слова. Це можуть бути критерій альфакоду, критерій довжини слова, критерій збігу першої літери слова тощо [4].

Крок 4. Ухвалення гіпотез(u) як підказки до виправлення (блок 9 рис.1). Функції, застосування яких до множини гіпотез передбачено даним кроком, дозволяють однозначно визначати слова, які схожі (згідно обраних ознак) на спотворене слово і можуть бути вірним варіантом його написання.

Якщо пошук серед сусідніх до контекстного оточення слів не дав позитивних результатів, коло можливих варіантів можна розширити $W_{dict_context} = W_{dict_context} \cup W_{hyp_context}$ і повторити кроки 2-4 алгоритму знову (див. блоки 8, 10, 12 рис.1). Даний процес триватиме, поки не будуть отримані позитивні результати ($PRECISION > 0$), або поки відстань до контексту спотвореного слова, на якій проводиться пошук варіантів виправлення у словнику, не перевищить припустимий поріг $maxdist$ (блок 13 рис.1). Ітеративний характер алгоритму, що пропонується, сприяє підвищенню швидкості роботи орфокоректора, оскільки аналізується виключно необхідний мінімум слів з околу контексту.

Перш, ніж проводити практичне дослідження ефективності методу орфокорекції з урахуванням семантично-асоціативних зв'язків слів природної мови необхідним є обрання способу кількісного оцінювання відповідності за змістом гіпотези виправлення та контекстного оточення спотвореного слова.

2.4. Кількісне оцінювання семантичної близькості гіпотези виправлення контексту спотвореного слова

У даній статті дослідження ефективності інверсного контекстно-асоціативного методу орфокорекції проводиться із застосуванням двох мір семантичної близькості варіанта виправлення контексту спотвореного слова під час автоматизованої орфокорекції:

– *Miru1*, яка використовується у практиці машинного перекладу та синтактико-семантичного аналізу текстів [5, 7] і обчислюється як обернена величина сумарної довжини найкоротших шляхів від даного слова до усіх слів контекстного оточення за структурою лексико-семантичного ресурсу;

– *Miru2*, запропонованої у [6], яка дає змогу оцінювати близькість слова та контексту з огляду на наявність зв'язку цього слова хоча б з одним елементом його оточення і є оберненою величиною мінімальної з довжин найкоротших шляхів від даного слова до елементів контексту.

Основна відмінність *Miru 1* та *Miru 2* полягає у тому, що для обчислення сумарної довжини найкоротших шляхів від слова до елементів контексту у будь-якому випадку необхідно задіяти весь окіл контексту на відстані $maxdist$ (адже, наприклад, слово,

яке має шляхи до контексту довжиною 4 та 1 згідно *Miru 1* є більш семантично близьким до нього, ніж слово, яке має шляхи довжиною 3 та 3). Процедуру ж визначення слів із *Mירוю 2* можна організувати таким чином (вибрати такий напрям відбору слів із словника), що для її успішного завершення достатньо обробити вершини графу G , які лежать у радіусі $R_{min} = \min(R)$ від контексту, де:

$$R = \{r \in \mathbb{N} \mid (r \leq \maxdist) \wedge (((\bigcup_i \alpha x_i^r \cup \alpha x_i^{-r}) \cap W_x) \neq \emptyset)\} \quad (5a)$$

$$R = \{r \in \mathbb{N} \mid (r \leq \maxdist) \wedge (F_A(\bigcup_i \alpha x_i^r \cup \alpha x_i^{-r}) \neq \emptyset)\} \quad (5b)$$

У формулі (5a) $W_x \subset W_{dict}$ - результат попереднього відбору гіпотез, які потрібно перевірити на семантичну відповідність контексту. Якщо ж процес корекції починається із пошуку семантично пов'язаних з контекстом слів, доцільно використовувати формулу 5б, де F_A - абстрактна функція перевірки елементів заданої множини слів на відповідність іншим критеріям схожості із спотвореним словом. В обох записах x_i - слово контекстного оточення; $\alpha x_i^r, \alpha x_i^{-r}$ - відображення r -го ступеня вершини x_i графу G (пряме та зворотне).

3. Аналіз результатів експериментальних досліджень ефективності інверсного контекстно-асоціативного методу автоматизованої орфокорекції

Для експериментальної апробації ефективності інверсного контекстно-асоціативного методу орфокорекції використано масив словосполучень, які є запитами користувача до інформаційно-пошукової системи [8] та характеризуються різною потужністю множин слів, що складають контекст спотвореного слова; різною кількістю помилок, припущених у слові (1 та 2 помилки); різною силою семантичного зв'язку контексту із спотвореним словом.

