УДК 004.421

П.С.Гончаренко, Т.М.Заболотня, А.Ю.Михайлюк, В.П.Тарасенко

ОРГАНІЗАЦІЯ ПРОГРАМНОГО ОРФОКОРЕКТОРА ДЛЯ ТЕКСТООРІЄНТОВАНОЇ ІНФОРМАЦІЙНО-АНАЛІТИЧНОЇ СИСТЕМИ

Національний технічний університет України «КПІ», Київський університет імені Бориса Грінченка, м.Київ, tatiana104@yandex.ru, may-62@ukr.net

В статті досліджується проблема нарощення ефективності текстоорієнтованих інформаційно-аналітичних систем за рахунок підвищення точності та швидкості попередньої орфокорекції природномовних текстових даних. Пропонується орієнтована на інтегрування до складу інформаційно-аналітичної системи архітектура програмного агента-орфокоректора, який реалізує комбінацію контекстно-асоціативного та імовірнісного підходів до автоматизованого виправлення орфографічних помилок.

Вступ

В умовах суспільства, заснованого на знаннях, важливим чинником успішної професійної діяльності практично в будь-якій галузі стає можливість оперативного отримання достовірної інформації з глобального електронного інформаційного простору. З огляду на це, а також з врахуванням динамічності, гетерогенності, неструктурованості та несистематизованості глобального електронного інформаційного ресурсу інформаційно-аналітичні системи (ІАС), зокрема текстоорієнтовані, набувають характеру ІТ підвищеного попиту. У зв'язку з цим надзвичайно актуалізується задача розробки теоретичної бази та відповідного комп'ютерного інструментарію для ефективної реалізації всіх форм інтелектуального аналізу природномовних текстових інформаційних об'єктів, у тому числі орфокорекції [1].

Концепція побудови програмного орфокоректора

У доповіді розглядається питання побудови програмного орфокоректора, який входить до складу текстоорієнтованої ІАС. Для підтримки конкурентоспроможності таких систем їх розробляють як відкритий програмний продукт, придатний до масштабування. В основу реалізації ІАС часто буває покладений агентоорієнтований підхід, згідно якого орфокоректор, що входить до складу відкритої системи, повинен бути побудований як програмний агент, а сама система при цьому відіграє роль зовнішнього середовища [2].

Аналіз кола задач орфокоректора, а також вибір критеріїв ефективності роботи даного програмного продукту як точності та швидкості виправлення помилок [3] визначили доцільність проектування агента-орфокоректора реактивним по відношенню до зовнішньої ІАС. Для забезпечення продуктивної роботи коректора пропонується обробляти вхідні дані як за допомогою використання підключених лінгвістичних ресурсів, так і на основі накопичення власної статистики сумісного використання слів у внутрішній базі даних (БД). Кожного разу, коли агенту надходить команда на виправлення спотвореного слова, слід проводити оновлення вмісту БД на основі отриманої інформації. Таким чином, з кожним наступним виправленням агент зможе збільшувати точність своєї роботи за рахунок використання статистичної інформації, взятої з більшої кількості текстів.

Алгоритм роботи агента-орфокоректора

З огляду на вищезазначене, процес роботи орфокоректора, який функціонує на основі результатів аналізу статистично-семантичних даних, пропонується розділити на два етапи – етап тренування, який виконується одноразово перед початком роботи агента, та робочий етап.

Етап тренування:

[Крок 1] Підрахувати кількість входжень кожного слова до тренувального тексту.

[Крок 2] Для кожного слова підрахувати, скільки разів воно зустрічається поруч з кожним із слів, які знаходяться на відстані $\pm k$ слів від даного.

[Крок 3] Видалити дані щодо слів, які є неінформативними і не можуть допомогти при виборі виправлення, та зберегти статистичні дані щодо всіх інших слів.

Робочий етап.

[Крок 1] Отримати від ІАС спотворене слово та його контекст.

[Крок 2] Перевірити контекст на наявність помилок чи слів зі «стоп-списків». Якщо виявлено помилки, але контекст не можна перевизначити, робота коректора закінчується.

[Крок 3] Сформувати множину гіпотез виправлення спотвореного слова зі слів, що знаходяться на відстані редагування 1-2 від нього та містяться в словнику.

[Крок 4] На основі статистичних даних для кожного слова c_i із сформованої множини визначити значення M_i – загальну кількість входжень слова-гіпотези c_i в тренувальний текст, та m_i – число входжень слова c_i в текст в межах $\pm k$ слів контексту спотвореного слова.

[Крок 5] Відкинути контекстні слова, які не допомагають при виборі виправлення.

[Крок 6] Для кожного слова c_i з множини гіпотез обчислити ступінь семантичної близькості c_i та контексту спотвореного слова (K_i).

[Крок 7] Для кожного слова c_i обчислити $S_i = \frac{m_i}{M_i} \cdot \mathrm{P}(c_i) \cdot \frac{1}{K_i}$, де $\mathrm{P}(c_i)$ – ймовірність появи слова c_i в тексті, яка обчислюється як відношення кількості разів появи слова c_i в тексті до числа слів в останньому.

[Крок 8] Вибрати з множини гіпотез слово c_i, для якого значення виразу S_i є максимальним. Це слово будемо вважати найімовірнішим варіантом виправлення.

Структура агента-орфокоректора

У структурі орфокоректора передбачимо складові, які реалізують функції виправлення помилок, та складові, що виконують характерні для агента функції взаємодії з іншими модулями ІАС. Модулі, які відповідають за орфокорекцію, доцільно включити до складу виконавчого блоку агента. Виконання функцій взаємодії агента-коректора та інших модулів ІАС покладемо на інтерфейс із зовнішнім середовищем та координатор дій.

