УДК 004.421

КОНТЕКСТНО-АСОЦІАТИВНИЙ ПІДХІД ДО АВТОМАТИЗОВАНОГО ВИПРАВЛЕННЯ ОРФОГРАФІЧНИХ ПОМИЛОК

В.П. ТАРАСЕНКО, А.Ю. МИХАЙЛЮК, Т.М. ЗАБОЛОТНЯ

Показано, що введення фільтрації варіантів виправлення помилок за семантичним критерієм до схеми орфокорекції забезпечує високий рівень точності роботи програмного коректора та уможливлює його функціонування у реальному часі. Відступ від традиційної послідовності аналізу текстових даних дозволяє проводити контекстно-асоціативну обробку оточення спотвореного слова на будь-якому етапі орфокорекції. Запропоновано спосіб оцінювання точності роботи алгоритму виправлення помилок. Визначені актуальні питання реалізації контекстноорієнтованої орфокорекції, які мають бути вивчені при подальшій розробці подібних програмних засобів.

ВСТУП

Характер постановки та шляхи вирішення задачі автоматизованого виправлення орфографічних помилок у текстових даних різняться у залежності від масштабу та призначення відповідних інформаційних систем. Основою функціонування більшості автокоректорів є використання морфологічних моделей частин природної мови та результатів синтаксичного аналізу контексту слова з помилкою. Перевірка узгодженості за змістом варіантів виправлення спотвореного слова з його контекстним оточенням, як правило, не входить до функціональних профілів систем реального часу через високий ступінь складності алгоритмів її реалізації.

Сучасні досягнення у галузі створення lingware дозволяють вивести на якісно новий рівень розв'язання задачі встановлення семантичної відповідності варіантів виправлення спотвореного слова його контексту. У даній роботі доводиться доцільність використання контекстно-асоціативного підходу до відбору варіантів виправлення під час проведення орфокорекції в реальному часі, а також пропонується модифікація загальноприйнятої схеми корекції для підвищення точності виправлення орфографічних помилок прикладними програмними засобами із покращенням часових характеристик їх роботи.

СУЧАСНИЙ СТАН ПРОБЛЕМИ ПОБУДОВИ ПРОГРАМНИХ АВТОКОРЕКТОРІВ

На сьогоднішній день більшість систем автоматичної обробки текстів (АОТ), зокрема орфокоректори, працюють відповідно до класичної послідовної схеми аналізу даних (морфологічний, синтаксичний, семантичний рівні аналізу, причому «результати кожного попереднього рівня є вихідною

інформацією для наступних» [1]). Звідси перевірка семантичної узгодженості варіантів виправлення із контекстним оточенням спотвореного слова (якщо вона взагалі передбачена) має розміщуватися наприкінці алгоритму орфокорекції [2]. Але, не дивлячись на сучасний прогрес у галузі побудови lingware, розробники систем реального часу найчастіше взагалі уникають використання семантичного аналізу даних та віддають перевагу підвищенню ефективності роботи коректорів за рахунок створення нових алгоритмів формального підбору варіантів виправлення спотвореного слова. На жаль, у такий спосіб не вдається істотно покращити точність отримуваних результатів, тому коректори повертають користувачеві список усіх варіантів виправлення, які задовольняють формальним критеріям близькості слів, але за змістом не відповідають контексту [3–5]. У таких випадках остаточний вибір правильного варіанту покладається на людину.

Між тим, фахівці у галузі побудови систем АОТ наголошують на відсутності функціональної ізольованості етапів аналізу природномовного тексту. Згідно з цим морфологічний аналіз може не лише надавати вихідні дані для синтаксичного та семантичного аналізу, але й використовувати результати їх роботи [6–9]. Звідси, на думку авторів, порушення класичної схеми аналізу тексту повинно сприяти використанню у повній мірі можливостей семантичного рівня аналізу для підвищення точності та швидкості роботи програмного забезпечення виправлення орфографічних помилок.

ВИХІДНА СХЕМА АВТОМАТИЗОВАНОГО ВИПРАВЛЕННЯ ОРФОГРАФІЧНИХ ПОМИЛОК

Загальноприйнята схема автоматизованої корекції спотвореного слова [10] передбачає реалізацію таких етапів:

- висунення гіпотез (вірогідних варіантів виправлення помилки);
- перевірка 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. Для реалізації даного етапу використовуються *найпростіші* та *найшвидші* методи пошуку варіантів корекції слова (наприклад, підбір гіпотез за критерієм альфакоду, довжини слова, збігу першої літери слова тощо) [10].

