

УДК 004.421

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

Національний технічний університет України «Київський політехнічний інститут»,
Київський університет імені Бориса Грінченка, м. Київ

tatiana104@yandex.ru, may-62@ukr.net

Організація програмного орфокоректора для текстоорієнтованої інформаційно-аналітичної системи

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 сфері стає можливість оперативного отримання з глобального електронного інформаційного простору достовірної інформації, яка чітко відповідає критерію поточного пошукового інтересу людини-фахівця [1]. При цьому слід відзначити, що з усіх типів даних найбільш актуальним в переважній більшості випадків сьогодні є природномовний текст. Однак, внаслідок швидкого розростання глобального електронного текстомісткого інформаційного ресурсу, а також з огляду на його об'єктивну гетерогенність, неструктурованість, нестабільність та несистематизованість, традиційні засоби спрямованого доступу до інформації на зразок Internet-каталогів, стандартних інформаційно-пошукових систем тощо тепер вже неспроможні задовольнити надзвичайно жорстким вимогам користувачів. У зв'язку з цим все більшої популярності на світовому ринку ІТ-продуктів набувають програмні засоби моніторингу електронного текстового ресурсу за змістом, семантично-орієнтованого пошуку та інтелектуального аналізу текстових даних – т.з. інтелектуальні інформаційно-аналітичні системи (ІАС) [2]. В залежності від галузі застосування функціональність таких систем може суттєво варіювати [3] щодо конкретного набору аналітичних операцій, котрі реалізуються відповідною системою, виходячи зокрема з таких потенційних можливостей як [4]:

- реалізація повного спектру пошукових операцій:

- повнотекстовий пошук,
- пошук за атрибутами,
- семантичний пошук,
- квазісемантичний пошук (із залученням тезаурусів, онтологій тощо),
- асоціативний пошук тощо,

- наявність механізмів ранжування пошукового відгуку (як за релевантністю, так і за пертинентністю),
 - наявність механізмів логічної компенсації дублювань у пошуковому відгуку;
- реалізація функцій змістовного аналізу текстових даних:
- структурний аналіз як структурованих, так і неструктурованих текстових інформаційних об'єктів,
 - автоматична класифікація текстових інформаційних об'єктів за відповідними ознаками (тематика, змістова тональність, авторство та інші атрибути),
 - автоматична кластеризація текстових інформаційних об'єктів,
 - автоматизоване виділення атрибутів текстових інформаційних об'єктів (змістова тональність, персоналії, асоціативні зв'язки з іншими інформаційними об'єктами тощо),
 - автоматизоване анотування текстових інформаційних об'єктів,
 - автоматизоване реферування текстових інформаційних об'єктів,
 - попереднє виявлення логічних взаємозв'язків та залежностей на змістовному рівні, виявлення причинно-наслідкових логічних зв'язків тощо,
 - оцінка рівня оригінальності тексту (виявлення плагіату, тавтології тощо),
 - комп'ютерний переклад та інформаційна підтримка навчального перекладу (словники, засоби ідентифікації граматичних конструкцій, ідіоматичних зворотів тощо);
- спрямований моніторинг інформаційного ресурсу зокрема з метою оперативного виявлення оновлень;
- можливість логічного упорядкування та агрегування гетерогенного текстомісткого інформаційного ресурсу;
- компенсація на логічному рівні дублювань у текстомістких інформаційних об'єктах.

Однак незаперечним є той факт, що обов'язковою умовою ефективної роботи практично будь-яких засобів аналітичної обробки електронного природномовного тексту є адекватна попередня орфокоорекція останнього [5].

Таким чином, виходячи із наведених вище міркувань, **метою** даного дослідження стало підвищення сумарної ефективності автоматизованого аналізу текстових природномовних електронних інформаційних об'єктів інтелектуальними інформаційно-аналітичними системами за рахунок інтегрування до їх складу програмного агента-орфокооректора підвищеної швидкодії та точності, архітектура котрого розробляється.

