DOI:
UDC 004.02
I.A. Karpov, postgraduate, karpovilya5@gmail.com
S.V. Antonenko, candidate of engineering sciences, associate professor, szemlyanaya@gmail.com
Oles Honchar Dnipro National University, Dnipro

USAGE OF MULTI-AGENT SYSTEM TO SOLVETEXT PROCESSING PROBLEM

In this article an approach to the analysis of the text and obtaining information is proposed. A model of a multi-agent system was proposed, which allows you to process text documents and perform semantic text processing. A model for describing the process of extracting information using a text coverage system is proposed.

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

У даній роботі запропоновано підхід до аналізу тексту та отримання інформації з нього, використовуючи знання про моделі лексичної мови. 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.*

Problem's Formulation

The development of parallelization methods for word processing is becoming increasingly important due to the increase in the volume of text data, including on Internet sites. A long stage of text processing is its conceptual or semantic analysis, and it is for this stage that it makes sense first of all to use the means of intelligent multi-threaded optimization.

One of the means for organizing the process of parallel data processing is multi-agent systems. They are used, among other things, for processing natural language texts and extracting information from the Internet.

A multi-agent system assumes a community of autonomously acting agents. However, in the overwhelming majority of works on this topic, agents are entities that rather direct data flows using standard algorithmic modules for their processing than directly implement their processing. Thus, the Multi-agent approach is applied to the organization of the word processing process as a whole, but does not directly affect the semantic analysis, it is nevertheless implemented sequentially, and, therefore, a significant performance gain cannot be achieved.

Analysis of recent research and publications

Decision-making issues have been dealt with by many domestic and foreign scientists from different countries and universities. A big amount of different multi-agent systems for analysing text documentsalready exist.

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.

Formulation of the study purpose

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.

Presenting main material

The most important reason for using multi-agent systems in the design of an information system is that some domain domains require this. In particular, if there are different people or organizations with different (possibly conflicting) goals and their own information, then for their interaction a multi-agent system is needed. Even if each organization wants to model its internal affairs using a single system, organizations will not be given the authority of one separate person to build a system that represents them all: different organizations will need their own systems that reflect their capabilities and priorities [1].

The knowledge model in this article is considered in two aspects. Firstly, the data / information model used in the process of generating knowledge from text sources. The results of each processing stage are projected onto the text, allows you to visually interpret the results and highlight fragments that are contextually associated with each element of the information received

Secondly, a model of knowledge about the context within which text processing is carried out. Such knowledge includes dictionaries of subject vocabulary, models of facts describing the ways of expressing information accepted in a given field of knowledge, as well as knowledge of the types and genres of considered text sources and subject knowledge that already exists in the database, for example, obtained earlier in time processing other sources.

All the knowledge that is used in this approach, to one degree or another, is based on the domain model, which captures the concepts and relationships that interest the system user in the form of an ontology. Thus, the ontology determines which information should be extracted from available data sources.

A feature of the approach, it is considered that the use of knowledge in accordance with the subject area and the predominant use of lexical and semantic information to extract information from text does not exclude the use of partial parsing and syntactic restrictions that are imposed on the semantic framework of conceptual factual schemes.

Semantics-syntactic models. One way to describe the syntax of a language is an approach based on so-called control models. The essence of this approach is to establish certain rules that correspond to the lexeme or group of the same type of lexemes, which describes the necessary selective attributes of related words (valencies).

The semantic-syntactic model limits the syntactic compatibility and consistency of grammatical and semantic features of terms (vertices of syntactic groups) in accordance with the rules of coordination and management. Such models are described in the form of an actant structure associated with one or more generalized tokens [2]. A generalized lexeme means either the term of the dictionary (or its form), or a group of lexemes described in terms of grammatical and semantic categories without indicating a normal form. An actant structure describes a set of actants that characterize the corresponding valency, in terms of semantic and grammatical characteristics, which are limitations on dependent words.