Координатор дій отримує розібрані вхідні повідомлення від *модуля обробки вхідних та вихідних повідомлень*. Якщо вхідна команда є службовою, модуль оновлює параметри *внутрішнього стану агента*. Якщо вхідна команда є командою на корекцію слова, координатор отримує спотворене слово та його контекст. Також модуль керує *чергою слів на виправлення*, яка необхідна тоді, коли приходить команда на виправлення слова, а виконавчий блок зайнятий виправленням попереднього слова.

Внутрішній стан агента визначається станом його модулів; переліком підключених словників; критеріями оцінки ефективності роботи агента; даними про слова, які не вдалося виправити. Ця інформація не оновлюється при кожній ініціалізації агента, тому її необхідно зберігати у БД агента.

Виконавчий модуль призначений для генерації гіпотез виправлення спотвореного слова, а також для відбору з них слів, які є найімовірнішими варіантами виправлення. Він обробляє статистичні дані про гіпотези виправлення, а також оцінює міру їх семантичної близькості до контексту.

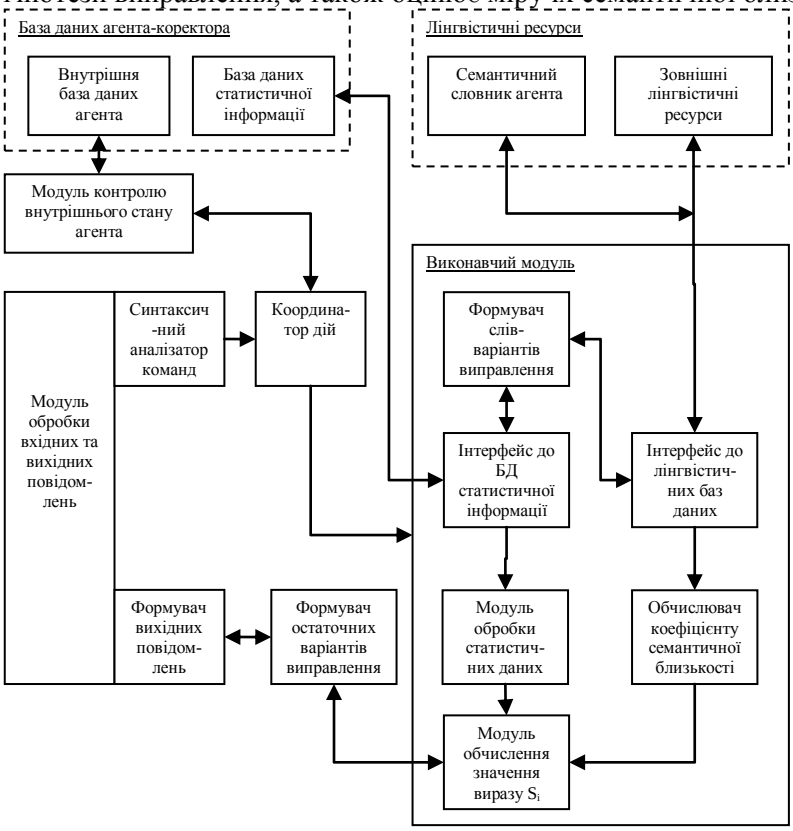

Рисунок 1 – Узагальнена структурна схема агента-орфокоректора

Альтернативні варіанти реалізації агента-орфокоректора

При реалізації орфокоректора у багатопроцесорному середовищі доцільним є забезпечення паралельного отримання та обробки ним непов'язаних між собою статистичних та семантичних даних. Після формування множини гіпотез виправлення координатор дій буде створювати два окремі незалежні потоки і в цих потоках запускати отримання та обробку статистичної та семантичної інформації. Результати своєї роботи потоки передаватимуть основному потоку, в якому формується множина остаточних варіантів виправлення. Таким чином, за рахунок розпаралелювання незалежних процесів можна підвищити швидкодію агента в багатопроцесорному середовищі.

Ще одним варіантом реалізації орфокоректора є його побудова у формі сукупності агентів. Якщо ІАС функціонує в багатокомп'ютерному середовищі, реалізація коректора у вигляді реактивного агента може знизити ефективність його роботи, оскільки швидкість виправлення помилок зменшиться через втрату часу на передачу даних каналом зв'язку. В такому випадку доцільно розробити компоненти, які відповідають за отримання і обробку статистичних та семантичних даних, у формі допоміжних агентів, кожен з яких виконуватиме свій етап роботи.

Висновки

В роботі запропоновано спосіб структурно-алгоритмічної організації програмного агента-орфокоректора, що реалізує комбінований контекстно-асоціативний метод виправлення орфографічних помилок з врахуванням накопичених під час фази тренування статистичних даних щодо сумісного використання слів, а також вмісту семантичних словникових ресурсів. Завдяки цьому значно підвищується точність орфокорекції і, як наслідок, всіх подальших етапів аналізу текстової інформації засобами ІАС.

Література

1. Тарасенко В.П., Михайлюк А.Ю., Сніжко М.В., Бігун Л.М. Функціональність спеціалізованих інформаційно-аналітичних систем для підтримки інформаційно-навчальної діяльності // Проблеми інформатизації та управління . – Зб. наук. праць. – К.: НАУ, 2009. – № 3 (27). – С. 123-130
2. Бугайченко Д.Ю., Соловьев И.П. Абстрактная архитектура интеллектуального агента и методы её реализации // Системное программирование. – СПб.: СПбГУ, 2005, №1. – С. 36–67.
3. Заболотня Т.М., Михайлюк А.Ю., Михайлюк О.С. Інверсний контекстно-асоціативний метод автоматизованої орфокорекції// "Штучний інтелект" - 2008. - №3. - с.78-88