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

- визначення концепції побудови програмного агента-орфокооректора, придатного до інтегрування в програмне середовище текстоорієнтованої інформаційно-аналітичної системи в ході функціонального масштабування останньої;
- алгоритмізація функціонування програмного агента-орфокооректора;
- розробка логічної структури програмного забезпечення агента-орфокооректора;
- аналіз варіантів архітектурної організації програмного агента-орфокооректора.

Концепція побудови програмного орфокоректора

У статті 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 [6].

Аналіз кола задач орфокоректора, а також вибір критеріїв ефективності роботи даного програмного продукту як точності та швидкості виправлення помилок [7] 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) [8]. 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] Сформувані 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 $S_i = \frac{m_i}{M_i} \cdot P(c_i) \cdot \frac{1}{K_i}$, де $P(c_i)$ –

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

Для забезпечення можливості підключення орфокоректора до існуючих відкритих систем автоматизованої обробки текстів доцільним є наділити його властивостями програмних агентів [9]. Тому у структурі орфокоректора передбачимо складові, які реалізують функції виправлення помилок, та складові, що виконують характерні для агента функції взаємодії з іншими модулями ІАС (див. рис.1). Модулі, які відповідають за виправлення спотвореного слова, доцільно включити до складу виконавчого блоку агента [9, 10]. При цьому до бази даних слід додати таблиці лексико-семантичного словника та статистичні дані про взаємне використання слів. Виконання функцій взаємодії агента-коректора та інших модулів ІАС покладемо на інтерфейс із зовнішнім середовищем та координатор дій. Нижче більш детально розглянемо основні функції структурних елементів орфокоректора.

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.

Принцип роботи черги такий: при надходженні команди на виправлення слова модуль аналізу ситуації перевіряє значення спеціального параметру, що характеризує стан виконавчого модуля. Якщо виконавчий модуль вільний, модуль аналізу ситуації передає спотворене слово та його контекст напряму виконавчому модулю. Якщо ж виконавчий модуль в момент надходження команди на виправлення зайнятий, модуль аналізу ситуації додає відповідну команду до черги повідомлень. При цьому виконавчий модуль має спеціальний лічильник, який відображує кількість команд в черзі. Цей лічильник автоматично збільшується на одиницю при вставці команди в чергу та зменшується при вилученні команди виконавчим модулем для її обробки. Слід зазначити, що використовувати чергу повідомлень доцільно, по-перше, для виключення випадків, коли агент чекає виправлення слова і через це не може прийняти чергову команду від зовнішнього середовища, а по-друге – для того, щоб виконавчий модуль (якщо він вільний) постійно не перевіряв, чи є необроблені команди на корекцію слів, а мав відповідні дані через прийняття спеціального сигналу від модуля аналізу ситуації.

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.

