УДК 681.3.093:044.3 DOI 10.36910/6775.24153966.2019.68.25

В.С. Яременко, А.С. Худяков
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського

МОДЕЛЬ МУЛЬТИАГЕНТНОЇ СИСТЕМИ ДЛЯ СЕМАНТИЧНОГО АНАЛІЗУ ТЕКСТІВ

У даній роботі запропоновано підхід до аналізу тексту та отримання інформації з нього використовуючи знання про моделі лексичної мови. Була запропонована модель мультиагентної системи, що дає можливість паралельно обробляти текстові документи та виконує семантичну обробку тексту. Запропоновано модель опису процесу видобутку інформації за допомогою системи покриття тексту.

Ключові слова: *Мультиагентні системи, глибинне навчання, інтелектуальний аналіз тексту, обробка великих масивів текстових даних.*

В.С. Яременко, А.С. Худяков

МОДЕЛЬ МУЛЬТИАГЕНТНОЙ СИСТЕМЫ ДЛЯ СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ

В этой работе предложен подход для глубокого обучения текста и получения из него информации на основании знаний о моделях лексического языка. Была создана мультиагентная система, которая дает возможность параллельной обработки текстовых документов и выполняет семантический анализ текста. Также предложена модель описания процесса добычи информации с помощью использования системы покрытий текста

Ключевые слова: *Мультиагентные системы, глубокое обучение, интеллектуальный анализ текста, обработка больших массивов. текстовых данных.*

V.S. Yaremenko, A.S. Khudiakov

MODEL OF MULTIAGENT SYSTEM FOR SEMANTIC TEXT ANALYSIS

The approach to the deep text analysis and mining information on the basis of knowledge about the model of lexical language is proposed. A model for describing the process of extracting information using a system of text processing is proposed. This model enables parallel processing of text documents. With this system, we can improve the process that analysis a text document as a whole, and it does execute semantic analysis as well. The most significant advantage of using a multiagent system is the ability to simultaneously process a text document, and this system can also help to remove repetitions from the text. The downside is that during the process, the algorithm generates multiple agent conversations, as well as breaking existing connections and establishing new ones. This behavior of the model requires a considerable amount of computing resources. The model receives a text document. The result of the model is the object coverage of the text. The set of information objects received is subsequently refined and a resulting set of objects is formed that describes the content of the document in terms of the ontology of the subject area. All the knowledge used in this approach is, to one degree or another, based on a domain model that captures the concepts and relationships of interest to the user of the system in the form of an ontology. Thus, the ontology determines what kind of information should be extracted from the available data sources. The results of each stage of processing are projected onto text, which allows to interpret the obtained results clearly and to distinguish fragments that are contextually related to each element of the received information.

Keywords: *Multiagent systems, deep learning, intellectual text mining, processing large text data arrays.*

Постановка проблеми. Розробка методів розпаралелювання процесу обробки текстів стає все більш актуальною в зв'язку з наростанням обсягів текстових даних, в тому числі на інтернет-сайтах. Найдовшим етапом обробки тексту є його концептуальний або семантичний аналіз і саме для даного етапу має сенс у першу чергу застосовувати засоби інтелектуальної багатопотокової оптимізації.

Одним із засобів організації процесу паралельної обробки даних є мультиагентні системи. Вони використовуються в тому числі, і для обробки текстів природною мовою [2,3] і вилучення інформації з мережі Інтернет [4].

Мультиагентна система передбачає наявність співтовариства автономно діючих агентів. Однак у переважній більшості робіт з даної тематики агенти є сутностями, які швидше направляють потоки даних, використовуючи для їх обробки стандартні алгоритмічні модулі, ніж безпосередньо реалізують їх обробку. Тим самим, мультиагентний підхід застосовується до організації процесу обробки текстів у цілому, але не зачіпає безпосередньо семантичний аналіз, який все одно реалізується послідовно, і, отже, істотного виграшу в продуктивності досягатися не може.

