UDC 004.42 https://doi.org/10.15407/pp2022.03-04.196

A DIALOGUE SYSTEM BASED ON ONTOLOGY AUTOMATICALLY BUILT THROUGH A NATURAL LANGUAGE TEXT ANALYSIS

Anna Litvin, Vitalii Velychko, Vladislav Kaverynskyi

Створено комплексний підхід до розробки природномовних діалогових систем, в основі яких лежить графова база даних онтологічного типу. Онтологія має визначену регулярну структуру, що містить типізовані семантичні відносини між поняттями, а також пов'язані з ними контексти, що також можуть мати багаторівневу структуру і додаткову типізацію. Онтологія створюється автоматично за рахунок семантичного аналізу природно-мовного тексту за допомогою спеціально розробленої оригінальної програми, яка налаштована насамперед на роботу з мовами флективного типу, зокрема української. Опис онтології зберігається у форматі OWL. Для роботи у складі діалогової системи онтологія переноситься до графової системи управління базами данних Neo4j. Для формальних запитів використовується мова Cypher. Вихідні репліки користувача системи підлягають спеціальному методу семантичного аналізу, за допомогою якого визначається вигляд формального запита до бази даних. Сутність аналізу полягає в тому, що текст фрази користувача проходить через ряд перевірок. За їх результатами визначається набір базових шаблонів формальних запитів, а також додаткові конструкції, що приєднуються до базового шаблону. Певні перевірки можуть також повертати поняття для підстановки у певні зазначені позиції формального запиту. Формальні запити можуть повертати як контексти, так і списки понять з онтології. Окрім понять, запити також можуть повертати інформацію про конкретні семантичні предикати, що їх пов'язують, що спрощує синтез природно-мовних відповідей. Синтез відповідей відбувається за спеціальними шаблонами, вибір яких напряму пов'язаний з відповідним шаблоном формального запиту.

Ключові слова: онтологія, Neo4j, Cypher, аналіз тексту, автоматична генерація онтології, семантичний аналіз, синтез природно мовного тексту, машинна обробка природно мовного тексту, машинне розуміння природно мовного тексту.

An integrated approach is created to the development of natural-language dialogue systems driven by an ontological graph database. Ontology here has a defined regular structure that contains typed semantic relationships between concepts, as well as related contexts, which may also have a multilevel structure and additional typing. The ontology is created automatically due to the semantic analysis of a natural language using a specially developed original software, which is set up to work with inflected languages, in particular Ukrainian. The ontology description is serialized in OWL format. To work as part of the dialog system, the ontology is transferred to the graph database Neo4j. The Cypher language is used for formal queries. The original phrases of the user are subject to a special method of semantic analysis, which determines the type of formal query to the database. The essence of the analysis is that the text of the user phrase goes through a series of checks. Based on their results, a set of basic templates for formal requests is determined, as well as additional constructions that are attached to the basic template. Some of the checks may also return the notion of substitution to certain specified positions of the formal query. Formal queries can return both contexts and lists of ontology concepts. In addition to concepts, queries can also return information about specific semantic predicates that connect them, which simplifies the synthesis of natural language responses. The synthesis of answers is based on special templates, the choice of which is directly related to the corresponding template of the formal query.

Keywords: ontology, Neo4j, Cypher, text analysis, automatic ontology generation, semantic analysis, natural language text synthesis, natural language processing, natural language understanding.

Introduction

Creating a dialog system that can be "trained" using natural language texts without regular structure or prior markup is an important problem. Automation of the process will greatly help to work with a significant amount of information stored as text or collected over the World Wide Web. Such a system could help users to find answers to their questions in the form of the appropriate contexts extracted from the texts or even as conclusions drawn from semantic data obtained from the analyzed text. The current study is devoted to the development of such kind of a dialogue system. The main feature of the proposed method is the automatic building of the ontological graph through the semantic analysis of a natural language text. Another part of the system is the natural language user's interface for the graph database, which provides the conversion of user phrases into formal queries to the ontology. The system also includes a module for the synthesis of natural language responses based on the results of a formal request. It should be noted that the current study is primarily aimed at inflectional languages, which include East Slavic languages, in particular, Ukrainian (for which the examples of implementation of the developed method are given here).

The automatic creation of a database using natural language text in this case can be considered as a particular kind of machine learning. The core of the system is an ontology which is represented as a graph database dedicated to a specific topic. This ontology must have a predefined structure to make easier and more predictable its integration with programs. Nevertheless, the specific content of the ontology is not predetermined and depends on the information from the text submitted as the input data. Thus, the certain results (answers) given by the system and their subject area depend only on the texts used as material for its "learning".

The important notice is that the system proposed in this paper is designed to work primarily with the grammatically and orthographically correct text of scientific and technical style.

Analysis of modern achievements in the field of natural language processing methods for working with ontological knowledge bases

As it was mentioned in the introduction, the design and development of natural language dialogue systems is a complex task, which includes building a database and modules for interaction with it, a semantic analyzer of natural language text, procedures for the answers forming, and providing content in the context of dialogue. The ontology creation is not the main subject considered in the present work. The main topic here is the creation of formal queries and the formation of natural language responses, which mainly form a natural language interface of a graph database. More information on the ontology structure and methods of its automatic creation could be found in the work [1], the problem of staying inside the dialogue context is considered in [2]. Natural language dialogue systems, so-called chat-bots, have a long history and a number of approaches. Below we are to consider some interesting examples of dialog systems developed in recent years, in particular those that in one way or another use an ontology in their structure.

A good example of a natural-language dialogue system is described in works [3, 4]. Like most of the others, it deals with the English language and its structural features. In the means of the analysis of the user's source phrase, it assumes that sentences in English have quite a regular structure that can be expressed through a rather restricted set of templates. The constant part of such a template corresponds to its semantic type ("intention"), the variable parts show the places in the phrase, which concepts to be extracted from. These placeholders are specified according to the certain expected "intentions" of the extracted concepts. For example, there is a template: "Show me {@M} by {@D} for {@V}.". The curly brackets here mark the places where the concepts are expected. The markers in the placeholders here show the following: @M corresponds to the main requested concept, @D is the selected category for concepts such as @M, @V is the filter parameter. For example, there is a phrase "Show me admits by major diagnostic category for 2017", which fully satisfies the above template. The main concept that the user asks to show it is admitted (in this case it is "number of hospitalizations"), the category of selection and sorting is a major diagnostic category (basic diagnostic categories), and the filtering parameter is "2017" in the pattern of which year concept could be guessed. The structure and the constant part of the query determine its "intention". Fore each "intention" there exist a certain package of queries to databases and instructions on how to visualize and present their results in the user interface. Databases containing basic information in this case are mostly relational. However, the system also contains an ontology, which serves to structure the categorization of types and measurements of data stored in the main database. The "intentions" and concepts derived from the source phrase of the user are compared with the ontology to determine the closest to the requested dimensions and categories from those available in the databases. That is, ontology in this case plays quite a secondary role. The ontology is created automatically based on a relational data model. The authors note the ability of the system to stay in the context of dialogue. Their approach is mostly focused on pronouns substitution. If the variables of the analysis template appear pronouns or merely empty, then the program uses the relevant data from the last of the previous queries. In the case when there is no information in previous queries, the default values are substituted. These default values are formed based on the most common requests gathered during the system usage. Currently, the system does not contain automated learning, although the authors have declared the possibility of its development in the future.

The main features of the system from works [3, 4] can be briefly described as follows: works only with English and adapted to its features; analysis of output phrases is based on patterns; not capable of automatic learning; the main data is stored in a relational database, the ontology exists, but plays a supporting role, and is created automatically on the basis of a relational database; has a set of specified "intentions" and related schemes of information presentation (in the form of tables, diagrams and graphs); does not generate natural language responses; implements methods for staying in the context of dialogue.

Dialogue systems which use ontology as the main knowledge base are usually merely natural-language interfaces of a graph database. As a language for formal queries SPARQL is often used. The main task appearing during their development is the conversion of a user's natural language request into a formal one. Below are presented some examples of such converters that have been developed in recent years.

One of the examples of automated conversion of natural language queries into SPARQL frameworks is the PAROT [5]. It uses an approach that generates the most probable RDF triple based on the user's request. The triplet is then checked by a special module containing a dependency analyzer to process user requests to RDF triplets. Then the RDF triplets obtained in this way are to be transformed into ontological triplets using a special thesaurus. The generated ontological triplets are used to build a SPARQL query, which is used for answers obtained from the ontology. Testing of the PAROT framework by the authors [5] showed that for simple questions it shows an accuracy of about 81 – 82 %, for complex ones – about 43 - 56%, and for a specific thematic data set (geography) accuracy raised up to 88%.

Another example of natural language conversion into SPARQL techniques implementation is FREyA [6]. It is available on the GIT-hub [7]. FREyA offers an interactive native language interface for ontology queries. It uses parsing combined with ontology-based search to interpret questions and, if necessary, engages the user. User selection is used to train the system, which improves the accuracy of its operation. This system is currently implemented for English only. In [7] some examples are given which illustrate how questions in natural language could be converted to SPARQL using FREyA. It should be noticed that the FREyA configuration can be tuned for a certain ontology structure.

Also it seems to be worthy to remind the LODQA (Linked Open Data Question Answering) system presented in [8]. It accepts a query in natural language as the input and returns SPARQL queries along with the corresponding responses as a result. The system consists of several modules. The first module processes the request in natural language. It is responsible for parsing and creating a graphical representation of the query, called a pseudographic template. The pseudographic template contains nodes and links. The nodes usually correspond to the basic noun groups and the links to the dependencies between them. In addition, the pseudographic template indicates which node of the ontological graph is the focus of the query, i.e. what the user is going to get as a response to the query. A pseudographic template is a search graph template of a target graph of RDF subgraphs that match it. However, it is called a pseudographic template because it is not yet based on the target data set. No sooner than the first module has generated a pseudographic template from the given natural language query, the next module is activated, which is responsible for finding URIs and nodes values in the pseudographic template. URIs and values must be present in the target data set. To normalize, each node of the pseudographic template is associated with the URI of the dataset. The concept in natural language could be normalized (reduced to the initial grammatical form) in more than one way because of possible ambiguity. Therefore, more than one template could be obtained from one pseudographic template. The third module for the created pseudographic template performs a search in the target data set for the relevant parts, taking into account possible changes that may occur in the data set. To account for the structural differences between the bound pseudographic template and the actual structure of the target data set, this module attempts to generate SPARQL queries for all possible structural variations. SPARQL queries are then sent to the target endpoint, where responses are to be obtained and then sent to the user. These query arguments can be a primitive type, such as S, N or NP, or complex, such as S \ NP, or NP / N. A slash means that the argument should be displayed on the right, and a backslash means that the argument should be displayed on the left. The system uses the following notation of parts of speech, for example: NN – noun, DT – definition (adjective), VB – verb. To facilitate the identification of RDF-triplets, the words in the sentence are lemmatized and assigned with the appropriate grammatical characteristics. The considered LODQA system is focused on working only with English. Detailed features of its functioning in [8] are not given, limited to a general description and analysis of examples of work.

Although the development of dialog systems, as well machine processing and "understanding" of natural language text are mostly carried out for the English language, they are not limited to it. For example, in [9] is presented a dialog system for the German language. This work seems to be interesting because it also involves ontology. In this case, the ontology acts as a dialog manager (OntoDM), which maintains the state of the conversation. Ontology is also used here as a knowledge base. These roles are combined. Subject area knowledge is used to track objects of interest, i.e. ontology nodes (classes) that are products and services represented in the ontological knowledge base. In this way, there was introduced the ability of the conversation's history memory. Also, much of the work [9] is devoted to the peculiarities of linguistic problems of German language processing. By the time of publishing [9], the research work was still proceeding and the quality assessing criteria for the system was not yet obtained. The work [10] is an example of developing a dialogue system for the Korean language, which is fundamentally different from the European type.

One of the most promising graph DBMS is Neo4j [11], which provides fairly high performance and scalability, and is suitable for working with large amounts of data. It is also currently one of the most popular graph DBMS. The language of formal queries adopted in Neo4j is Cypher. It has a wide range of capabilities, is quite flexible and open for extra functionality through plug-ins, for instance, for the implementation of typical algorithms on graphs. However, at present, unlike SPARQL, there are not many developments to convert natural language queries into formal queries on Cypher. Among the few examples could be considered the works [12, 13]. The system proposed in [13] is quite primitive. Requests must have a predefined structure. In fact, this approach is close to that presented in the above-mentioned work [3]: a set of sentence templates in natural language, where some fragments are replaced by special notation, as places from which the concepts are to be extracted for substitution into a query template. Each such template sentence corresponds to a specific query pattern on Cypher. The described approach has its advantages and disadvantages. The main advantage is its simplicity. And the main disadvantage is that a real dialog system requires a large number of such sentences-templates, which include all possible options for asking. Moreover, this approach is justified for languages with a regular sentence structure, such as English, where fewer phrase patterns are needed. Inflective languages, such as Ukrainian, have a complex sentence structure with quite a free word order. This fact significantly increases the required number of templates.

Thus, the **main purpose of this work** was to develop a natural language dialogue system based on ontology, which is created automatically through semantic analysis of natural language text, taking into account the peculiarities of inflective languages, in particular, Ukrainian, and uses Neo4j and Cypher query language to work with its knowledge base.

Below we will consider each of the three main parts of the system: automatic creation of the ontology using natural language text, natural language interface of the graph database and synthesis of answers in natural language using the results of the formal query.

A brief description of the assumed in the system ontology structure

A detailed description of the ontology automatic creation technique based on a natural language text is given in our work [1]. Let us consider the ontology structure itself in the terms of OWL.

The ontology has the following root classes:

- Action – actions expressed by verbs;
- Adjective – adjectives and participles;
- Adverb – adverbs and gerunds;
- Name – proper names;
- Number – numbers (uncertain also) and digit symbols;
- Preposition – prepositions;
- Term – nouns and nouns groups. Has a hierarchical structure from more common (from one word) to more certain terms;
- Negation – negative particles;
- UndefinedEntities – all the entities from the text that class doesn't suitable for any of the ones listed above;
- PhraseType – types of the linked word groups. It has two child classes: MainNarration (the main part of the sentence) and SubordinatePhrase. These classes do not have descendants but are used as "Domain" value for the ontology properties responsible for the groups' characterization.
- SubordinatePhraseType – is used for subordinate phrases classification. For the moment it has the following subclasses: movement_in, actor, Participial, AdverbialPhrase, movement_out, goal, place, consequence, object, subject, cause, condition, instrument. The listed subclasses do not have descendants but are used as «Range» for the properties that have SubordinatePhrase as their «Domain» value.
- SentTypes – sentences typing. It has the following subclasses: Narration, Interrogative та Imperative. These classes do not have descendants but are used as «Range» for the properties responsible for the sentences' characterization.

The properties of the ontology are devised in the following three root groups:
- WordsLink – used for single entities linking;
- Groups – linked word groups;
- SentenceGroups – sentences.

The "WordsLink" property has descendants that match semantic types. In a more primitive version of the ontology, the descendants are merely the semantic types themselves but only those that have been found in parsed text. For example, "action addressing", "object entry", "quality change", "tool", "quantity", "adjacent localization", "destination", "separation", "object-action", "transfer","compatibility", etc. In a more complicated version of ontology, the descendants of "WordsLink" have an additional structure with a hierarchy. The scheme of "WordsLink" descendants structure is given as a tree in figure 1.

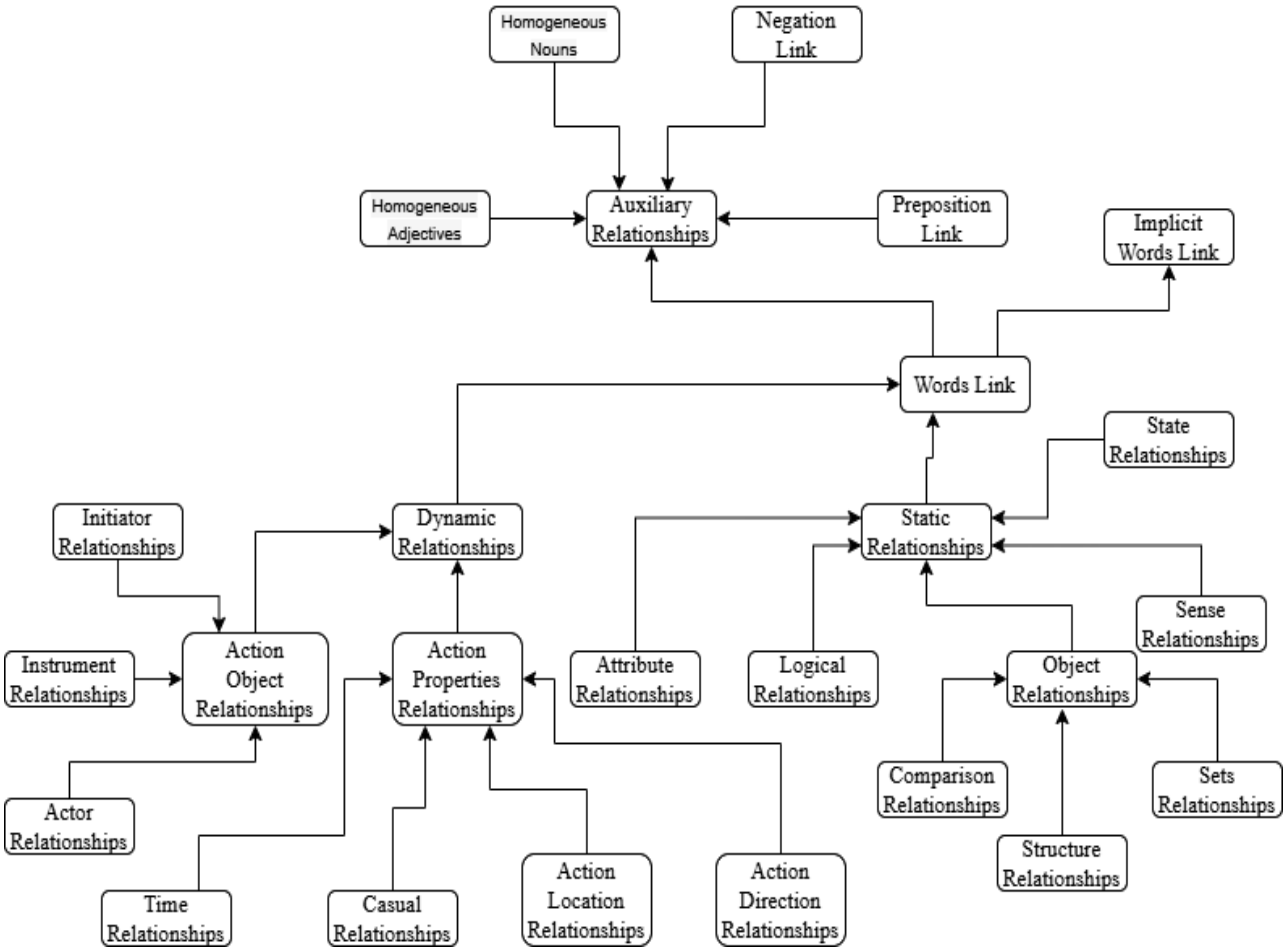

Figure 1. – Scheme of "WordsLink" properties higher-level hierarchy in the ontology

The given here structure covers only the higher level of the semantic relationships hierarchy typing that remains the same for all built ontologies. The presence of lower-level entities depends on their presence in the considered text. They also could be hierarchy structured. For the moment the developed system operates with about 80 possible final semantic categories and this is obviously not the limit.

The final descendants of "WordsLink" property correspond to the specific types of connections between certain concepts. Each of them occurs only once in the ontology, even if it could be found several times in the considered text. "Domain" of such property refers to the main concept of the linked pair, and "Range" to the dependent one. In addition, these properties are also heirs of the groups where the pair linked in this way is observed.

The "Groups" property characterizes groups of linked words. The descendant properties of Groups correspond to certain groups. Sub-properties of the groups are the above-mentioned properties, which show the connections between the concepts. The sub-properties of "SentenceGroups" correspond to sentences. As a label parameter, they contain the full text of the sentence (context). Their descendants are properties of the "Groups" type that correspond to the groups in the given sentence.

Neo4J DBMS could be used to work with the ontology of the described type. For this purpose, an OWL file is to be loaded to it using "Neosemantics" plug-in. In this case, classes and properties become the graph nodes of the corresponding type which are "Class" and "Relationship". Relationships between the nodes can have the following types: SCO – subclass of; SPO – sub-property of; DOMAIN; RANGE. The Cypher language is used for the queries.

Building an ontological graph based on natural language text is perhaps the most important part of a system that is responsible for collecting and structuring information. The construction of a semantically structured database requires semantic analysis of the considered text. Thus, an important part of the study was the development of its methodology, adapted for the East Slavic languages, which are of inflectional type. The important peculiarity of these languages is that words connection appear mainly mostly through a combination of certain flections (variable word endings). Behind these variations of word forms, which belong to the relevant parts of speech and the combination of concepts with prepositions, there lies a huge amount of semantic information. Moreover, there are other factors, such as word order. The order of words in inflectional languages is not very strict and might be considered as quite a secondary factor. Nevertheless, the words even in such types of languages do not go completely randomly. Moreover, in some cases, it may even become even a defining feature.

Analysis of the user's input phrase for the formal queries to the ontology creation

As proposed in our previous work [14] tree-based method of the query template determining through the analysis of words sequence is quite demanding for the effort and time needed for such a tree development. At the same time in inflective languages the word order is a less significant factor. A more valuable one is just presence of the certain words in specific forms. Thus, it seems that often enough would be merely testify the considered phrase through a number of criteria. Grounding on the test results it might be possible not only to determine the most appropriate formal query template or the group of such templates but also to select the input entities for them. In the simplest test version of the system, which exists now, there are the 4 following main checks:

1 – question word – 6 lists + absence of such word. The result is the number of the sufficient list from 1 to 6 or 0, if there is no question word in the sentence.

2 – the presence of a word from given lists (most of them are specific verbs) – 6 lists + absence of words from all of the lists. The result is the number of the sufficient list from 1 to 6, of 0 – if there are no such words in the sentence.

3 – the presence of a noun in the nominative case except words from the check (2) if any. The result may be 1 – such a word exist (+ the word itself) or 0 – there is no such word. Several entities could be selected.

4 – the presence of a verb, except ones from lists in the check (2) if any. The result may be 1 – such a word exist (+ the word itself) or 0 – there is no such word. Several entities could be selected.

Even this brief test has quite enough options for its results that make it possible to have a number of templates or various types of templates.

Then an additional test is to be performed. Its procedure is as follows: adjectives linked to the word from the clause (3) that must be close to it and fit it with number and gender; nouns in indirect cases (they form the base of the additional circumstances) and adjectives linked to them; and the last but not the lease is the check of presence or absence of negation predicates. An additional test is needed for the modifier templates adding.

The template are stored as XML-files of a special structure. Here is an example of one of such (the simplest ones) templates:

```
<template>
 <verbose_name>Common information</verbose_name>
 <id>1</id>
 <type>base</type>
 <variables>
       <variable>
             <name>INPUT_VALUE_1</name>
             <destination>input</destination>
       </variable>
       <variable>
             <name>CONTEXT</name>
```

```
            <destination>output</destination>
      </variable>
</variables>
<match>
      (inp:Class)-[]-(n:Relationship),
      (n:Relationship)-[]-(x:Class),
      (n)-[:SPO]->(rel_group),
      (rel_group)-[:SPO]->(rel_sent),
      (rel_sent)-[:SPO]-(sent_super)
</match>
<where>
      inp.label = "INPUT_VALUE" and
      sent_super.name = "SentenceGroups"
</where>
<return>
      DISTINCT rel_sent.label as CONTEXT;
</return>
</template>
```

In the given example it is possible to explain the common structure of the query template. The XML-template chapters <match>, <where>, and <return> correspond to the certain sections of a Cypher query [11]. Some parts of the chapter's content are the template variables. The variables themselves are described in the chapter <variables>. For each of the variables are defined its name and destination in the appropriate XML-containers <name> and <destination>. The destination can have values "input" or "output". The input variables are to be substituted with the input parameters values and the output ones define the parameters that should be obtained as a result of the query execution. The container <id> is needed for the finding and identity of the template. Moreover, here is a tag <verbose_name> that helps to identify a template not only by machine but also by a human during the system development. Further, most of the query template examples here shall be given in a simplified mater – without XML-tags. Tag <type> shows the type of a template – base or additional. Above is given an example of a base one. Let us consider the structure of the additional templates. Here is an example of one of them:

```
<template>
<verbose_name>Adjective linked to subject</verbose_name>
<id>1</id>
<type>additional</type>
<variables>
      <variable>
            <name>INPUT_VALUE_ADJ</name>
            <destination>input</destination>
      </variable>
      <variable>
            <name>ADJ_PLUS</name>
            <destination>intermediate</destination>
      </variable>
      <variable>
            <name>INP_ADJ</name>
            <destination>intermediate</destination>
      </variable>
</variables>
<block_union>and</block_union>
<next_item_union>or</next_item_union>
<match>
      (inp:Class)-[]-(ADJ_PLUS:Relationship),
      (ADJ_PLUS:Relationship)-[]-(INP_ADJ:Class),
      (ADJ_PLUS)-[:SPO]->(rel_group)
</match>
<where>
      INP_ADJ.label = "INPUT_VALUE_ADJ"
</where>
<return></return>
</template>
```

The template also has blocks <match>, <where> and <return>. However, the content of them is not independent but is to be added to the appropriate parts of a query formed through a base template. Some of the chapters in this case could be merely empty. The main feature of an additional template are presence of the chapters <block_union> and <next_item_union>. Tag <block_union> shows the manner of how the block <where> must be united to the query formed by a base template. Tag <next_item_union> determines the union type for the repeated elements of the block <where> in

a case when the appropriate variable is presented as a list (array). For instance, for the given above template, the variable INPUT_VALUE_ADJ could correspond to a number of adjectives linked with the object. The values of <block_union> and <next_item_union> could be "and" or "or". Also the variables of the additional templates can have the third type of <destination> - "intermediate". Such variables neither take part in transferring values into the forming query nor in the results returning. They are just needed to mark the template parts that are not to be duplicated during the part repeating. Instead, they are implemented with an order number, for example: ADJ_PLUS_1, ADJ_PLUS_2, ADJ_PLUS_3, …, etc.

Let us consider in more detail the structure of the formal queries and the manner of their formation. The presented structure of the ontology makes it possible to search for contexts or individual terms. Not only has it allowed just the presence of some entities in the context considering, but also their relationships according to a certain semantic category. In the presented scheme there are a base query template, aimed to obtain information of a certain type in a given form, and additional modifiers templates that optionally adds the description of extra circumstances. Let us consider some types of queries. The already given above template is aimed to a context obtaining which includes a specific term (word). However, the term must not only be presented in the context but form a link with others. This could guarantee that the term is "organically" implemented into the context.

Cypher queries are devised into three main parts: MATCH, WHERE, and RETURN. The MATCH block gives a linking pattern of the nodes in the oriented graph. In the WHERE part the conditions are given that characterize the entities (nodes and relationships) from the MATCH case. The RETURN block shows what is to be returned as a result and with what name (alias). In the presented example there is a class marked by the variable "inp". It the WHERE block a condition for it is added, which says that the "label" field of the node "inp" must be equal to a specific value (here and below INPUT_VALUE is the text of the input value). From the MATCH block, it is clear that "inp" is a node because of parentheses and it must have the type "Class". It must be linked with another node "n" of type "Relationship", which corresponds to an ontology property from OWL. The link type is undefined in this case (square brackets are empty), and the direction of the link is also not specified. So, the node could be linked either as a "DOMAIN" or "RANGE". There is no need to specify the link direction in this case because it is known that such links always come from a property to a class. Also it is given that this property must be linked with some class "x". Further is given that the property linking this classes must have a relation to a sentence "rel_sent". The condition "sent_super.name = «SentenceGroups»" guarantees that the "rel_sent" shall be a sentence. As a result of the query is to be returned "rel_sent.label", which contents the sentence context with the alias "CONTEXT".

Let us come to a more complicate example. Here we are to request the characteristics (properties) of an INPUT_VALUE entity included in the ontology. . Ми хочемо запросити відомі в онтології характеристики (визначення) об'єкта INPUT_VALUE. In other words, what the INPUT_VALUE is or could be. The query is as follows:

```
MATCH (inp:Class)-[]-(n:Relationship),
    (n:Relationship)-[]-(x:Class),
    (n)-[:SPO]->(prop_type_1),
    (n)-[:SPO]->(rel_group),
    (rel_group)-[:SPO]->(rel_sent),
    (rel_sent)-[:SPO]-(sent_super)
WHERE
    inp.name = "INPUT_VALUE" and
    (prop_type_1.label = "object propery" or
    prop_type_1.label = "action property" or
    prop_type_1.label = "action separately" or
    prop_type_1.label = "action level")
     and
    sent_super.name = "SentenceGroups"
RETURN DISTINCT x.label as result,  rel_sent.label as context;
```

Compared to the previous example an extra statement is added to the MATCH block: (n)-[:SPO]->(prop_type_1). This gives information that the property "n" must be a child of "prop_type_1". Here the link direction is specified. In the WHERE block is given sufficient values of "label" field of "prop_type_1". To make the query template more universal, as it is not known whether INPUT_VALUE is noun or verb, a number of options are given for the possible "prop_type_1.label" value united with logical "OR". If the ontology has a semantic categories hierarchy, the construction could be simplified as follows:

```
MATCH (inp:Class)-[]-(n:Relationship),
    (n:Relationship)-[]-(x:Class),
    (n)-[:SPO]->(prop_type_1),
    (n)-[:SPO]->(rel_group),
    (rel_group)-[:SPO]->(rel_sent),
    (rel_sent)-[:SPO]-(sent_super),
    (prop_type_1)-[:SPO]->(prop_type_category)
WHERE
    inp.name = "INPUT_VALUE" and
    prop_type_category.label = "entities properties"
     and
    sent_super.name = "SentenceGroups"
RETURN DISTINCT x.label as result,  rel_sent.label as context;
```

As a result of the query "label" field of "x" node is to be returned. That will be the characteristics of an "inp" object. Also the contexts are requested to recognize the circumstances where the entity's property is mentioned.

In a close manner actions of an object could be requested. For this purpose, it is just needed to set another value for "prop_type_1.label" in WHERE block, namely: `prop_type_1.label = «object-action»`.

If there are several possible options of relationship in the query (prop_type_1.label) the result may include its certain value, which then helps in the answer synthesis. The next example illustrates a query of an object localization without its type concretization ("Where is INPUT_VALUE?").

```
MATCH (inp:Class)-[]-(n:Relationship),
     (n:Relationship)-[]-(x:Class),
     (n)-[:SPO]->(prop_type_1),
     (n)-[:SPO]->(rel_group),
     (rel_group)-[:SPO]->(rel_sent),
     (rel_sent)-[:SPO]-(sent_super)
     (prop_type_1)-[:SPO]->(prop_type_category)
WHERE
      inp.label = " INPUT_VALUE " and
     prop_type_category.label = "localization" and
     sent_super.name = "SentenceGroups"
RETURN DISTINCT x.label as result, rel_sent.label as context,
            prop_type_1.label as predicate;
```

The main peculiarity here is the statement "prop_type_1.label as predicate" in the RETURN block. That makes it to return the certain semantic type of the obtained result.

In some cases instead of predicates lists of some entities (verbs, nouns, adjectives) could be included in a query. The peculiarity here is that conditions are given for the node of ontogrhaph linked with "x". Thus, the requested object not only must be linked with some term "x" through the specific relationship, but this term must be from a certain list. If the terms (or actions) are additionally classified in the ontology, the condition for the term will be merely being a descendant of a specific category.

A special mention should be made of modifier templates – fragments that could be added to the main query templates. Let us consider an example where the input parameter is not a single word, but a noun group. So, there are linked nouns and adjectives. To link to the input adjective concept there must be added the appropriate statements to the MATCH block:

```
     (inp:Class)-[]-(adj_plus:Relationship),
     (adj_plus:Relationship)-[]-(inp_adj_1:Class),
     (adj_plus)-[:SPO]->(rel_group)
```

and in WHERE block:

```
      and
     inp_adj_1.label = "INPUT_VALUE_ADJ"
```

For the extra adjectives the same blocks are to be added but with variables inp_adj_2, inp_adj_3 etc.

It is also possible to add a condition of a noun in indirect case presence through the following statements:

in MATCH block:

```
     (inp_noun_1:Class)-[]-(noun_plus:Relationship),
     (noun_plus)-[:SPO]->(rel_group)
```

and inWHERE block:

```
      and
     inp_noun_1.label = "INPUT_VALUE_NOUN"
```

Here in the example there is a condition of presence of one noun in the same group where the main concept is included. Nevertheless, conditions of adjectives presence linked with this noun also could be added:

in MATCH block:

```
     (inp_noun_1:Class)-[]-(adj_plus_add:Relationship),
     (adj_plus_add:Relationship)-[]-(inp_adj_add:Class),
     (adj_plus_add)-[:SPO]->(rel_group)
```

and in WHERE block:

```
      and
     inp_adj_add.label = "INPUT_VALUE_ADJ_ADD"
```

Also in some cases a negation predicate should be added to a query. For this purpose the following construction must be added to it:

in MATCH block:

```
     (neg:Class)-[]-(neg_rel:Relationship),
     (neg_rel)-[:SPO]->(rel_group)
```

and in WHERE block:

```
      and
     (neg.label = "no" or
 neg.label = "not" or
 neg.label = «forbidden» or
```

```
neg.label = «impossible» or
neg.label = «cant» or
neg.label = "unable")
```

Synthesis of natural language answers based on the results of formal queries execution

The user interface of a dialog system, which displays merely the results of a formal query, even being pretty designed, may not look so friendly, and sometimes could be even not quite understandable for a person. Therefore, the next important problem is the synthesis of natural language answers. Some principles of the approach of answers formation, based on information taken from the results of formal queries, and the analysis of the source phrase using templates-instructions are described in our previous work [14]. In general, during the system development, making the decision of how the answer ought to appear in the user's interface is to be balanced between providing ready-made contexts and text synthesis. For example, to provide some tables, or graphical objects, or other media illustrating the answer, the best option is to use ready-made contexts containing links to the relevant files. In the current study, we omit representation and creation methods of graphical and tabular materials (charts, graphs, diagrams) based on the results of queries in the user interface, although this approach is quite desirable in certain types of systems and, as demonstrated by [4], may well be implemented. Contextual responses may be the best option if you need to provide detailed information. The synthesized answers provide greater ease of perception for more specific questions, which formal response is just a list of entities from the ontology. Here are provided some examples of answers synthesizing instruction templates for some typical cases. These templates also give user contexts (sentences) that illustrate and confirm the statement. The templates below are presented in human-readable form (a kind of meta-language). In a software implementation, they are software entities (classes with methods) in the Python language that are attached to the system in a specific module file. An attempt was also made to add response templates in the form of XML descriptions, which, however, led to greater complexity and lower performance of the software.

Let us consider an example of a question about entity characteristics (properties). Here is the answer template:

```
Repeat for each result:
      if INPUT_VALUE noun:
            INPUT_VALUE + може бути + result (fit the genger)
            + context
      is INPUT_VALUE verb:
            INPUT_VALUE + можна + result
            + context
```

For the word's morphological characteristics determination (part of speech, gender, case, etc.) and for word form fitting PyMorphy2 library methods are used [16]. In the simple example above part of speech of INPUT_VALUE must be checked. It could be a noun or verb. If it is a noun, the "result" value must be fitted in gender with INPUT_VALUE.

Let us consider a more complicated example. Here the subject of the query is an object localization. The certain localization predicate is not specified in the input query parameters, but appears in its results. As it was mentioned above, a certain semantic predicate could be used in an answer synthesis.

```
Repeat for each result:
INPUT_VALUE + знаходиться +
      if predicate = "localization in set":
            + серед + result (plural, genitive case)
            + context
      if predicate = «localization near»:
            + біля + result (genitive case)
            + context
      if predicate = "objective localization":
            + на + result (locative case)
            + context
      if predicate = " objective entering":
            + у + result (locative case)
            + context
      if predicate = "localization between objects":
            + між + result (plural, instrumental case)
            + context
      if predicate = "localization behind object":
            + за + result (instrumental case)
            + context
      if predicate = "localization in front of object":
            + перед + result (instrumental case)
            + context
      if predicate = "localization under object":
            + під + result (instrumental case)
            + context
```

```
if predicate = "localization above object":
      + над + result (instrumental case)
      + context
if predicate = "localization in object":
      + всередині + result (genitive case)
      + context
```

From the given example we can see that the certain type of semantic predicate (localization in this case) determines the appropriate preposition and case for the value of "result" variable for the Ukrainian language.

Conclusions and further prospective

An approach and the corresponding software toolkit are developed for the construction of natural-language dialogue systems on the basis of automatically through a natural-language text semantic analysis built ontology. Within the framework of this approach an analysis technique is developed. It deals with an initial user's phrase adapted for inflective languages, in particular Ukrainian, aimed at formation on its basis of formal queries in the Cypher language. The essence of the method is a series of checks for the presence in the initial phrase of certain words and/or word forms. Depending on the set of the test results, the main query template (or group of such templates) is selected. Components from modifier templates are added to the main template (to its corresponding sections) as a result of additional checks, which make the appropriate clarifications and extensions to the query. Query variables are supplemented with concepts obtained when performing the appropriate checks. A number of queries (package) can be created on the basis of one initial phrase. Also proposed here is an approach to the synthesis of natural language responses using query results and the values of source variables. The peculiarity of the approach is the usage of including specific values of semantic predicates obtained as a result of the query to the ontology, which allows the program more accurately and correctly formulate the answer by using the appropriate prepositions and word forms. Also, these answer templates provide instructions for fitting word forms of concepts-results with the original concepts.

Based on the proposed approach, an experimental dialogue system was developed, which proved to be workable. It can become a prototype for the development of new more powerful dialogue styles able to be "learned" using natural language texts provided in the form of documents, or as search results obtained from the Internet. Further improvement of the system is to use the opportunity to create more detailed classified ontologies, expand the number of checks and variants of their results. Accordingly, a large number of basic and additional formal query templates and corresponding response synthesis templates can be created.

Bibliography

1. Litvin A. A new approach to automatic ontology creation from untagged text on natural language of inflective type / A. Litvin, V. Velychko, V. Kaverinsky // International conference on software engineering "Soft Engine 2022". – Kyiv, NAU. – 2022.
2. Litvin A. Development of natural language dialogue software systems / A. Litvin, V. Velychko, V. Kaverinsky // Information Theories and Applications. – Vol. 28. – No. 3. – 2021. – P/ 233 – 270.
3. Quamar A. Conversational BI: An Ontology-Driven Conversation System for Business Intelligence Applications / Abdul Quamar , Fatma Ozcan, Dorian Miller , Robert J. Moore , Rebecca Niehus, Jeffrey Kreulen // Proceedings of the VLDB Endowment, Vol. 13, No. 12. – P. 3369 – 3381.
4. Quamar A. An Ontology-Based Conversation System for Knowledge Bases / Abdul Quamar, Chuan Lei, Dorian Miller , Fatma Özcan1, Jeffrey Kreulen, Robert J. Moore1, Vasilis Efthymiou // Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. – 2020. – P. 361 – 375. doi:10.1145/3318464.3386139.
5. Ochieng P. PAROT: Translating natural language to SPARQL / P. Ochieng // Expert Systems with Applications. – 2020. – № 5. – P. 1–16.
6. Damljanovic D. FREyA: an interactive way of querying linked data using natural language / D. Damljanovic, M. Agatonovic, H. Cunningham // The Semantic Web: ESWC 2011 Workshops. – 2011. – P. 125–138.
7. GIT-hub: FREyA documentation [Електронний ресурс]. Режим доступу: https://github.com/nmvijay/freya
8. Shaik S. Transforming natural language query to SPARQL for semantic information retrieval / S. Shaik, P. Kanakam, S.M. Hussain, D. Suryanarayana // International Journal of Engineering Trends and Technology. – 2016. – № 7. – P. 347–350.
9. Duygu Altinok An Ontology-Based Dialogue Management System for Banking and Finance Dialogue Systems / Duygu Altinok // 1st Financial Narrative Processing Workshop. - Japan, Miyazaki. – 2018. https://doi.org/10.48550/arXiv.1804.04838
10. Jung H. Automated conversion from natural language query to SPARQL query / H. Jung, W. Kim // Journal of Intelligent Information Systems. – 2020. – Vol. 55. – P. 501–520.
11. