УДК 004.912

А. Б. Бєгунов[1], Т. М. Заболотня[2]
[1]НТУУ «КПІ», м. Київ, Україна, arxton@mail.ru
[2]НТУУ «КПІ», м. Київ, Україна, tatiana104@yandex.ru

АГЕНТНО-ОРІЄНТОВАНИЙ МЕТОД АВТОМАТИЗОВАНОГО ВИЛУЧЕННЯ ФАКТІВ З ПРИРОДНОМОВНИХ ТЕКСТОВИХ ДАНИХ

У статті наведено результати аналізу існуючих методів вилучення фактів, що засновані на цих методах. Визначено недоліки існуючих методів. Для їх усунення запропоновано агентно-орієнтований метод автоматизованого фактографічного аналізу природномовних текстових даних. Метод полягає у вилученні кожного типу фактів різними типами агентів.

МОВА ПРИРОДНА, АНАЛІЗ ФАКТОГРАФІЧНИЙ, ФАКТ, АГЕНТ, АГЕНТНО-ОРІЄНТОВАНЕ ПРОГРАМУВАННЯ

Вступ

Важливим напрямком досліджень у галузі автоматизованого аналізу природномовних текстових даних є вилучення фактів із заданого тексту (так званий фактографічний аналіз). Програмні засоби фактографічного аналізу дозволяють приводити невпорядковану інформацію до вигляду, придатного для подальшого перетворення або використання стандартними методами обробки текстових даних.

На даний момент для отримання фактів використовуються такі базові методи їх автоматизованого вилучення: методи вилучення на основі ознак; методи вилучення, які використовують функції-ядра; методи, засновані на співставленні зразків; методи, засновані на фразових зразках [1].

Слід зазначити, що переважна більшість цих методів характеризується жорсткою прив'язкою контексту вилучення до певної предметної галузі, що призводить до необхідності реалізації окремих систем вилучення фактів для різних предметних галузей. Звичайно, така розробка вимагає значних витрат часу, а також матеріальних ресурсів.

Аналіз же існуючих систем, в яких передбачено можливість налаштування на різні предметні галузі, показав, що їх істотним недоліком є складність налаштування за умови потреби у додаванні нових структур даних, що описуватимуть нові предметні галузі. Це, в свою чергу, призводить до іншої проблеми – неможливості динамічного навчання системи на наборах даних із вже заданих предметних галузей з метою підвищення точності та повноти вилучення фактів [1].

Усунення визначених вище недоліків потребує модифікації існуючих методів фактографічного аналізу. Таким чином, пошук шляхів щодо забезпечення абстрагування механізму вилучення фактів від предметної галузі оброблюваного тексту, а також щодо підтримки можливості динамічного навчання систем фактографічного аналізу видається актуальним питанням.

1. Постановка задачі

Перш, ніж сформулювати задачі дослідження, дамо визначення основних понять, якими будемо оперувати у статті.

Природномовними текстовими даними (текстом) будемо називати сукупність речень, складених будь-якою природною мовою .

Під *фактом* будемо розуміти атомарну одиницю даних в тексті, що несе інформацію про певну сутність або об'єкт [1]. Факти розділяють на види за їх змістом (події, опис, твердження, дії тощо).

Фрейм – структура, що складається з поіменованих елементів (слотів). Фрейми і слоти наділяються певною семантикою в залежності від предметної області, в якій вони використовуються. Фрейми слугують для структурованого подання фактів.

Під *фактографічним аналізом* будемо розуміти процес обробки природномовних текстових даних із застосуванням певних способів та методів вилучення фактів.

Під *текстовим доменом* будемо розуміти частину текстової інформації, що містить в собі інформацію про деякий факт.

Агент – сутність, що функціонує в певному середовищі, сприймає своє середовище за допомогою сенсорних механізмів і впливає на це середовище за допомогою виконавчих механізмів для досягнення певної мети.