Виклад основного матеріалу дослідження. Найважливіша причина використання мультиагентних систем при проектуванні інформаційної системи полягає в тому, що деякі доменні області цього вимагають. Зокрема, якщо є різні люди чи організації з різними (можливо,

суперечливими) цілями та власною інформацією, то для їх взаємодії потрібна мультиагентна система. Навіть якщо кожна організація хоче моделювати свої внутрішні справи за допомогою єдиної системи, організації не дадуть повноважень жодній окремій особі будувати систему, яка представляє їх усіх: різним організаціям знадобляться власні системи, що відображають їх можливості та пріоритети [7].

Модель знань в даній статті розглядається в двох аспектах. По-перше, модель даних/інформації, яка використовується в процесі породження знань з текстових джерел. Результати кожного етапу обробки проектуються на текст, що дозволяє наочно інтерпретувати отримані результати і виділяти фрагменти, контекстно пов'язані з кожним елементом отриманої інформації

По-друге, модель знань про контекст, в рамках якого здійснюється обробка тексту. До таких знань відносяться словники предметної лексики, моделі фактів, що описують способи вираження інформації, прийняті в даній області знань, а також знання про типи та жанри, розглянутих текстових джерел і предметні знання, які вже є в базі даних, наприклад, отримані раніше під час обробки інших джерел.

Всі знання, які використовуються в даному підході, в тій чи іншій мірі спираються на модель предметної області, яка фіксує поняття і відносини, що цікавлять користувача системи, у вигляді онтології. Таким чином, онтологія визначає, яку саме інформацію слід витягати з доступних джерел даних.

Особливістю розвиваючого підходу до вилучення інформації з тексту є застосування знань стосовно предметної області (ПО) та переважне використання лексико-семантичної інформації, що не виключає застосування часткового синтаксичного аналізу і синтаксичних обмежень, що накладаються на семантичний каркас концептуальних схем фактів.

Семантико-синтаксичні моделі. Одним із способів опису синтаксису мови є підхід, в основі якого лежать так звані моделі управління. Суть цього підходу полягає у встановленні відповідності лексемі або групі однотипних лексем деякого правила, яке описує необхідні селективні ознаки пов'язаних слів (валентності).

Семантико-синтаксична модель обмежує синтаксичну сполучуваність і узгодженість граматичних і семантичних ознак термінів (вершин синтаксичних груп) відповідно до правил узгодження і управління. Такі моделі описуються у вигляді актантної структури, пов'язаної з однією або декількома узагальненими лексемами [5]. Під узагальненою лексемою розуміється або термін словника (або його форма), або група лексем, описаних в термінах граматичних і семантичних категорій без вказівки нормальної форми. Актантна структура описує набір актантів, що характеризують відповідну валентність, в термінах семантичних і граматичних характеристик, які є обмеженнями для залежних слів.

Формально, семантико-синтаксична модель, яка визначається щодо словника V, характеризується парою SS = < lg, A >, де

$lg = L_V, S_V, M_V >$ узагальнена лексема, що характеризує групу термінів словника $L_V \subseteq V$що володіє набором семантичних ознак S_V та морфологічних атрибутів M_V; [6]

A = <a1,…an> – послідовність актантів, що описують модель, де кожен актант ai = <Si, Mi>, представлений множиною альтернативних семантичних ознак Si, а для кожної ознаки sij є Si задається набір морфологічний обмежень mij ⊆ Mi.

Запропонована структура семантико-синтаксичні моделей надає широкі можливості моделювання мовних зв'язків у тексті. Так, модель може не містити синтаксичних обмежень і представляти собою онтологічні відносини, або описуватися без семантичних характеристик і відповідати чисто синтаксичним моделям управління. Узагальнення лексем в моделях дозволяє компактно визначити декілька мовних конструкцій, варіанти взаємозв'язку слів у виразах і словникові групи.