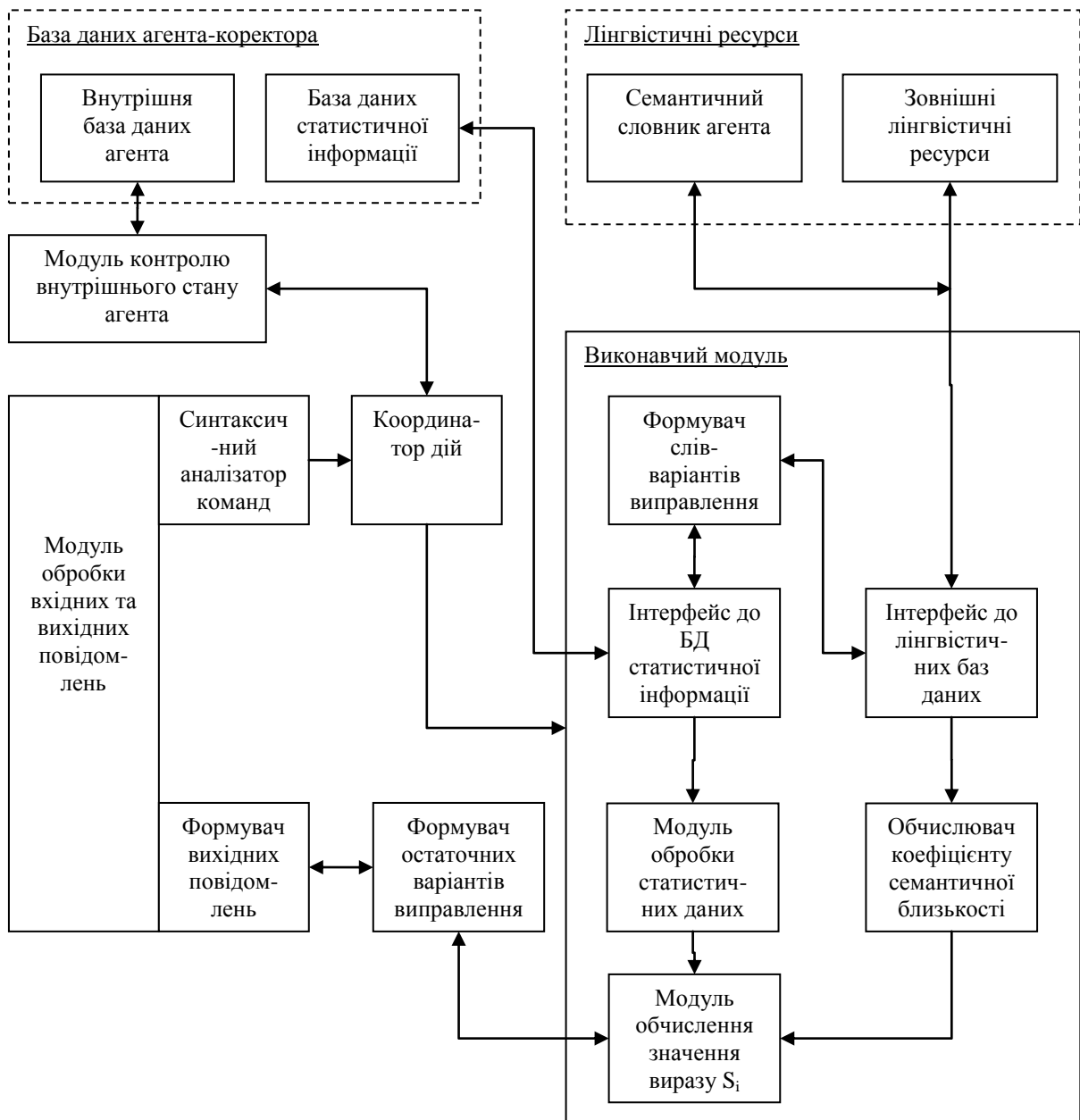


Рисунок 1 – Узагальнена структурна схема агента-орфокоректора

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, які були збережені на етапі тренування а також оцінює міру їх семантичної близькості до контексту.

Лінгвістичні ресурси. До набору програмних словникових ресурсів, які використовує агент-коректор, не потрібно включати всі словники, які можуть знадобитися при роботі з різними типами текстoorієнтованих ІАС, адже певна їх частина завжди буде невикористана. Пропонується ввести до структури

орфокоректора лише необхідний мінімум словникових ресурсів, які могли б підтримувати його працездатність, а складні та масштабні словники підключати як компонент ІАС, до якого матимуть доступ й інші модулі системи. В якості семантичного ресурсу доцільно використовувати словник, який має формат подібний до словника *WordNet* [11], оскільки він є безкоштовним, легко локалізується, має просту структурну організацію.

Альтернативні варіанти реалізації агента-орфокоректора

Реалізація агента в багатопроцесорному середовищі. Якщо текстоорієнтована ІАС, яка слугує зовнішнім середовищем для агента, функціонує на багатопроцесорному комп'ютері, доцільним є забезпечення паралельного отримання та обробки орфокоректором непов'язаних між собою статистичних та семантичних даних. В такому випадку 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 орфокоректора в багатопроцесорному середовищі.

Реалізація орфокоректора у формі сукупності агентів. Якщо ІАС функціонує в багатокомп'ютерному середовищі (наприклад, лінгвістичні ресурси такої системи розташовані на різних комп'ютерах), 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. Схема подібної реалізації агента-орфокоректора зображена на рис.2.