Formally, the semantic-syntactic model, which is determined by the V dictionary, is characterized by a pair$SS =< lg, A >$, where $lg = L_V, S_V, M_V$ is generalized token characterizing a group of vocabulary terms$L_V \subseteq V$possessing a set of semantic attributes$S_V$and morphological attributesM_V; $A =< a_1, \dots, a_n >$ is a sequence of actants describing the model, where each actant $a_i =< S_i, M_i >$, of representations by a set of alternative semantic attributes S_i,and for each trait $s_{ij} \in S_i$ a set of morphological restrictions is settrait $m_{ij} \subseteq M_i$.

The given structure of semantic-syntactic models provides ample opportunities for modeling language relationships in the text. So, the model may not contain syntactic restrictions and represent ontological relations, or be described without semantic characteristics and correspond to purely syntactic control models. The generalization of lexemes in models allows one to compactly define several language constructions, variants of the relationship of words in expressions and dictionary groups.

Models of facts. The fact model generates knowledge about the coordination of existing linguistic knowledge with subject knowledge. In a simplified form, without a semantic-syntactic component, this model was proposed in [3]. The fact model is defined by a structure similar to the SS actant structure. It is described either in terms of ontology classes, or in terms of semantic features of a dictionary and is associated with a fragment of ontology. Additionally, restrictions are placed on ontological features of structural elements and their relative position in the text.

Text model. In the process of processing the text of his presentation is gradually changing, enriching at each stage with new knowledge. According to the results of the inspection of works [1-5], to describe the change in the sequences of representations, we offer the concept of text coverings, when each cover is represented by a set of elements of the same type with given text positions (intervals). The following types of coatings are distinguished:

1. Grafematic coverage - is a breakdown of the text into elementary components, such as a word, punctuation mark, paragraph, number and the like.

2. The terminological cover consists of the vocabulary terms found in this text, taking into account possible homonymy and relationships of verbose terms.

3. segment coverage reflects the structural division of the text into logical (paragraph, sentence, title, etc.) and genre fragments.

4. The thematic coverage defines the textual boundaries of thematically related text areas for each subject under consideration.

5. Object coverage describes the information found in the form of a semantic network of domain objects.

Thus, the text model is determined by the combination of coverages $< G, L_C, S_C, T_C >$, where

G is graphematical coverage, determines the text position of the model elements;

L_C is terminological coverage, ordered by text position sequence of lexical objects form $l =< v, m_v, s_v, pos >$, where

$v \in V$ is thesaurus term;

m_v is many morphological characteristics of the term v;

s_v is many semantic attributes of the$v$;

pos is text position of the v [5];

S_C is segment coverage, including a hierarchically ordered set of the form segments $s =< t_s, pos, R_S >$, where each segment is determined by the type t_s, the text positions pos and the relationships of R_S with other segments determine their relative position in the text;

T_C is thematic coverage;

I_C is object coverage, defines many ontological objects and indicates text fragments in which their descriptions were found.

The graphematical coverage of the text is the result of graphematical analysis, in which the input linear text is divided into elementary atoms. The main task of this stage is to group symbols of the same type in a sequence and give them the necessary interpretation: a word of a certain alphabet, number, symbol. For counters that work with markup (for example, html-texts), you can optionally set the typification of tags or labels. An important property of this stage is that the coating elements specify all possible limits of the elements for all subsequent representations, that is, during further processing, no atom can be “divided” [4, 5].

The terminological cover of the text is a lexical model of the text, which is based on the lexical model of the sublanguage, and includes the terms found in the text with reference to the position in the text. After the term is found in the text (more precisely, in the graphematical cover), a

lexical object is formed, which is provided by a set of attributes specified in the thesaurus for the found term [4, 5].

Segment coverage is the result of text segmentation and one way to display the formal structure of the text. In this approach, segmentation is considered at the macro level, that is, at the level of the entire text (as opposed to local sentence analysis) and the allocation of a set of interconnected fragments (clauses) that are considered in the framework of sentence parsing) and is based both on formal-textual and on genre features of the document [4, 5], which are transmitted by dividing the text into conceptual parts. When analyzing text, dividing into genre fragments helps narrow the scope of the search for information of a certain type and, thereby, improve the quality of analysis. The problems of determining the genre relevance of documents obtained from unknown sources, for example, when searching the Internet, are also being solved.