Для підтвердження досягнення найкращих показників роботи коректора за умови застосування фільтрів до вмісту лексико-семантичного словника в *інверсному* порядку проаналізуємо результати функціонування відповідного програмного забезпечення у випадках, коли алгоритмом його роботи передбачено:

1) використання *Miru1* та *Miru2* сили семантичного зв'язку контексту спотвореного слова та варіантів виправлення;

2) використання семантичної функції f_{cont} на початку та наприкінці послідовності фільтрів (прямий та інверсний порядок фільтрації);

3) проведення спроб виправлення одно- та двократних помилок;

4) встановлення радіуса околу контекстного оточення спотвореного слова за структурою графа словника рівним від 1 до 5 переходів ($maxdist = 1..5$).

Показники ефективності роботи створеного орфокоректора порівняно із аналогічними показниками

модуля, вбудованого до пакету MS Word, функціональність якого найчастіше використовується для обробки текстів. Робота коректора базується на використанні WordNet 3.0 для англійської мови, оскільки кінцевого варіанту локалізації словника даного формату для української мови поки що не існує.

Різноманітність вхідних текстових даних створює незручності при оцінюванні часових показників роботи коректора в секундах (чи інших абсолютних одиницях виміру часу). Тому результати вимірювань часу орфокодекції наведено у вигляді відношення часу роботи коректора при заданих вихідних умовах (t) до відповідного показника роботи коректора, коли він виконує виправлення *однократних* помилок із застосуванням фільтрів вмісту словника в *прямому* порядку ($t_{\text{оаз}}$).

$$K_t = t * 100 / t_{\text{оаз}} \quad (6)$$

Першим показником ефективності роботи програмних засобів орфокодекції, на основі значень якого можна робити висновки щодо швидкості виправлення помилок, є час, необхідний для виконання семантичної функції f_{cont} . У статті [2] показано, що для забезпечення більш швидкої роботи коректора при перенесенні f_{cont} до початку послідовності *FILTERS*, необхідним є виконання нерівності $t_{f_{\text{cont}}(w_{\text{dict}})} \leq t_{f_{\text{cont}}(w_m)}$, де W_{dict}, W_m - множини слів, які містить словник, та слів, які передаються на обробку функції f_{cont} , якщо остання стоїть наприкінці послідовності фільтрів. Перевіримо справедливості даного твердження. На графіку 2а та 2б наведені результати вимірювання часу виконання f_{cont} , в основі роботи якої лежать *Mipa1* та *Mipa 2* семантичної близькості варіантів виправлення та контексту спотвореного слова.

■ Прямий порядок, 1 помилка ■ Інверсний порядок, 1 помилка ▨ Прямий порядок, 2 помилки ▩ Інверсний порядок, 2 помилки

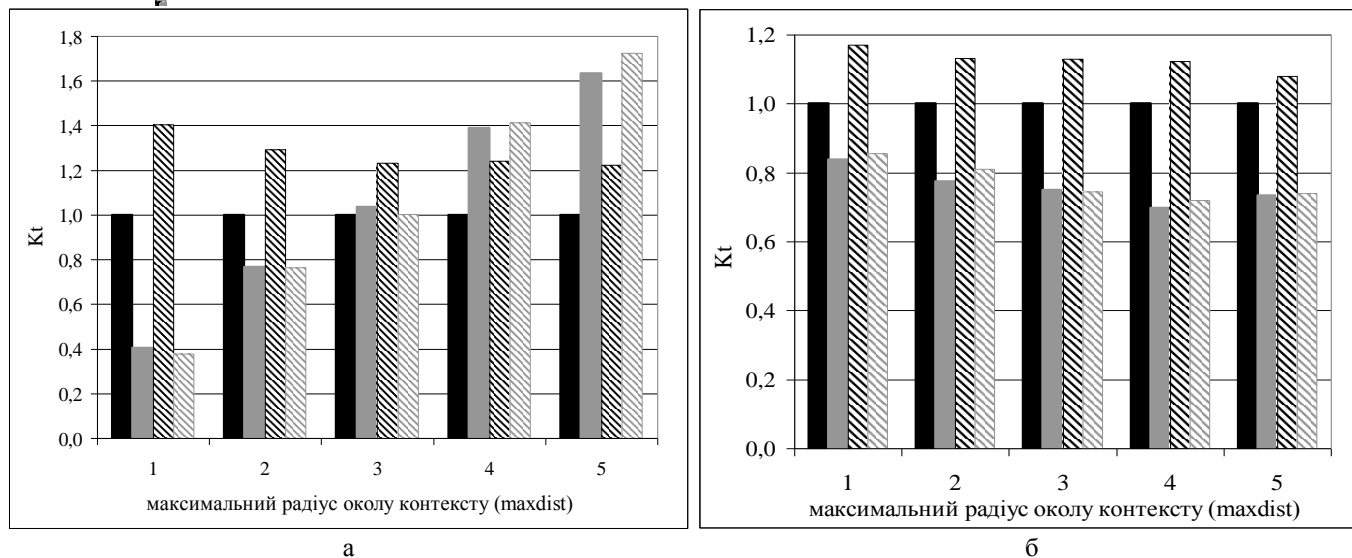


Рис. 2. Графік залежності відносного часу виконання f_{cont} від особливостей реалізації алгоритму орфокодекції (для *Mipa1* (а) та *Mipa2* (б))

Як можна побачити на графіку 2а, у межах виправлення однократних помилок при обробці околу контексту на відстані менше трьох переходів за структурою графу словника виконання функції f_{cont} є швидшим для інверсного порядку фільтрів. Дана закономірність підтверджує теоретичні відомості щодо стрімкого зменшення сили семантичного зв'язку при збільшенні дистанції між словами, яке приводить до зростання кількості нерелевантних оброблюваних гіпотез виправлення та погіршення швидкості роботи програмних засобів орфокодекції.

При спробі виправлення двократних помилок (за умови застосування фільтрів вмісту словника в інверсному порядку) час роботи f_{cont} практично збігається з аналогічними показниками, отриманими для даної функції у межах корекції слів з однократною

помилкою. При застосуванні фільтрів у прямому порядку час виконання f_{cont} для виправлення двократних помилок є більшим, ніж час роботи f_{cont} для корекції слів з однократними помилками.

На відміну від *Mipa1*, використання *Mipa2* семантичної близькості варіантів виправлення та контекстного оточення спотвореного слова при проведенні фільтрації вмісту словникового ресурсу в інверсному порядку (див. рис.2б) приводить до покращення часових характеристик роботи функції f_{cont} (порівняно із застосуванням f_{cont} наприкінці процесу фільтрації) незалежно від радіусу пошуку варіантів виправлення у графі словника.

Робота f_{cont} (з *Mipa2*) при виправленні двократних помилок у разі застосування фільтрів в інверсному порядку займає приблизно однаковий час з роботою

семантичної функції для корекції слів, які мають однократні помилки. Час же виконання f_{cont} за умови прямого порядку фільтрації у випадку виправлення двократних помилок є більшим, ніж при виправленні однократних спотворень.

Таким чином, можна зробити висновок про те, що у разі перенесення на етап висунення гіпотез семантична функція виконується швидше, ніж тоді, коли вона входить до послідовності фільтрів етапу перевірки гіпотез (для *Mipu1* та *Mipu2*). Отже, умова $t_{f_{cont}}(W_{dict}) \leq t_{f_{cont}}(W_m)$ виконана, і загальний час проведення орфокоорекції повинен бути меншим при застосуванні фільтрів до вмісту словникового ресурсу в інверсному порядку. На основі рис. 3 визначимо, чи виконується дане твердження.

■ Прямий порядок, 1 помилка ■ Інверсний порядок, 1 помилка
 ▨ Прямий порядок, 2 помилки ▨ Інверсний порядок, 2 помилки

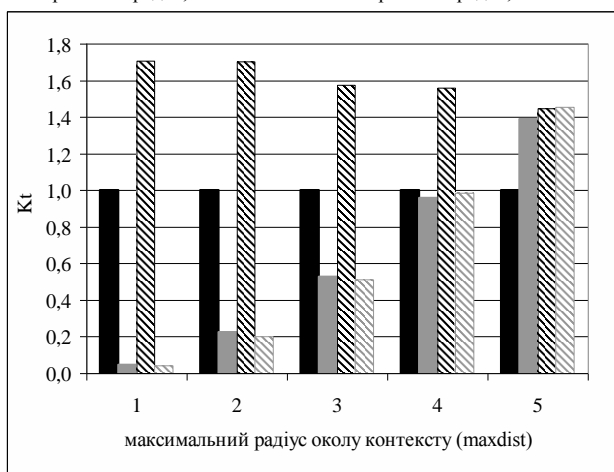


Рис. 3. Графік залежності відносного загального часу орфокоорекції від особливостей реалізації її алгоритму (для *Mipu1*)

Якщо порівняти графіки, подані на рис. 3 та 2а, можна помітити, що вони мають багато спільного: орфокооректор працює швидше у разі інверсної послідовності застосування фільтрів до вмісту словника на відстані пошуку гіпотез, яка не перевищує 4 переходи за структурою графу; час, який витрачається на проведення спроби виправлення двократних помилок, є практично однаковим з часом, потрібним для визначення варіантів виправлення однократних помилок (для інверсного порядку фільтрації).

Таким чином, виходячи з наведеного на рис. 3 графіка, можна стверджувати, що перенесення семантичної функції (яка використовує *Mipu1* близькості слів заданому контексту) на початок послідовності фільтрів вмісту словника сприяє підвищенню загальної швидкодії орфокооректора.

Всі тенденції, згадані вище для рис. 3а, також мають місце і у результатах дослідження загального часу орфокоорекції за умови використання *Mipu2*.

Порівняємо тепер найкращі часові показники роботи орфокооректора із застосуванням кожної з двох

мір семантичної близькості заданого слова до контексту та аналогічний показник роботи модуля корекції помилок, який входить до складу редактора MS Word (див. рис. 4). За $t_{баз}$ (див. (6)) прийнятий час орфокоорекції із застосуванням *Mipu1* семантичної близькості слів до контексту. Як видно з рис. 4, використання *Mipu2* забезпечує найшвидшу роботу орфокооректора, а MS Word демонструє гірший результат.

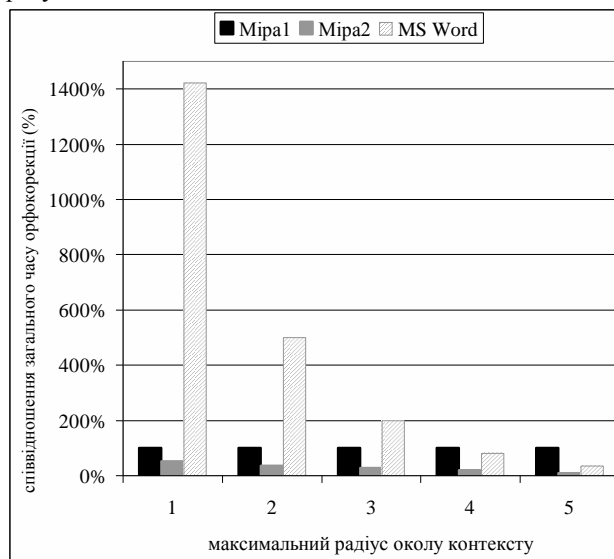


Рис. 4. Графік залежності відносного загального часу орфокоорекції від особливостей реалізації її алгоритму (для *Mipu1*)

Таким чином, на основі рис. 2–4 можна зробити висновки про досягнення найкращих часових показників роботи програмних засобів орфокоорекції за таких умов їх функціонування:

- при виконанні фільтрації вмісту словника в інверсному порядку;
- при залученні для реалізації f_{cont} *Mipu2* семантичної близькості варіантів виправлення та контексту спотвореного слова;
- при неперевищенні радіусом околу контексту, в якому здійснюється пошук варіантів виправлення, значення $maxdist=3$.

Крім того, показано, що при виконанні наведених вище умов роботи орфокооректора час на спробу виправлення однократних та двократних помилок є майже однаковим.

Проаналізуємо тепер, яким чином змінюється точність орфокоорекції в залежності від різних вихідних умов проведення останньої (див.рис.5).

Практичні дослідження підтвердили справедливості теоретично доведеної комутативності композиції функцій *filter*: від зміни порядку застосування фільтрів точність орфокоорекції не змінюється. Разом з тим, очевидно є перевага розроблених програмних засобів орфокоорекції за критерієм точності над модулем виправлення помилок MS Word. Найімовірнішою

причиною таких результатів є припущення про те, що останній не реалізує перевірки варіантів виправлення на семантичну узгодженість з контекстом спотвореного слова [1]. Так, наприклад, спотворене слово «*furm*» в залежності від контексту розробленим коректором виправляється на «*farm*» (при контексті «*animals grow*»), «*form*» (контекст – «*geometrical*») або «*firm*» (контекст – «*business*»). Модуль орфококорекції MS Word пропонує у всіх трьох випадках однакову множину варіантів виправлення: «*farm, form, firm*».

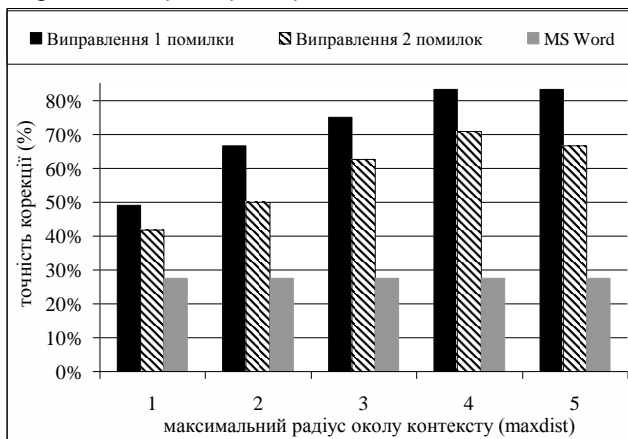


Рис.5. Графік залежності точності роботи програмних засобів від реалізації алгоритму орфококорекції

Факт, що розроблене ПЗ, яке реалізує інверсний контекстно-асоціативний метод автоматизованої орфококорекції, не характеризується 100% точністю роботи, пояснюється тим, що, з одного боку, при виправленні однократних помилок слова, які мають двократні помилки, залишаються невиправленими, а з іншого боку, спроба виправлення двократних помилок спричиняє додавання до результатів зайвих слів у випадку корекції слів, які мають однократну помилку.

Можна відмітити ще одну важливу закономірність: точність виправлення помилок перестає зростати при $maxdist > 4$. Отже, з точки зору досягнення найвищої точності роботи орфококоректора збільшувати далі радіус пошуку гіпотез недоцільно.

Якщо порівняти рівні точності орфококорекції, які забезпечують реалізації алгоритмів на основі *Miru1* та *Miru2* семантичної близькості варіантів виправлення та контексту спотвореного слова, можна зробити висновок про те, що робота коректора із залученням модифікованого словника (для виправлення одно- та двократних помилок) для обох мір є практично однаково точною (тому у статті наведений графік тільки для *Miru1* – див.рис.5).

Висновки

Таким чином, результати експериментального дослідження ефективності інверсного контекстно-асоціативного методу автоматизованої орфококорекції показали, що коректор працює краще за показниками точності та швидкості, якщо алгоритмом його роботи

передбачено застосування фільтрів до вмісту словника в *інверсному* порядку (починаючи з семантичного фільтра), а також використання *Miru2* семантичної близькості варіантів виправлення та контексту спотвореного слова для реалізації семантичного фільтра f_{cont} . Крім того емпірично отримані дані показали недоцільність перевищення відстані $maxdist=4$ при проведенні пошуку гіпотез у лексико-семантичному словнику, а також практично однакову швидкість та точність роботи коректора при спробах виправлення одно-та двократних помилок.

Отримані рекомендовані значення параметрів алгоритму виправлення помилок можуть бути уточнені при побудові програмних орфококоректорів для роботи зі спеціалізованими текстовими даними.

Список літератури

1. Лавошнікова Э.К. О компьютерной коррекции «популярных» ошибок в текстах на русском языке // НТИ, с.2. – 2003. – № 9. – с.28–34.
2. Заболотня Т.М., Михайлюк А.Ю., Михайлюк О.С. Инверсный контекстно-асоциативный метод автоматизованої орфококорекції // Искусственный интеллект – 2008. – № 3. – с.78-88.
3. Kukich K. Techniques for Automatically Correcting Words in Text // ACM Computing Surveys. – 1992. – Vol. 24, № 4. – P.377–439.
4. Файн В.С., Рубанов Л.И. Машинное понимание текстов с ошибками.–М.: Наука, 1991. – 151с.
5. Марченко О.О. Алгоритми семантичного аналізу природномовних текстів: Дис. канд. фіз.-мат. наук:01.05.01/ КНУ ім. Тараса Шевченка. – К., 2005. – 17с.
6. Заболотня Т.М. Оптимізація процесу контекстноорієнтованої орфококорекції шляхом спрощення обчислення міри семантичної близькості слів//Проблеми інформатизації та управління. Збірник наукових праць. Випуск 3(21).– К.: НАУ, 2007. – С.55-59.
7. Скороходько Э.Ф. Семантические сети и автоматическая обработка текста. – К.: Наукова думка, 1983. – 217 с.
8. Заболотня Т.М. Инверсный контекстно-асоциативный метод та програмні засоби автоматизованої орфококорекції природномовних текстових об'єктів: автореф. дис. канд. техн. наук: 05.13.05 /НТУУ "КПІ". — К., 2008. — 21с.

Надійшла до редколегії XX.XX.2009