На другому етапі виконується перевірка гіпотез на подібність до спотвореного слова за певними критеріями. Тут задіяні більш *складні*, але водночас і більш *точні* методи аналізу набору гіпотез (наприклад, відстань редагування В.Левенштейна) [5, 10, 11].

Таким чином, умовне віднесення методів визначення варіантів виправлення орфографічних помилок до певного етапу процесу орфокорекції здійснюється на основі їх характеристик (швидкості, точності тощо).

З іншого боку, всі методи перевірки гіпотез виправлення (на обох етапах) за своєю суттю є фільтрами заданої множини слів, адже в результаті застосування кожного з них відбувається звуження поточної множини варіантів корекції спотвореного слова. З огляду на це у даній роботі пропонується внести уточнення в подання вихідної схеми орфокорекції (рис. 1): бу-

демо вважати таким, що відноситься до етапу висунення гіпотез, тільки метод підбору гіпотез виправлення із словника; усі ж методи фільтрації множини слів, отриманої на першому етапі, перенесемо до другого етапу — етапу перевірки гіпотез.

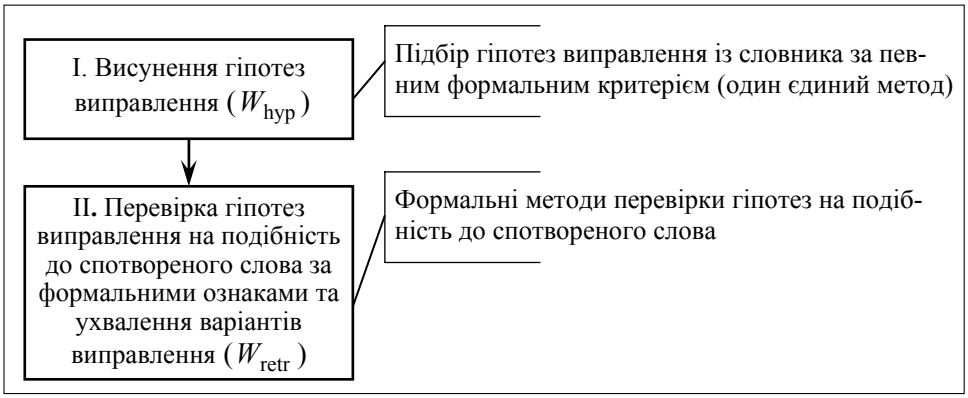

Рис. 1. Вихідна схема виправлення орфографічних помилок

Для оцінки ефективності роботи програмних засобів машинної корекції помилок введемо функцію фільтрації заданої множини слів за певною ознакою.

Визначення 1. Функція $\text{filter}: W_x \to W_y$ називається фільтром множини W_x, якщо за її допомогою з елементів W_x проводиться формування множини слів W_y, які відповідають певному критерію схожості із спотвореним словом error_word ($W_y \subseteq W_x$).

$$\text{filter}: W_x \to W_y,\ W_y \subseteq W_x, \tag{1}$$

де W_x, W_y — множини природномовних слів.

Властивості даної функції:

1) $\text{filter}(W_A \cup W_B) = \text{filter}(W_A) \cup \text{filter}(W_B)$; (2)

2) якщо $|W_A| < |W_B|$, то час, необхідний для виконання фільтрації даних множин, характеризується нерівністю

$$t_{\text{filter}(W_A)} < t_{\text{filter}(W_A)}; \tag{3}$$

3) при застосуванні композиції фільтрів $F = \text{filter}_n \circ \text{filter}_{n-1} \circ \ldots \ldots \circ \text{filter}_2 \circ \text{filter}_1 : W_x \to W_y,\ W_y \subseteq W_x$ до множини слів W_x від перестановки складових filter_i місцями результат W_y не змінюється. Тривалість виконання даних функцій, навпаки, змінюється у залежності від порядку їх застосування.

Оскільки функції, які застосовуються в межах етапу перевірки гіпотез (рис. 1), є фільтрами, їм притаманні властивості визначеної вище функції filter.

Позначимо функцію, за допомогою якої проводиться підбір гіпотез виправлення W_{hyp} зі словника, як

$$fI : W_{\text{dict}} \to W_{\text{hyp}}, W_{\text{hyp}} \subseteq W_{\text{dict}}. \tag{4}$$

Вважатимемо, що fI забезпечує висунення оптимальної (за показниками кількості слів, міри їх формальної схожості на спотворене слово error_word та швидкості отримання) множини гіпотез W_{hyp} для її ефективної перевірки на наступному етапі орфокорекції. Час, протягом якого триває виконання fI, позначимо $t_I = t_{fI(W_{\text{dict}})}$.

Фільтри, які використовуються на *етапі перевірки гіпотез*, позначимо

$$FII = fII_m \circ fII_{m-1} \circ \ldots \circ fII_i \circ \ldots \circ fII_2 \circ fII_1 : W_{\text{hyp}} \to W_{\text{retr}}, \quad m \geq 1, \qquad (5)$$

де $fII_i : WII_{i-1} \to WII_i$ $(i = 1,2,\ldots,m)$ — фільтр множини слів, отриманої у результаті виконання fII_{i-1} (для fII_1 — множини W_{hyp}); W_{retr} — множина слів, визначених коректором як можливі варіанти виправлення за формальними ознаками їх близькості до спотвореного слова.

Будемо вважати, що FII містить необхідний та достатній набір функцій, послідовне застосування яких до множини W_{hyp} забезпечує оптимальне співвідношення часу $t_{II} = t_{fII_1(W_{\text{hyp}})} + \sum_{k=2}^{m} t_{fII_k(WII_{k-1})}$, витраченого на виконання зазначених функцій та точності отриманого результату.

Оскільки визначення гіпотез виправлення здійснюється шляхом їх *пошуку* в словнику (а не за допомогою безсловникової генерації), при визначенні показників ефективності орфокорекції можна провести певні паралелі з оцінками результатів роботи програм у теорії інформаційного пошуку [12].

Визначення 2. Під *точністю* машинної орфографічної корекції спотвореного слова матимемо на увазі відношення числа запропонованих орфокоректором вірних варіантів написання слова (це одиниця або нуль) до загальної кількості підібраних слів.

$$\text{PRECISION} = \frac{|W_{\text{corr}} \cap W_{\text{retr}}|}{|W_{\text{retr}}|}, \qquad (6)$$

де W_{corr} — множина вірних варіантів корекції спотвореного слова у словнику.

Відповідно до формули (6), для того, щоб досягти високого показника точності роботи орфокоректора, необхідно, по-перше, забезпечити постійне входження вірного слова до сформованого масиву варіантів виправлення ($|W_{\text{corr}} \cap W_{\text{retr}}| = 1$), а по-друге — зменшити загальну кількість слів, які пропонуються програмою як найбільш вірогідні кандидати виправлення помилки (W_{retr}).

МІСЦЕ СЕМАНТИЧНОЇ СКЛАДОВОЇ У МОДИФІКОВАНІЙ СХЕМІ ВИПРАВЛЕННЯ ОРФОГРАФІЧНИХ ПОМИЛОК

Розглянемо можливі варіанти модифікації вихідної схеми орфокорекції шляхом введення до різних її етапів семантичної складової, а також проаналізуємо, як дані зміни вплинуть на показники точності та швидкості роботи відповідної програми.

Формування множини гіпотез виправлення за семантичним критерієм із заданого набору слів здійснюватимемо за допомогою функції f_{cont} . Визначимо дану функцію як фільтр вихідного набору слів для відбору тих лексем, що узгоджені з контекстним оточенням спотвореного слова.