Метою даного дослідження є покращення характеристик масштабованості систем вилучення фактів, їх здатності до динамічного навчання під час функціонування, а також універсалізація таких систем шляхом абстрагування від контексту певної предметній галузі.

З огляду на описану вище проблематику задачами дослідження стали:

1. Вивчення методів вилучення фактів з природномовних текстових даних на предмет виявлення можливостей щодо підвищення ефективності цього процесу за критеріями повноти та точності отримуваних результатів.

2. Розробка методу фактографічного аналізу текстових даних, реалізація якого програмним забезпеченням дозволить задовольнити такі вимоги як:

– абстрагування методу вилучення фактів від конкретної предметної галузі для підтримки налаштування вже існуючої системи на вилучення фактів будь-якого змісту;

– масштабованість системи, що забезпечить підтримку додавання нових даних для вилучення фактів відносно нових предметних галузей;

– можливість динамічного навчання системи на основі результатів, отриманих на попередніх циклах її роботи.

2. Розв'язок задачі

Нижче наведено ряд кроків, які пропонуються авторами до виконання для досягнення мети дослідження. Для кожного кроку визначимо необхідні структури даних та дії, спрямовані на вирішення поставлених задач.

2.1. Вибір способу подання фактів

Перш за все, необхідним кроком є визначення структури, що буде використовуватися для опису фактів.

На сьогоднішній день найбільш поширеною і зручною для передачі на подальшу обробку формою подання фактів в задачах вилучення знань є фрейм — структура з пойменованими елементами – слотами. Фрейми і слоти наділяються семантикою, що залежить від предметної галузі, в якій вони використовуються.

Декларативну частину знань про предметну галузь поділяють на дві групи: аксіоматичну і фактичну. У нашому випадку аксіоматика задає схему фреймового подання даних, тобто визначає склад кожного фрейма і область значень слотів. Вважатимемо, що вона задається людиною-експертом. Факти з точки зору фреймової моделі – це конкретні екземпляри фреймів із зазначеними слотами. Екземпляри фреймового подання можна подати у вигляді моделі [2]:

F – множина екземплярів фреймів;

S – множина слотів екземплярів фреймів;

T – множина значень слотів;

$R_{FST} \in F \times S \times T$ – відношення, що зв'язує з екземпляром фрейма певний слот і його значення.

Таким чином, завдання вилучення фактів зводиться до виявлення в аналізованому тексті деякого екземпляра фрейма f_i і заповнення значень його слотів $\{s_j\}$ фрагментами з тексту. Результатом вилучення є трійки $(f_i, s_j, t_k) \in R_{FST}$, де t_k – значення слота виявленого фрейма [3]. Тоді для певного фіксованого значення f_i фрейм можна подати як множину пар $\{(s_j, t_k)_1, ..., (s_j, t_k)_m\}$.

Крім концептуального рішення щодо використання фреймів для зберігання даних про вилучені факти, необхідно також визначити, які саме типи фреймів потрібні для подання довільного факту, а також коли і якими даними будуть заповнюватися їхні слоти.

Для цього спочатку проаналізуємо характер інформації, що міститься у будь-якому тексті.

В загальному випадку текст описує певні об'єкти, кожен з яких має набір характеристик. Текст також описує відношення між об'єктами та властивості цих відношень. На основі цього введемо два типи фреймів:

– фрейми першого типу – описують об'єкт; мають групу слотів для іменування заданого об'єкту та групу слотів, що містять опис об'єкта, тобто набір його властивостей. Об'єкт може не мати властивостей, тобто слоти, відповідальні за його опис, можуть бути пустими;

– фрейми другого типу – описують зв'язок одного об'єкта з іншим. Вважатимемо, що в такій парі зв'язок є однонаправленим і він завжди направлений від першого об'єкта до другого. Такі фрейми мають три слоти. Перший та другий слоти описують, відповідно, зв'язані між собою об'єкти. Третій слот описує відношення, що пов'язує ці об'єкти. Ці три слоти є фреймами першого типу. Отже, відношення також виступає в ролі об'єкта, який має назву і може мати властивості.

Проаналізуємо, якими даними можна заповнювати слоти фреймів визначених вище типів в залежності від виду виконуваної процедури вилучення інформації з природномовних текстових даних. Зазначимо, що існує 5 видів таких процедур, які відрізняються глибиною аналізу текста і типом інформації (видами фактів), що вилучаються [4].

1. Виділення іменованих сутностей – пошук та класифікація імен об'єктів в тексті. При цьому деякі об'єкти можуть бути подані в тексті групами слів. Ця процедура передбачає заповнення слотів, що відповідають за іменування об'єкта у фреймах першого типу.

2. Побудова елементів шаблону – зводиться до додавання описової інформації (значень деяких атрибутів) до об'єкта, знайденого на попередньому етапі аналізу. Ця процедура передбачає заповнення слотів, що відповідають за опис об'єкта у фреймах першого типу.

3. Розпізнавання зв'язків кореференції – визначення того, які імена об'єктів, знайдені в тексті, вказують на один і той самий об'єкт. Сюди входить як вилучення власне кореференції, тобто позначення одного і того ж об'єкта двома значущими лексичними структурами, так і аналіз анафоричних зв'язків (наприклад, коли при частому повторенні значуще слово замінюється займенником). Отримана інформація додається до об'єктів, знайдених на двох попередніх етапах аналізу. Ця процедура передбачає об'єднання фреймів, сформованих на попередніх етапах, що описують один і той самий об'єкт. Внаслідок цього утворюються фрейми першого типу, що більш повно описують об'єкт (наскільки це можливо в контексті певного тексту, аналіз якого проводиться).

4. Побудова зв'язків – виявлення зв'язків між

окремими об'єктами. Ця процедура передбачає встановлення зв'язків між об'єктами, для яких було сформовано фрейми на попередніх етапах. На основі цієї інформації будуються фрейми першого типу, що описують відношення.

5. Побудова сценарію – побудова повного опису події (факту) шляхом об'єднання результатів попередніх етапів. Ця процедура передбачає використання інформації, отриманої з попередніх етапів для побудови фреймів другого типу, що будуть виражати зв'язані між собою поняття з тексту.

Розглянемо, на яких етапах автоматизованого аналізу текстових даних можна отримати інформацію, необхідну для вилучення кожного з вищезгаданих видів фактів [5], а також оберемо структуру для подання природномовних текстових даних у вигляді, зручному для проведення фактографічного аналізу.

Графематичний аналіз. На цьому етапі з тексту виділяються структурні або синтаксичні одиниці. Характер і тип таких елементів залежить від поставленої задачі. Частіше за все таким елементом виступає слово. Цей етап слугує лише для підготовки даних для наступних етапів.

Морфологічний аналіз. Визначає морфологічну інформацію про словоформи, що були виділені на попередньому етапі. Дані, отримані на цьому етапі, використовуються для виділення іменованих сутностей.

Попередній синтаксичний аналіз. Відповідає за об'єднання окремих лексичних одиниць в одну синтаксичну. Так, наприклад, в єдину синтаксичну одиницю можуть об'єднуватися змінювані нерозривні словосполучення (наприклад, «бити байдики»). Ще одним завданням етапу попереднього синтаксического аналізу є проведення синтаксичної сегментації. Її метою є розмітка лінійного тексту на фрагменти, згідно з морфологічними правилами. На цьому етапі відбувається лише підготовка до наступного етапу і жодні процедури вилучення фактів не проводяться.

Синтаксичний аналіз. Результатом проведення даного етапу є набір дерев, вершинами яких є словоформи, а дугами – синтаксичні відношення. Для кожного речення будується своє, окреме дерево синтаксичного розбору. Виконання завдання ускладнюється величезною кількістю альтернативних варіантів, що виникають в ході розбору, пов'язаних як з багатозначністю вхідних даних (одна й та ж словоформа може бути отримана від різних нормальних форм), так і неоднозначністю самих правил розбору. Інформація, отримана на цьому етапі, використовується для побудови елементів шаблону. Також на цьому етапі процедури вилучення розпізнавання зв'язків кореференції та побудови сценарію можуть бути виконані частково, оскільки зв'язки між лексемами встановлюються в рамках кожного окремого речення.

Семантичний аналіз. На цьому етапі проводиться аналіз змісту тексту. З одного боку, семантичний аналіз відтворює семантичні зв'язки між лексемами, за допомогою яких відображається значення слів. З іншого боку, семантичний аналіз дозволяє відфільтрувати деякі значення слів або навіть цілі варіанти розбору як «семантично незв'язні». Семантична інформація застосовується для розпізнавання зв'язків кореференції та побудови сценарію, оскільки для цього потрібні дані не тільки про синтаксичні зв'язки між лексемами, але і їх семантичний зв'язок.

В залежності від обраного методу вилучення фактів в системі можуть бути реалізовані лише деякі етапи обробки тексту. Те, які етапи потрібно реалізувати, залежить від особливостей застосування конкретного методу, необхідної точності аналізу, використовуваної структури фреймів та зв'язків між ними.

Вибір структури для подання природномовних текстових даних у вигляді, зручному для проведення фактографічного аналізу, зроблений авторами на основі міркувань, наведених нижче.

Природномовний текст містить багато посилань з одного речення на інше. Такі зв'язки носять синтактико-семантичний характер. З одного боку, між реченнями можливий лише семантичний зв'язок. З іншого боку, такий зв'язок може бути встановлений через аналіз морфологічних та синтаксичних характеристик словоформ, що входять до складу речень. Встановивши такі синтактико-семантичні зв'язки між лексемами з різних речень, отримаємо структуру, що складається з синтаксичних дерев, які пов'язані між собою синтактико-семантичними зв'язками. Ця структура є орієнтованим графом, вершинами якого є лексеми або групи лексем, а дуги відображають синтактико-семантичні зв'язки у тексті. Таке подання текстових даних є зручним для застосування різних алгоритмічних підходів до вилучення фактів. В разі використання статистичних алгоритмів кожна вершина та дуга можуть мати певну вагу, що буде враховуватись при заповненні фреймів. В разі застосування методів, заснованих на правилах (шаблонах), заповнення фреймів зводиться до пошуку підграфів певної топологічної та семантичної структури.

Для отримання коректної інформації при заповненні фреймів типи слотів фреймів ставляться у відповідність синтактико-семантичним ролям підграфів (в окремому випадку – вершин-словоформ). Задача заповнення фреймів у такому випадку зводиться до пошуку певних заздалегідь заданих закономірностей на графі.

Вершини отриманого графа (лексичні вузли) можуть об'єднувати в собі декілька лексем. У випадку, коли певна група лексем виражає одне поняття, всі лексеми з цієї групи входять в однин лексичний вузол. Такі лексичні вузли утворюються, , наприклад, для власних назв, що описуються декількома словами. Певний лексичний вузол має наступну структуру:

лексема_1; ознака_11; ...; ознака_1n;
лексема_2; ознака_21; ...; ознака_2n;
.....;
лексема_m; ознака_m1; ...; ознака_mn.

2.2. Агентно-орієнтований метод фактографічного аналізу тексту

Для визначення основи нового методу фактографічного аналізу, здатного задовольнити всім вимогам, сформульованим у задачі дослідження, розглянемо існуючі у цій галузі методи та проаналізуємо їх переваги та недоліки.

1. Методи вилучення на основі ознак.

Методи вилучення на основі ознак мають в основі наявність фіксованого набору ознак (словник ознак) і ваг використання цих ознак в межах вилучених елементів тексту. Для кожного вилученого елемента визначається його вектор характеристик (ознак). Найбільш поширеними в даному класі є імовірнісні класифікатори Байеса і приховані Марківські моделі [6]. Вилучення фактів у межах цих методів зводиться до розпізнавання деякого сегмента тексту на основі ймовірнісного аналізу ознак, виявлених в околі цього сегмента. Недоліками такого підходу є використання обмеженого розміру околу (як правило, не більше 2-3 слів), необхідне для забезпечення заданої точності вилучення. Використання більшого розміру контексту призводить до зниження повноти розпізнавання і до збільшення розміру необхідної репрезентативної вибірки, на якій збирається статистика для розрахунку оцінок ймовірностей.

2. Методи вилучення, які використовують ядра.

Дані методи не мають частини недоліків попереднього класу методів за рахунок алгоритмічного визначення міри подібності між поданнями текстових сегментів, що зіставляються. Суть методів - замінити скалярний добуток векторів, що відображають ознаки розпізнаваних елементів, деякою функцією - ядром. Така функція задається алгоритмічно і враховує більш складне подання розпізнаваних елементів та їх контекстів, як правило деревоподібне, таке що описує синтаксичну структуру текстового сегмента. Для деревоподібних подань розрахунок ядра найчастіше зводиться до співставлення всіх вкладених дерев у вихідні дерева. Недоліком такого підходу є висока обчислювальна складність розрахунку ядер і визначення синтаксичної структури текстового сегмента [6].

3. Методи, засновані на співставленні зразків.

Дані методи оперують поняттям «зразок» та правилами співставлення зразків з фрагментами текстів. Зразки є ланцюжками обмежень (символи, слова, частини мови та семантичні класи) і свого роду шаблонами фраз. В цьому відношенні дані методи є аналогічними методам з використанням функцій-ядер за умови, що текстові сегменти мають «пласке» подання, а не деревоподібне. Особливість методів полягає у використовуваному способі навчання, в основі якого лежать прийоми індуктивного логічного програмування.

4. Методи, засновані на фразових зразках.

Методи, засновані на фразових зразках, є свого роду компромісом між методами вилучення, які використовують ядра, і методами, заснованими на співставленні зразків. Текстові сегменти так само, як і в підході з функціями-ядрами, подаються деревами синтаксичних зв'язків, але замість складного розрахунку ядер виконується більш проста з точки зору обчислювальної складності процедура співставлення з синтаксичним шаблоном, притаманна методам, заснованим на співставленні зразків, але доповнена морфологічними ознаками, що використовуються при співставленні синтаксичних зв'язків. Співставлення відбувається за певними, заздалегідь визначеними правилами для кожного зразка. Такі правила описують типи відношень між вузлами дерева, а також морфологічні та синтаксичні ознаки відношень та самих вузлів.

Виходячи з поставлених задач дослідження, за основу нового створюваного методу фактографічного аналізу доцільно обрати методи, засновані на фразових зразках. Такі методи використовуються для обробки текстових даних, поданих у вигляді графа синтактико-семантичних зв'язків, що відповідає обраній нами структурі внутрішнього подання даних.

У новому методі пропонується розширити поняття зразка (шаблона), з яким зазвичай працюють методи даного класу, шляхом введення поняття синтактико-семантичного шаблона. При цьому спосіб співставлення з шаблоном залишається незмінним, але до синтаксичних зв'язків та морфологічних властивостей додаються семантичні, що описують зв'язки між синтаксичними деревами і також можуть існувати в рамках окремого дерева синтаксичних зв'язків. Додавання нових фразових зразків для забезпечення вилучення нових типів фактів зводиться до додавання нових наборів правил для кожного зразка. Це дозволяє покращити масштабованість системи, побудованої на такому методі. В нашому випадку, за умови подання тексту у вигляді орієнтованого графа синтактико-семантичних відношень, фразовий зразок виглядає як шаблон підграфа, властивості якого відповідають певним вимогам. Кожна вершина цього підграфа описує ті ознаки, які може або повинен мати певний лексичний вузол. Кожна дуга цього підграфа описує можливі або необхідні відношення між відповідними лексичними вузлами.

Оскільки метод, заснований на фразових зразках, спирається на певні морфологічні, синтаксичні та семантичні властивості, він не пов'язаний з контекстом певної предметної галузі. Таким чином, застосування цього методу дозволяє побудувати гнучку систему, яка може бути налаштована на вилучення будь-яких видів фактів без модифікації структурної чи алгоритмічної складової самої системи.

На жаль, обрання за основу методу, заснованого на фразових зразках, є недостатнім кроком для забезпечення ефективного вирішення поставлених задач. Масштабованість системи по відношенню до великих обсягів тексту все одно залишається обмеженою. За допомогою розробленого методу фактографічний аналіз проводиться лінійно по структурам даних, що подають текст, і, як наслідок, потребує значного часу для пошуку відповідних фразових зразків. Проте, будь-який великий за обсягом природномовний текст може бути розділений на частини, що непов'язані або слабко пов'язані між собою синтаксично або семантично. Наявність слабких зв'язків або відсутність зв'язків взагалі дає змогу аналізувати такі частини тексту паралельно. Оскільки у вилученні різних видів фактів приймають участь різні типи зв'язків, слабко зв'язані або незв'язані частини тексту слід виділяти в межах кожного виду фактів окремо.

Відповідно до задач даного дослідження необхідно забезпечити можливість динамічного навчання системи для покращення показників ефективності її функціонування. Отже, потрібен певний механізм, за допомогою якого можна інкрементно навчати систему, використовуючи попередні результати вилучення фактів та оцінку їх якості. Такий механізм закладений в агентно-орієнтованому підході до програмування [7]. Цей підхід дозволяє, по-перше, організувати паралельну обробку агентами окремих слабозв'язних або незв'язних частин тексту і, по-друге, навчати систему вилучення фактів при її функціонуванні через взаємодію з середовищем та користувачем. Таким чином, для розв'язання задачі дослідження застосування агентно-орієнтованого підходу є доцільним.

2.3. Опис агентів

За середовище функціонування агента пропонується вважати текст, поданий у вигляді графа синтактико-семантичних зв'язків. Тоді підграфи цього графа, що слабо зв'язані або не зв'язані між собою, відповідають окремим частинам тексту, які можуть бути оброблені паралельно окремими агентами.

Зазначимо, що для вилучення кожного виду фактів необхідне виконання різних дій з боку агентів. Зазвичай агенти здійснюють свої дії за допомогою ефекторів. У нашому випадку ефекторами агента будуть виступати процедури обходу підграфа, в межах якого функціонує агент. При цьому для вилучення кожного виду фактів пропонується використовувати окремий тип агентів.

У межах даного дослідження метою агентів є заповнення фреймів для фактів відповідного виду. Кожен такий спеціалізований агент містить когнітивні структури даних, що відповідають за його поведінку і водночас слугують тими даними, на які спирається агент для виконання дій, спрямованих на заповнення фреймів. Такі структури даних є переконаннями агента, вони можуть змінюватись в процесі його діяльності. Це дає змогу проводити навчання агента під час його функціонування.

Для кожного виду фактів визначимо когнітивні структури агента, що буде проводити вилучення фактів цього виду, його мету і дії щодо вилучення фактів.

1. Виділення іменованих сутностей.

Когнітивними структурами даних для агентів, що виконують виділення іменованих сутностей, є морфологічні характеристики лексем, які задаються у фразовому зразку. Агент шукає всі збіги морфологічних характеристик лексичних вузлів із заданими в когнітивних структурах даних морфологічними характеристиками і заносить у відповідні слоти фреймів інформацію про лексичні вузли, що описують об'єкти. Процедура обходу підграфа у даному випадку полягає у лінійному перегляді всіх лексичних вузлів і їх співставленні з морфологічними ознаками, згідно з якими ведеться пошук іменованих сутностей. Агент певним чином позначає знайдені вузли як такі, що відповідають іменованим сутностям, для подальшого використання цієї інформації іншими агентами на наступних етапах методу.

2. Побудова елементів шаблону.

При вилученні фактів даного виду когнітивними структурами даних є синтаксичні та морфологічні ознаки лексем, які можуть бути пов'язані з іменованими сутностями. Процедура вилучення фактів, що є елементами шаблону, зводиться до проходження агентом вузлів, що прямо чи опосередковано зв'язані із знайденими на попередньому етапі вузлами і відміченими як іменовані сутності. Кількість вузлів, зв'язаних з іменованими сутностями, які необхідно співставити із зразком, заданим у когнітивних структурах даних, залежить від налаштувань агента і повинна задаватися користувачем як ступінь деталізації пошуку.

Метою агента у даному випадку є заповнення слотів фреймів першого типу, що відповідають за опис об'єкта. Агент помічає вже знайдені вузли для подальшого використання цієї інформації на наступних етапах вилучення фактів.

3. Розпізнавання зв'язків кореференції.

Когнітивні структури даних агента на цьому етапі вилучення фактів мають містити інформацію про необхідні синонімічні відношення та відношення кореференції між іменованими сутностями. Вилучення фактів зводиться до проходження агентом всіх вузлів, що представляють іменовані сутності, і співставлення їх один з одним для визначення того, чи є вони кореферентно пов'язаними. В результаті всі пов'язані таким чином вузли помічаються як одна й та сама іменована сутність. Відповідно, ознаки таких сутностей також об'єднуються.

Метою агента є утворення нових фреймів першого типу шляхом об'єднання фреймів, утворених на попередніх етапах.

4. Побудова зв'язків та побудова сценарію.

Когнітивні структури даних агента на цьому етапі повинні містити інформацію про фразові зразки. Ці зразки є певного виду синтактико-семантичними зв'язками між лексичними вузлами, а також визначеною морфологічною та синтаксичною інформацією про ці лексичні вузли. За даними зразками повинно здійснюватися вилучення фактів. На цьому етапі воно зводиться до проходження агентом всіх груп вузлів, визначених на попередньому етапі як іменовані сутності, з метою пошуку збігів між такими групами та іншими вузлами із фразовими зразками. При вдалому співставленні із фразовими зразками агент позначає певними чином всі лексичні вузли, що приймають участь у формуванні відношення між іменованими сутностями. Після цього агент намагається визначити опис вузлів, що формують відношення, переглядаючи вузли, зв'язані з ними прямо чи опосередковано. Кількість таких вузлів залежить від налаштувань агента і повинна задаватися користувачем як ступінь деталізації пошуку.

Метою агента на даному етапі є формування фреймів першого типу, що описують відношення між іменованими сутностями, та фреймів другого типу, що зв'язують між собою іменовані сутності.

Для забезпечення інкрементного навчання кожен тип агентів на кожному етапі взаємодіє з користувачем з метою отримання оцінки якості виконаного етапу фактографічного аналізу. Агент надає користувачу вилучені фрейми та відповідні фразові зразки, за допомогою яких було здійснено вилучення. Таким чином, користувач корегує правильність виконання вилучення фактів.

Висновки

В результаті проведеного дослідження для опису фактів у системі фактографічного аналізу природномовних текстових даних обрано фреймову модель подання інформації. Виділено два типи фреймів та визначено, якими даними можна заповнювати слоти цих фреймів в залежності від глибини аналізу текстових даних, здійснюваного у межах процедури вилучення фактів. Обгрунтовано вибір структури даних для подання природномовного тексту у вигляді, зручному для проведення фактографічного аналізу. Авторами доведена доцільність використання методів, заснованих на фразових зразках, як основи для нового методу фактографічного аналізу, що пропонується у статті.

Новий метод розроблено на основі агентно-орієнтованого підходу до програмування. Він дозволяє абстрагуватися від конкретної предметної галузі та налаштовувати існуючу систему вилучення фактів на роботу з будь-якими видами даних, а також надає можливість динамічного навчання системи в процесі роботи, спираючись на результати, отримані на попередніх циклах її роботи.

Подальшого дослідження, на думку авторів, потребують можливі модифікації базового способу функціонування агентів. Також перспективним напрямком продовження роботи над тематикою статті є підвищенння швидкодії запропонованого методу.

Список літератури: 1. *Ландэ, Д.В.* Интернетика. Навигация в сложных сетях: модели и алгоритмы [Текст] / Д.В. Ландэ, А.А. Снарский, И.В. Безсуднов. — М.: Либроком, 2009. — 264 с. **2.** *Bocharov V.* Ontological parsing of encyclopedia information [Text] / Bocharov V., Pivovarova L., Rubashkin V., Chuprin B. Lecture Notes in Computer Science. T. 6008 LNCS., 2010. — 650 p. **3.** *Пивоварова, Л. М.* Фактографический анализ текста в системе поддержки принятия решений [Текст] / Л. М. Пивоварова. Вестник Санкт-Петербургского университета. Сер. 9, Филология, востоковедение, журналистика. — 2010. — Вып. 4, декабрь. — С. 190-197. — Библиогр.: с. 196-197 (21 назв.). **4.** *Гаспаров, Б. М.* Язык, память, образ. Лингвистика языкового существования [Текст] / Б. М. Гаспаров. — М.: Новое Литературное Обозрение,1996. — 352 с. **5.** *Леонтьева, Н. Н.* Автоматическое понимание текста: системы, модели, ресурсы [Текст] / Н. Н. Леонтьева. — М.: Издательский центр «Академия», 2006. — 304 с. **6.** *Симаков К.В.* Модель извлечения знаний из естественно-языковых текстов [Текст] / А.М. Андреев, Д.В. Березкин, К.В. Симаков // Информационные технологии., 2007. — С. 57—63. **7.** *Tomas Salamon.* Design of Agent-Based Models. [Text] / Tomas Salamon. Czech Republic, Bruckner Publishing, 2011. — 220 p.

Поступила до редколегії 15.11.2012

УДК 004.912

Агентно-ориентированный метод автоматизированного извлечения фактов из естественно-языковых текстовых данных /А.Б. Бегунов, Т.Н. Заболотняя // Бионика интеллекта: науч.-техн. журнал. – 2013. – № 1 (80). – С. 29-34.

В данной работе предложен метод извлечения фактографической информации из естественно-языковых текстовых данных, основанный на сопоставлении фразовых образцов, а также агентно-ориентированном подходе к созданию программного обеспечения. Разработанный метод дает возможность абстрагироваться от предметной области текстов, для которых проводится извлечение фактов, и реализовать на его основе программную систему, обладающую возможностью добавления новых видов фактов для извлечения, а так же способностью к динамическому обучению. Описаны этапы работы метода. Представлены особенности строения и функционирования агентов, при помощи которых предлагается производить извлечение фактов.

Библиогр.: 7 items.

UDK 004. 912

The agent-oriented method of automatic fact extraction from natural language text/A. Begunov, T. Zabolotnia // Bionics of Intelligense: Sci. Mag. – 2013. – № 1 (80). – P. 29-34.

In this work the method of automated fact extraction from natural language texts based on phrase pattern matching and agent-oriented approach to software development is proposed. The developed method allows to abstract from the subject area of the processed texts and enables to implement the software system based on it supporting the ability to add a new types of facts to extract as well as the ability of dynamically learning. The steps of the method are described. The features of the structure and functioning of the proposed agents for performing the extraction of facts are presented.

Ref.: 7 items.