The thematic coverage defines many areas or text fragments that cover a set of specific topics. The formation of such areas is carried out on the basis of a dictionary in which a correspondence between terms and thematic features is specified. Thematic coverage is built on terminological coverage. We define a topic cover element or topic layer as a piece of text that includes a cluster of terms related to a single topic within the formal segment (or sequence of segments) of the segment coverage. Similar to segments, thematic layers can narrow the search area for information of a certain kind. However, as a rule, this type of coverage is used in problems of thematic clustering and text classification, and therefore is beyond the scope of this paper.

Multi-agent system model. After analyzing the existing models of multi-agent systems, a new model was created that, unlike others, can process text documents and perform semantic text processing. The created model receives a text document. The result of the model is an object coverage of the text. The set of obtained information objects is refined and the resulting set of objects is formed, describes the content of the document in terms of the ontology of the subject area.

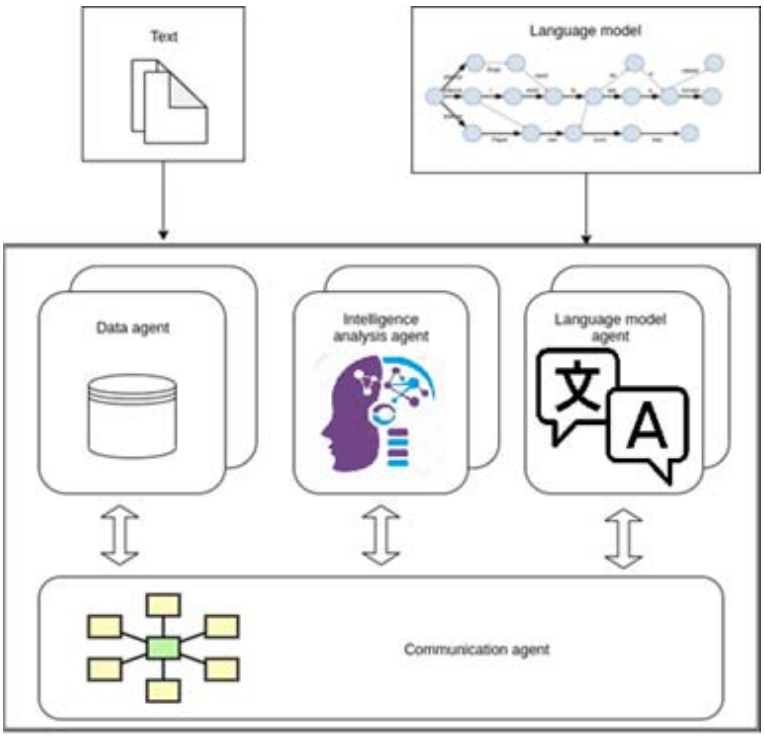

Fig. 1. Multi-Agent System Model

This system consists of four types agents:

1. Data agent. The agent receives a document and extracts textual information, provides a unification of heterogeneous data coming from various sources (for example, from a database).

Performs preliminary data processing and finds the appropriate relationships. The result of the agent is linear text with basic formatting and metadata.

2. Mining agent. The agent receives a pre-processed document and performs an analysis of contextual information. Each context object is generated based on the language model. According to the received data, they can generate new data agents, as well as reveal the value of their attributes.

3. Agent communicator. In the communication process, agents agree on the correspondence of the language model and the corresponding ontology tokens. Also, on the basis of finding new tokens, the agent generates new agents of the language model for replenishing the ontology. The work of the agent-communicator also consists in a sequential analysis of the work of other agents. If all agents except him are inactive, he ends the operation of this algorithm.

4. Agent language model. The agent analyzes each individual token, that is, establishes a correspondence between the classes of a given ontology and text units. Has the ability to replenish his vocabulary, expanding ontology.