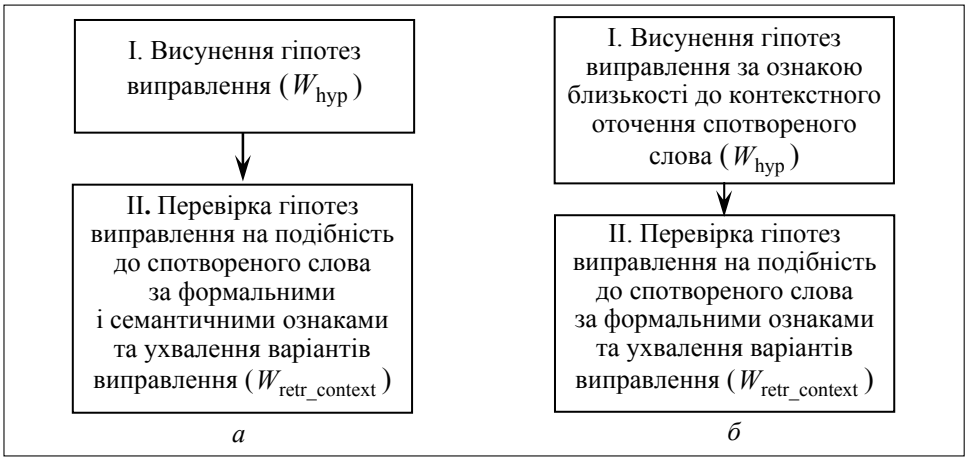

Рис. 2. Варіанти модифікації схеми виправлення орфографічних помилок

І варіант (рис. 2,*а*) **— введення контекстно-асоціативної фільтрації до етапу перевірки гіпотез виправлення.**

Оскільки склад композиції функцій FII за визначенням (5) є необхідним та достатнім для ефективної обробки гіпотез виправлення, будь-які зміни у ньому спричинять зниження ефективності роботи орфокоректора хоча б за одним із показників. Крім того, повна заміна формальних процедур перевірки слів FII семантичною f_{cont} неможлива, через те що варіанти виправлення мають відповідати як вимогам контекстної близькості, так і формальним критеріям схожості слів. Тому проаналізуємо можливість поєднання f_{cont} та FII без внесення змін до складу останньої.

$$FII' = fII_m \circ \ldots \circ fII_i \circ f_{\text{cont}} \circ fII_{i-1} \circ \ldots \circ fII_1 : W_{\text{hyp}} \to W_{\text{retr\_context}},\ m \geq 1, \quad (7)$$

де $W_{\text{retr\_context}}$ — множина слів, визначених як можливі варіанти виправлення спотвореного слова з урахуванням семантики його контексту.

Твердження 1. Введення функції f_{cont} до послідовності формальних фільтрів FII сприяє підвищенню точності роботи коректора (PRECISION).

Доведення. Нехай WII_{i-1} — результат фільтрації множини W_{hyp} із використанням композиції функцій $fII_{i-1} \circ \ldots \circ fII_2 \circ fII_1 : W_{\text{hyp}} \to WII_{i-1}$ (для $i = 1$ роль WII_{i-1} виконує безпосередньо W_{hyp}). Для FII та FII' вміст WII_{i-1} є однаковим, адже вихідна множина гіпотез і набір функцій, які до неї застосовуються, у цих двох випадках не відрізняються. f_{cont} за визначенням є фільтром, тому справедливе твердження $f_{\text{cont}} : WII_{i-1} \to WII_{\text{cont}\,i\text{-}1}$, $WII_{\text{cont}\,i\text{-}1} \subseteq WII_{i-1}$, де $WII_{\text{cont}\,i\text{-}1}$ — результат фільтрації слів з WII_{i-1} за ознакою близькості за змістом до контекстного оточення спотвореного слова.

Звідси маємо

$$WII_{\text{cont}_{i\text{-}1}} \cup \Delta W_{\text{cont\_out}} = WII_{i-1}, \tag{8}$$

де $\Delta W_{\text{cont\_out}}$ — частина множини WII_{i-1}, яка була виключена із подальшої обробки через невідповідність семантичному критерію фільтрації слів.

Перевірка множин WII_{i-1} та $WII_{\text{cont}_{i\text{-}1}}$ за допомогою функцій, які входять до складу композицій FII та FII', відповідно, проводиться, починаючи з фільтру fII_i.

$$WII_{i-1} \xrightarrow{fII_i} WII_{i,} \quad WII_i \subseteq WII_{i-1},$$

$$WII_{\text{cont}_{i\text{-}1}} \xrightarrow{fII_i} WII_{\text{cont}_i}, \quad WII_{\text{cont}_i} \subseteq WII_{\text{cont}_{i\text{-}1}}.$$

Відповідно до (2) та (8) можна записати

$$fII_i(WII_{i-1}) = fII_i(WII_{\text{cont}_{i\text{-}1}}) \cup fII_i(\Delta W_{\text{cont\_out}}) \Rightarrow$$

$$\Rightarrow WII_i = WII_{\text{cont}_i} \cup fII_i(\Delta W_{\text{cont\_out}}) \Rightarrow \left|WII_{\text{cont}_i}\right| \Leftarrow \left|WII_i\right|.$$

Застосування фільтру fII_{i+1} характеризується аналогічно.

$$WII_i \xrightarrow{fII_{i+1}} WII_{i+1,} \quad WII_{i+1} \subseteq WII_i,$$

$$WII_{\text{cont}_i} \xrightarrow{fII_{i+1}} WII_{\text{cont}_{i+1}}, \quad WII_{\text{cont}_{i+1}} \subseteq WII_{\text{cont}_i}.$$

Звідси

$$fII_{i+1}(WII_i) = fII_{i+1}(WII_{\text{cont}_i}) \cup fII_{i+1} \circ fII_i(\Delta W_{\text{cont\_out}}) \Rightarrow$$

$$\Rightarrow WII_{i+1} = WII_{\text{cont}_{i+1}} \cup fII_{i+1} \circ fII_i(\Delta W_{\text{cont\_out}}) \Rightarrow \left|WII_{\text{cont}_{i+1}}\right| <= \left|WII_{i+1}\right|.$$

У результаті отримуємо

$$fII_m(WII_{m-1}) = fII_m(WII_{\text{cont}_{m\text{-}1}}) \cup fII_m \circ fII_{m-1} \circ \ldots \circ fII_i(\Delta W_{\text{cont\_out}}) \Rightarrow$$

$$\Rightarrow W_{\text{retr}} = WII_m = WII_{\text{cont}_m} \cup fII_m \circ fII_{m-1} \circ \ldots \circ fII_i(\Delta W_{\text{cont\_out}}) =$$

$$= W_{\text{retr\_context}} \cup fII_m \circ fII_{m-1} \circ \ldots \circ fII_i(\Delta W_{\text{cont\_out}}) \Rightarrow \left|W_{\text{retr\_context}}\right| \le \left|W_{\text{retr}}\right|.$$

Отже, відповідно до (6) введення семантичної функції f_{cont} до послідовності формальних фільтрів FII забезпечує підвищення точності роботи коректора (PRECISION), завдяки проведенню більш ретельної фільтрації гіпотез виправлення, що і необхідно було довести.

Відмітимо, що місце розташування функції f_{cont} у композиції фільтрів FII, згідно з властивістю (3) функції filter, не впливає на точність роботи відповідної програми.

Проаналізуємо, як зміниться *швидкодія* машинного орфокоректора при доповненні композиції FII фільтром f_{cont}.

Твердження 2. Для збереження швидкодії даної модифікованої схеми необхідно виконати нерівність

$$t_{f_{\text{cont}}}(WII_{i-1}) \Leftarrow t_{fII_m \circ fII_{m-1} \circ \ldots \circ fII_i}(\Delta W_{\text{cont\_out}}). \tag{9}$$

Доведення. Будемо порівнювати час виконання FII та FII', починаючи від наступної за fII_{i-1} функції (fII_i та f_{cont} відповідно), адже частина $fII_{i-1} \circ ... \circ fII_2 \circ fII_1 : W_{\text{hyp}} \to WII_{i-1}$ є спільною для обох композицій. Для того щоб швидкість роботи коректора за наведеною модифікованою схемою була не нижчою за швидкість роботи вихідної схеми, має виконуватися нерівність

$$t_{f_{\text{cont}}(WII_{i-1})} + \sum_{k=i}^{m} t_{fII_k(WII_{\text{cont}_{k-1}})} \Leftarrow \sum_{k=i}^{m} t_{fII_k(WII_{k-1})} . \qquad (10)$$

Вище було доведено, що $\left|WII_{\text{cont}_k}\right| \Leftarrow \left|WII_k\right|$, де $k = i-1, i, \ldots, m$. Тому, виходячи з властивості функції filter (3), отримуємо $\sum_{k=i}^{m} t_{fII_k(WII_{\text{cont}_{k-1}})} \Leftarrow$ $\Leftarrow \sum_{k=i}^{m} t_{fII_k(WII_{k-1})}$. А на основі того, що WII_{cont_k} відрізняється від WII_k на множину $fII_k \circ ... \circ fII_{i+1} \circ fII_i(\Delta W_{\text{cont\_out}})$, де $k \geq i$, можна зробити такий висновок: час, витрачений на фільтрацію f_{cont} має бути компенсовано за рахунок того, що певна частина гіпотез з WII_{i-1} потрапила до $\Delta W_{\text{cont\_out}}$ і не буде оброблятися наступними функціями, що і відображено у (9).

Виконанню нерівності (9) сприятиме невисока (така, що не перевищує складності формальних фільтрів) складність алгоритму семантичної фільтрації.

Як наслідок даного твердження можна розглядати таку залежність: чим ближче до початку послідовності формальних фільтрів FII розташовано семантичну функцію f_{cont}, тим більше функцій входять до композиції $fII_m \circ ... \circ fII_{i+1} \circ fII_i(\Delta W_{\text{cont\_out}})$ з правої частини нерівності (9) і, отже, тим вища ймовірність успішної компенсації часу $t_{f_{\text{cont}}(WII_{i-1})}$.

Звідси розташування f_{cont} наприкінці послідовності FII (тобто $f_{\text{cont}} \circ fII_m \circ fII_{m-1} \circ ... \circ fII_1(W_{\text{hyp}}))$ не забезпечує покращення швидкодії орфокоректора, оскільки на виконання функції f_{cont} витрачається додатковий час. Отже, такий варіант модифікації схеми орфокорекції є окремим випадком введення контекстно-асоціативної фільтрації до етапу перевірки гіпотез виправлення і може бути використаний при побудові коректорів, для яких високий показник точності результатів має вищий пріоритет, ніж швидкість роботи програми.

Таким чином, введення семантичного фільтру до етапу перевірки гіпотез забезпечує підвищення точності орфокорекції, а за виконання умови (9) і прискорення роботи програми. При цьому множина гіпотез виправлення формується шляхом підбору лексем із словника за *формальною ознакою* схожості із спотвореним словом.

Розглянемо інший випадок модифікації схеми автоматизованої орфокорекції, коли визначення варіантів виправлення error_word починається з виконання f_{cont} (тобто, коли f_{cont} виконує роль fI і, відповідно, розташована на етапі висунення гіпотез).

II варіант (рис. 2,*б*) **— висунення гіпотез виправлення за ознакою семантичної близькості до контекстного оточення спотвореного слова error_word.**

Необхідно зазначити, що при введенні f_{cont} замість fI до етапу перевірки гіпотез має бути додана функція фільтрації множини варіантів виправлення за формальним критерієм, відповідно до якого виконувалося висунення гіпотез. Це обумовлено тим, що відсутність перевірки лексем за критерієм подібності до error_word, яку реалізовувала fI під час підбору гіпотез із словника, може негативно вплинути на точність роботи орфокоректора.

Окремо зупинимося на тому, що ефективне висунення гіпотез виправлення за ознакою семантичної близькості до контексту спотвореного слова можливе лише за умови використання коректором якісно укладеного лексико-семантичного словника. Цей лінгвістичний ресурс, як правило, має просту та зрозумілу форму опису знань і подається у вигляді орієнтованого графа $G = (W_{\text{dict}}, E)$, вершинами якого є лексеми природної мови W_{dict}, пов'язані між собою лексико-семантичними відношеннями з множини E [13]. Така архітектура словника відповідає принципам організації пам'яті людини, є близькою до семантичної структури природномовних фраз, а також дозволяє кількісно обчислювати міру близькості слів за змістом.

Твердження 1а. Введення семантичної функції f_{cont} до етапу висунення гіпотез схеми орфокорекції сприяє підвищенню точності роботи коректора (PRECISION).

Доведення. Відправною точкою доведення є факт, що потужність множини $W_{\text{dict\_cont}}$, отриманої шляхом аналізу вмісту словника функцією f_{cont}, є меншою, ніж вміст цілого словника, а значить можна стверджувати, що $W_{\text{dict\_cont}} \cup \Delta W_{\text{cont\_out}} = W_{\text{dict}}$. Даний вираз є подібним до (8). Звідси подальше доведення твердження 1а відбувається аналогічно до доведення твердження 1.

Твердження 2а. Для прискорення функціонування коректора, алгоритм роботи якого передбачає висунення гіпотез за ознакою семантичної близькості до контексту спотвореного слова, у порівнянні із коректором, що працює за вихідною схемою, необхідна справедливість нерівності

$$t_{f_{\text{cont}}(W_{\text{dict}})} - t_{fI(W_{\text{dict}})} \Leftarrow t_{fII_m \circ \ldots \circ fII_1(\text{W}_{\text{hyp}})} - t_{fII_m \circ \ldots \circ fII_1 \circ fI(\text{W}_{\text{hyp\_cont}})}. \quad (11)$$

Доведення. Мета, відповідно до якої ми модифікуємо схему орфокорекції, — це зменшення сумарного часу виконання етапів виправлення помилок.

$$t_{f_{\text{cont}}(W_{\text{dict}})} + t_{fII_m \circ \ldots \circ fII_1 \circ fI(\text{W}_{\text{hyp\_cont}})} \Leftarrow t_{fI(W_{\text{dict}})} + t_{fII_m \circ \ldots \circ fII_1(\text{W}_{\text{hyp}})}, \quad (12)$$

де $W_{\text{hyp\_cont}}$ — множина лексем, відібраних за семантичним критерієм із словника. Перенесення певних доданків з однієї частини нерівності до іншої дозволяє отримати запис (11), що і потрібно було довести.

Визначення позиції семантичної функції у загальній схемі машинної орфокорекції, яка забезпечила б оптимальне співвідношення швидкодії орфокоректора та рівня точності результатів виправлення, слід проводити, виходячи з того, який принцип формування набору слів, близьких за змістом до заданого контексту, покладений в основу роботи функції f_{cont}.

На основі властивості 2 функції filter, а також згідно з (6) можна зробити висновок про те, що показники швидкості та точності роботи програмного забезпечення орфокорекції залежать від потужності множин слів, які ним обробляються. Таким чином, враховуючи особливості реалізації f_{cont}, а також характеристики текстових даних та лексико-семантичних ресурсів, неважко визначити ефективний варіант модифікації схеми орфокорекції для кожного конкретного випадку.

Наприклад, у коректорі, що працює на базі лексико-семантичного ресурсу формату WordNet 3.0, а міру семантичної близькості до контексту обчислює як мінімальну з довжин найкоротших шляхів від заданого слова до елементів контексту за структурою графа G, семантичну функцію доцільно застосовувати на етапі висунення гіпотез виправлення.

Зробимо декілька зауважень відносно подальшого вивчення контекстно орієнтованого підходу до визначення варіантів виправлення спотвореного слова.

1. Залучення елементів семантичного аналізу тексту на початкових кроках процесу корекції у жодному разі не виключає подальшого проведення синтаксичного та семантичного аналізу тексту. Це пояснюється тим, що помилки, які перетворюють слово на іншу лексему, присутню у словнику, можуть бути виявлені та виправлені виключно на синтактико-семантичному рівні аналізу тексту. Отже, з точки зору розробки кінцевого програмного продукту практичний інтерес становить вивчення можливості використання допоміжних даних, отриманих під час орфокорекції на наступних кроках автоматизованої обробки тексту.

2. Вважаємо, що сферою застосування підходу до висунення гіпотез виправлення за семантичним критерієм, у межах якої він ефективний, є алгоритми роботи інформаційно-пошукових систем (ІПС).

По-перше, корекція слів у такому випадку не потребує синтаксичного узгодження варіантів виправлення, адже для ІПС важливим є визначення базової форми слова.

По-друге, у ролі контексту можуть виступати всі слова запиту. Відносно невелика кількість слів у запитах (~ 71% запитів складається з 2–4 слів [14]) не є перешкодою для застосування семантичного аналізу, тому що навіть одне вірно написане ключове слово може визначити область пошуку варіантів виправлення.

По-третє, користувач під час складання запиту до ІПС намагається вживати ключові слова, які найбільш адекватно відображають його інформаційну потребу та є максимально семантично навантаженими. Тому ймовірність швидкої та точної обробки пошукових запитів є високою.

3. Алгоритми роботи lingware часто є евристичними і базуються на емпіричних дослідженнях [7, 8]. Тому доцільно проведення практичного вивчення закономірностей у послідовності вживання типів семантичних відношень у процесі руху словником.

4. Визначення оптимальної комбінації фільтрів, використання якої покращувало б роботу орфокоректора за показниками швидкодії та точності, є багатокритеріальною задачею, що не має універсального розв'язку. Звідси її потрібно вирішувати, виходячи з конкретних умов роботи програми.

5. Для налаштування автокоректора на роботу з текстами певної предметної галузі у відповідному словнику необхідно ввести додаткове ранжування слів за критерієм відповідності їх тематиці галузі.

ВИСНОВКИ

1. Обґрунтовано доцільність відхилення від класичної схеми аналізу текстових даних у межах машинного виправлення орфографічних помилок, а отже і введення контекстно-асоціативного аналізу оточення спотвореного слова до будь-якого етапу корекції.

2. Дано визначення показника ефективності функціонування орфокоректора — точності результатів його роботи (як і швидкодія, вона залежить від кількості слів, що обробляється під час корекції).

3. Доведено факт підвищення точності та визначено умови покращення часових характеристик роботи відповідної програми при введенні до схеми орфокорекції додаткової функції відбору варіантів виправлення за семантичним критерієм. Таким чином, показана можливість реалізації семантичної складової в алгоритмах роботи орфокоректорів у реальному часі.

4. Розглянуто перспективні напрямки подальшого вивчення проблеми контекстно-асоціативного визначення варіантів виправлення спотвореного слова.

ЛІТЕРАТУРА

1. *Грязнухіна Т., Дарчук Н., Олексієнко Л.* Система автоматичного аналізу українського наукового тексту // Проблеми українізації комп'ютерів: Тези доп. наук. конф. — Л., 1991. — С. 19–20.
2. *Johannes Schaback and Fang Li.* Multi-Level Feature Extraction for Spelling Correction In Proceedings of the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data. — Hyderabad, India. — January 8, 2007. — P. 79–86.
3. *Лавошникова Э.К.* Об организации системных словарей компьютерных орфокорректоров // НТИ: Сер. 2. — 2004. — № 9. — С. 31–38.
4. *Лавошникова Э.К.* О компьютерной коррекции «популярных» ошибок в текстах на русском языке // НТИ: Сер. 2. — 2003. — № 9. — С. 28–34.
5. *Кондратюк Д.* Корекція орфографічних помилок в українському тексті // Проблеми українізації комп'ютерів: Матеріали 2-ї міжнар. конф. — Львів, 29 вересня–1 жовтня 1992 р. — Ін-т кібернетики ім. В.М. Глушкова. — Київ. — 1992. — С. 51–55.
6. *Марченко О.О.* Алгоритми семантичного аналізу природномовних текстів: Автореф. дис.канд.фіз.-мат. наук. — Київ, 2005. — 150 с.
7. *Леонтьева Н.Н.* «Политекст»: информационный анализ политических текстов // НТИ: Сер.2. — 1995. — №4. — С. 4–17.
8. *Экспериментальная* система автоматизированного обнаружения и исправления орфографических ошибок в текстах / Г.Г. Белоногов и др. // НТИ: Сер. 2. — 1984. — № 3. — С. 20–22.
9. *Бондаренко М.Ф., Осыка А.Ф.* Автоматическая обработка информации на естественном языке. — Киев: УМК ВО, 1991. — 144 с.
10. *Файн В.С., Рубанов Л.И.* Машинное понимание текстов с ошибками — М.: Наука, 1991. — 151 с.
11. *Михайлюк А.Ю., Заболотня Т.М.* Комбінований метод виправлення орфографічних помилок у текстових даних // Вісн. Хмельницького національного ун-ту. — 2007.— **2,** № 2. — С. 21–26.
12. *Пещак М.М.* Нариси з комп'ютерної лінгвістики. — Ужгород: Закарпаття, 1999. — 200 с.
13. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. — СПб.: Питер, 2000. — 384 с.
14. *Ландэ Д.В.* Поиск знаний в Internet. — М.: Диалектика, 2005. — 271 с.

Надійшла 13.12.2007