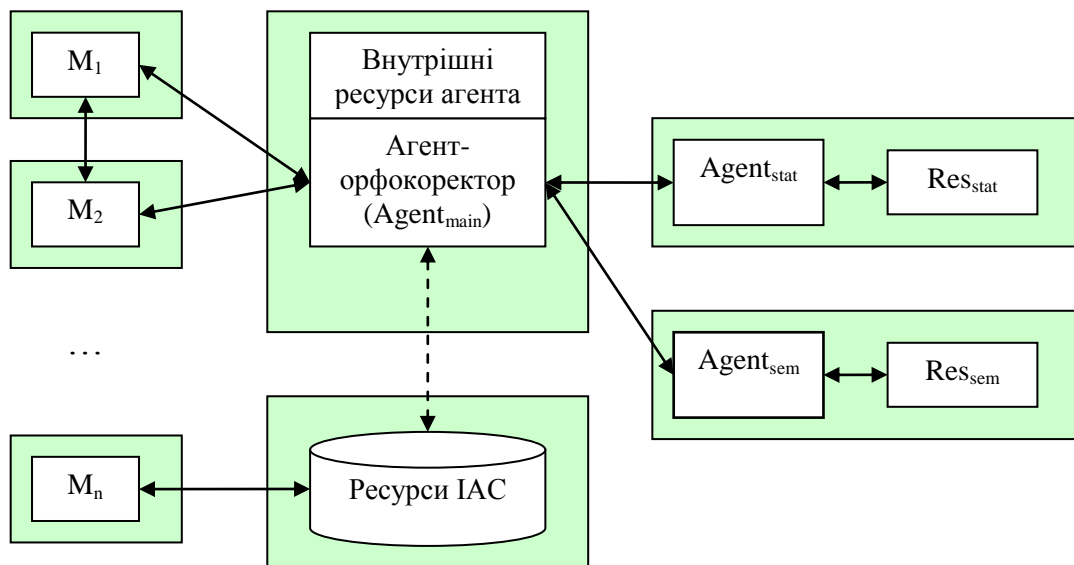


Рисунок 2 – Схема організації агента-коректора у формі сукупності агентів

Текстоорієнтована ІАС, зображена на рис.2, має розподілену структуру. Модулі системи (M_1, M_2, \dots, M_n) функціонують не на одному комп'ютері, а на багатьох. Допоміжні агенти $Agent_{stat}$ та $Agent_{sem}$, які відповідають, відповідно, за обробку статистичної інформації та за роботу із семантичними ресурсами, розташовуються окремо від головного агента $Agent_{main}$, котрий координує дії допоміжних агентів. Кожен з допоміжних агентів в разі необхідності має змогу використовувати ресурси системи через головного агента.

При такій організації в структурі агента, до якого звертається ІАС, мають залишитися лише складові, які безпосередньо відповідають за виправлення помилок, такі як модуль обчислення значення виразу S_i та формувач остаточних варіантів виправлення. Такий агент відіграватиме роль супервізора по відношенню до допоміжних агентів: він буде керувати роботою цих агентів, а також здійснюватиме формування будь-яких повідомлень до цих елементів орфокоректора. При цьому взаємодія агента-супервізора із системою обробки текстів має залишитися на реактивному рівні.

Якщо говорити про розташування структурних елементів такого розподіленого орфокоректора, то агент-супервізор доцільніше розташовувати на головному серверному комп'ютері, на якому функціонують основні модулі системи обробки текстів. Тут справа в тому, що управляючий агент постійно взаємодіє із системою, отримуючи від неї команди на виправлення слів та надаючи варіанти виправлення, і при його розміщенні на деякому віддаленому комп'ютері неодмінно будуть втрати часу на передачу цих команд, що є неприпустимим.

Що стосується допоміжних агентів, то їх доцільно розташовувати на тих серверах або робочих станціях, на яких розташовуються відповідні сховища даних, з якими працюють такі агенти (мова йде про базу даних статистичної інформації та про базу даних, в якій зберігається лексико-семантичний словник). Якщо таке розташування з будь-яких причин неможливе, необхідно намагатися розмістити допоміжні агенти таким чином, щоб між робочими станціями, на яких вони розташовані, та серверами з базами даних було якомога менше транзитних комп'ютерів. Таке розташування допоміжних агентів дозволить знизити втрати часу на отримання необхідної інформації з баз даних.

Висновки

В роботі аргументовано, що якісна робота текстоорієнтованих інтелектуальних ІАС можлива за умови наявності в їх складі ефективних засобів корекції орфографічних помилок.

Проаналізовано концепцію інтегрування програмного агента-орфокоректора до середовища відкритої текстоорієнтованої інтелектуальної ІАС.

Розроблено 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.

Запропоновано два варіанти організації агента-орфокоректора на архітектурному рівні: у формі єдиної повнофункціональної сутності та у формі сукупності вузькофункціональних програмних агентів.

Література

1. Згуровський М.З., Родіонов М.К., Жилиєв І.Б. Розвиток інформаційного суспільства в Україні: Правове регулювання у сфері інформаційних відносин.-К: НТУУ «КПІ», 2006. – 542 с.
2. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А.А.Барсегян, М.С.Куприянов, В.В.Степаненко, И.И.Холод. – СПб.:БХВ-Петербург, 2007. – 384с.
3. 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
4. Кебало О.С., Михайлюк А.Ю., Тарасенко В.П. Функціональні профілі спеціалізованих інформаційно-аналітичних систем // Науковий вісник Чернівецького університету: Збірник наук. праць. Вип. 423: Фізика. Електроніка.: Тематичний випуск «Комп'ютерні системи та компоненти». Частина I. – Чернівці: ЧНУ, 2008. – С. 117-123.
5. Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. – М.: Издательский центр "Академия", 2006. – 304 стр.
6. Бугайченко Д.Ю., Соловьев И.П. Абстрактная архитектура интеллектуального агента и методы её реализации // Системное программирование. – СПб.: СПбГУ, 2005, №1. – С. 36–67.
7. Заболотня Т.М., Михайлюк А.Ю., Михайлюк О.С. Інверсний контекстно-асоціативний метод автоматизованої орфокоорекції// "Штучний інтелект" - 2008. - №3. - с.78–88
8. Михайлюк А.Ю., Заболотня Т.М. Комбінований метод виправлення орфографічних помилок у текстових даних // Вісник Хмельницького національного університету. – 2007. – № 2, Т. 2. – С. 21-26.
9. Рассел С., Норвиг П. Искусственный интеллект: Современный подход. – М.: Издательский дом «Вильямс», 2006. – 1408 с.
10. Кльшинский Э.С. Агентные системы: классификация и применение // САПР и графика. – 1999. – № 8. – С. 90–96.
11. Wordnet – a Lexical Database for English. Princeton University, Princeton, NJ, 2001 // <http://wordnet.princeton.edu/>

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

Организация программного орфокооректора для текстоориентированной информационно-аналитической системы

В статье исследуется проблема наращивания эффективности текстоориентированных информационно-аналитических систем за счет повышения точности и скорости предварительной орфокоорекции естественных языковых текстовых данных. Предлагается ориентированная на интегрирование в состав информационно-аналитической системы архитектура программного агента-орфокооректора, который реализует комбинацию контекстно-ассоциативного и вероятностного подходов к автоматизированному исправлению орфографических ошибок.

P.Goncharenko, T.Zabolotnia, A. Mykhailyuk, V. Tarassenko

Organization of software spelling corrector for text-oriented information-analytical system

The article examines the problem of increasing the efficiency of text-oriented information-analytical systems by improving the accuracy and speed of the preliminary spelling correction of natural language text data. The architecture of software agent-corrector which implements a combination of context-associative and probabilistic approaches to the automatized correction of spelling errors is proposed. It's focused on the corrector's integration to the information-analytical system.

Статья поступила в редакцию XX.06.2010.