Моделі фактів. Модель фактів формує знання про узгодження наявних лінгвістичних знань з предметними знаннями. У спрощеному вигляді, без семантико синтаксичного компонента, дана модель була запропонована в роботі [6]. Модель фактів задається структурою, аналогічною актантній структурі SS. Вона описується або в термінах класів онтології, або в термінах семантичних ознак словника і зв'язується з фрагментом онтології. Додатково накладаються обмеження на онтологічні ознаки елементів структури і їх взаєморозташування в тексті.

Модель тексту

У процесі обробки тексту його уявлення послідовно змінюється, збагачуючись на кожному етапі новими знаннями. Для опису зміни запропонована концепція покриттів тексту, коли кожне покриття представляється набором однотипних елементів з заданими текстовими позиціями (інтервалами). Виділяються наступні типи покриттів.

(1) Графематичне покриття – являє собою розбиття тексту на елементарні складові, такі як слово, розділовий знак, абзац, число і т.п.;

(2) Термінологічне покриття складається з словникових термінів, знайдених в даному тексті, з урахуванням можливої омонімії і перетинів багатослівних термінів;

(3) Сегментне покриття відображає структурний поділ тексту на логічні (абзац, речення, заголовок тощо) і жанрові фрагменти;

(4) Тематичне покриття визначає текстові кордону тематично пов'язаних областей тексту для кожної розглянутої тематики;

(5) Об'єктне покриття описує знайдену інформацію у вигляді семантичної мережі об'єктів предметної області.

Таким чином, модель тексту визначається <G, LC, SC, TC >, де

G – графематичне покриття, що визначає текстові позиції елементів моделі,

LC – термінологічне покриття, впорядкована за текстовими позиціями послідовність лексичних об'єктів виду l = <v, mv, sv, pos>, де v ϵ V термін тезаурусу; mv – множина морфологічних характеристик терміна v; sv – множина семантичних ознак v; pos – текстова позиція v[6],

SC – сегментне покриття, що включає ієрархічно-упорядкований набір сегментів виду s=<ts, pos, Rs>, де кожен сегмент визначається типом ts, текстовими позиціями pos і зв'язками Rs з іншими сегментами, що визначають їх взаєморозташування в тексті,

TC – тематичне покриття, IC – об'єктне покриття, задає множину онтологічних об'єктів і вказує текстові фрагменти, в яких були знайдені їх опису[6].

Графематичне покриття тексту є результатом графематичного аналізу, в процесі якого вхідний лінійний текст розбивається на елементарні атоми. Основне завдання даного етапу згрупувати символи одного типу в послідовності і дати їм необхідну інтерпретацію: слово певного алфавіту, число, символ. Для рахівників, що працюють з розміткою (наприклад, html-тексти), можна додатково задати типізацію тегів або послід. Важливою властивістю даного подання, є те, що елементи покриття задають всі можливі межі елементів для всіх наступних вистав, тобто при подальшій обробці жоден атом не може бути «розділений».

Термінологічне покриття тексту – це лексична текст модель, яка будується на основі лексичної моделі підмови, і включає знайдені в тексті терміни з прив'язкою до позиції в тексті. Після того, як термін знайдений в тексті (точніше в графематичному покритті), формується лексичний об'єкт, який забезпечується набором атрибутів, заданих в тезаурусі для знайденого терміну.

Сегментне покриття є результатом сегментації тексту і одним із способів відображення формальної структури тексту. У даному підході сегментація розглядається на макрорівні, тобто на рівні всього тексту (на відміну від локального аналізу пропозиції та виділення сукупності взаємопов'язаних фрагментів (клауз), що розглядаються в рамках синтаксичного аналізу речень) і спирається як на формально-текстові, так і на жанрові особливості документа [6], які передаються поділом тексту на концептуальні частини. При аналізі тексту розбиття на жанрові фрагменти допомагає звузити область пошуку інформації певного виду і, тим самим, підвищити якість аналізу. Також вирішуються завдання визначення жанрової релевантності документів, отриманих з невідомих джерел, наприклад, при пошуку в інтернет.

Тематичне покриття задає множину областей або фрагментів тексту, що покривають набір певних тем. Формування таких областей здійснюється на основі словника, в якому задано відповідність між термінами і тематичними ознаками. Тематичне покриття будується над термінологічним покриттям. Ми визначаємо елемент тематичного покриття або тематичний шар як фрагмент тексту включає кластер термінів, що відносяться до однієї теми, в межах формального сегмента (або послідовності сегментів) сегментного покриття. Аналогічно сегментам, тематичні шари можуть звужувати область пошуку інформації певного виду. Проте зазвичай, даний вид покриттів використовується в задачах тематичної кластеризації та класифікації тексту і тому виходить за рамки розгляду даної роботи.

Модель мультиагентної системи.

Модель отримує тектовий документ. Результатом роботи моделі є об’єктне покриття тексту. Множина одержаних інформаційних об’єктів згодом уточнюється та формується результуюча множина об’єктів, що описує контент документа в термінах онтології предметної області.

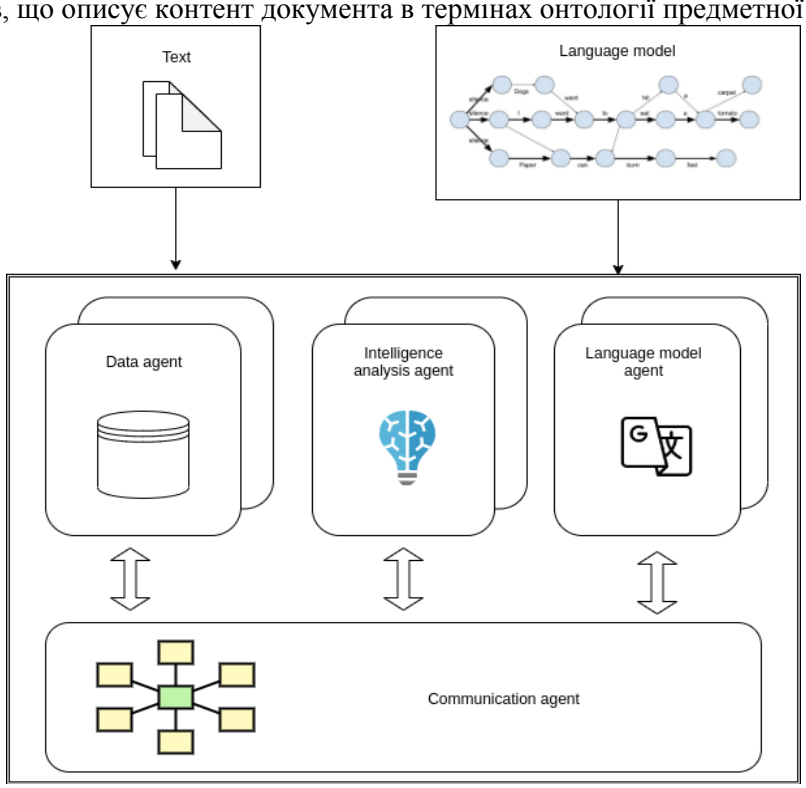

Рис 1. **Модель мультиагентної системи**

Дана система складається з агентів чотирьох видів:

1. Агент даних

Агент отримує документ та витягує текстову інформацію, забезпечує уніфікацію різнорідних даних, які надходять з різних джерел (наприклад з Бази даних). Виконує попередню обробку даних та знаходить відповідні зв'язки. Результатом роботи агента буде лінійний текст з базовим форматуванням та метаданими.

2. Агент інтелектуального аналізу

Агент отримує попередньо оброблений документ та виконує аналіз контекстної інформації. Кожен об'єкт контексту породжується на основі мовної моделі. Відповідно до отриманих даних вони можуть породжувати нових агентів даних, а також виявляти значення їх атрибутів.

3. Агент комунікатор

У процесі комунікації агенти домовляються про відповідність токенів мовної моделі та відповідної їй онтології. Також на основі знаходження нових токенів агент породжує нові агенти мовної моделі для поповнення онтології. Робота агента комунікатора також полягає в послідовному аналізу роботи інших агентів. Якщо всі агенти, крім нього є неактивними, він закінчує роботу даного алгоритму.

4.Агент мовної моделі

Агент виконує аналіз кожного окремого токена, тобто, встановлює відповідність між класами заданої онтології та текстовими одиницями. Має можливість поповнювати свій словниковий запас розширюючи онтологію.

Агенти взаємодіють за допомогою повідомлень двох видів:

1. Передача інформації про нові дані відбувається за допомогою агента комунікатора та виконується між агентом даних та агентом інтелектуального аналізу. Метою такого запиту є отримання інформації про певні атрибути та зв’язки між ними для кожного окремого документа.

2. Повідомлення токена. Такими повідомленнями обмінюється агент інтелектуального аналізу даних та агент мовної моделі за допомогою агенту комунікатора. Даний запит виконується для поповнення мовної моделі та аналізу кожного окремого слова.

Опис протоколів роботи агентів, способів розуміння один одного, та способів комунікації представлений в роботі [1]. Всі агенти можуть працювати паралельно доки вони не перейдуть в

стан очікування. Момент зупинки визначається агентом комунікатором. Найбільш вагомою перевагою використання мультиагентної системи є можливість паралельної обробки текстового документа, також дана система може допомогти з видаленням повторів з тексту. Недоліком може бути те, що у процесі роботи алгоритм породжує численні переговори агентів, а також розрив існуючих зв'язків і встановлення нових. Така поведінка моделі може вимагати додаткових обчислювальних ресурсів.

Висновок. Алгоритм, що наведений у даній роботі надає можливість паралельно добувати інформацію з текстових файлів за допомогою використання мультиагентної системи. За допомогою цієї системи ми можемо пришвидшити процес обробки текстового документа в цілому, і також вона може виконувати семантичний аналіз, який зазвичай відбувається послідовно. Тобто, для паралелізації процесу семантичної обробки тексту можливо використовувати також даний метод, який покриває дана стаття.

Список використаних джерел :

1. Michael Wooldridge. An Introduction to Multi Agent Systems.– University of Liverpool: Wiley, 2009.

2. Aref, M.M. A Multi-Agent System for Natural Language Understanding} International Conference on Integration of Knowledge Intensive Multi-Agent Systems, 2003, 36.

3. C.T. dos Santos, P. Quaresma, I. Rodrigues, R. Vieira A Multi-Agent Approach to Question Answering // In Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006) LNAI 3960, 13-17 de Maio de 2006, Berlin/Heidelberg: Springer Verlag, pp. 131-139.

4. Cheng X., Xie Y., Yang T. Study of Multi-Agent Information Retrieval Model in Semantic Web // In Proc. of the 2008 International Workshop on Education Technology and Training and 2008 International Workshop on Geoscience and Remote Sensing (ETTANDGRS'08), 2008, Vol. 02, P. 636-639.

5. Яковчук Е.И., Сидорова Е.А. Обобщенные семантико-синтаксические модели в задачах обработки текста // Труды рабочего семинара «Наукоемкое программное обеспечение НПО-2011». Ершовская конференция по информатике. –Новосибирск: ИСИ СО РАН, 2011. –С.287-292.

6. Гаранина Н.О., Сидорова Е.А. Мультиагентный подход к извлечению информации из текстов и пополнению онтологии.

7. MultiAgentSystems [Електронний ресурс]. – Режим доступу доресурсу: https://www.cs.cmu.edu/afs/cs/usr/pstone/public/papers/97MAS-survey/node2.html. - Дата доступу: 4.10.2019.