Agents interact using two types of messages:

1. Information about new data is transmitted using the communicator agent and is performed between the data agent and the data mining agent. The purpose of such a request is to obtain information about certain attributes and the relationship between them for each individual document.

2. Token message. These messages are exchanged between the data mining agent and the language model agent using the communicator agent. This query is performed to replenish the language model and analyze each individual word.

A description of the agents protocols, ways of understanding each other, and methods for communication are presented in [5]. All agents can work in parallel until they go into a wait state. The stopping moment is determined by the communicator agent. The biggest advantage of using a multi-agent system is the possibility of a text document parallel processing, and this system can also help with the removal of repetitions in the text. The disadvantage may be that in the process, the algorithm generates numerous negotiations of agents, as well as breaking existing relationships and establishing new ones. Such model behavior may require additional computing resources.

Conclusions

The algorithm presented in this paper allows parallel processing of text documents and semantic processing of text using the process of extracting information using a text coverage system and knowledge for lexical language models.

References

[1] Aref, M.M. A Multi-Agent System for Natural Language Understanding. – International Conference on Integration of Knowledge Intensive Multi-Agent Systems, 2003, 36.

[2] 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.

[3] 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.

[4] Yakovchuk E.I., Sidorova E.A. Generalized semantic-syntactic models in text processing tasks // Proceedings of the workshop “High-tech software NGO-2011”. Ershov conference on computer science. –Novosibirsk: ISI SB RAS, 2011. –P.287-292.

[5] Michael Wooldridge. An Introduction to MultiAgent Systems. – University of Liverpool: Wiley, 2009.

ВИКОРИСТАННЯ МУЛЬТИАГЕНТНОЇ СИСТЕМИ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ ОБРОБКИ ТЕКСТУ

Карпов І.А., Антоненко С.В.

Реферат

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.

Були розглянуті способи організації мультиагентної системи з використанням різних моделей опису даних та структур семантико-синтаксичних моделей, які надають широкі можливості моделювання мовних зв'язків у тексті. Також були розглянуті різні покриття тексту. Після аналізу існуючих моделей мультиагентних систем була створена нова модель, яка, на відмінну від інших, може обробляти текстові документи та виконувати семантичну обробку тексту. Створена модель отримує текстовий документ. Результатом роботи моделі є об'єктне покриття тексту. Множина одержаних інформаційних об'єктів згодом уточняється та формується результуюча множина об'єктів, що описує контент документа в термінах онтології предметної області. 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).
2. Агент інтелектуального аналізу. Агент отримує попередньо оброблений документ та виконує аналіз контекстної інформації.
3. Агент-комунікатор. У процесі комунікації агенти домовляються про відповідність токенів мовної моделі та відповідної їй онтології.
4. Агент мовної моделі. Агент виконує аналіз кожного окремого токена, тобто, встановлює відповідність між класами заданої онтології та текстовими одиницями.

Агенти взаємодіють за допомогою повідомлень двох видів:

1. Передача інформації про нові дані відбувається за допомогою агенту-комунікатора та виконується між агентом даних та агентом інтелектуального аналізу.
2. Повідомлення токена. Такими повідомленнями обмінюється агент інтелектуального аналізу даних та агент мовної моделі за допомогою агенту-комунікатора.

Алгоритм, що наведений у даній роботі, надає можливість паралельно обробляти текстові документи та виконувати семантичну обробку тексту, використовуючи процес видобутку інформації за допомогою системи покриття тексту та знання про моделі лексичної мови.

Література

1. Aref, M.M. A Multi-Agent System for Natural Language Understanding. – International Conference on Integration of Knowledge Intensive Multi-Agent Systems, 2003, 36.
2. 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.
3. 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.
4. Yakovchuk E.I., Sidorova E.A. Generalized semantic-syntactic models in text processing tasks // Proceedings of the workshop "High-tech software NGO-2011". Ershov conference on computer science. –Novosibirsk: ISI SB RAS, 2011. – P. 287–292.
5. Michael Wooldridge. An Introduction to MultiAgent Systems. – University of Liverpool: Wiley, 2009.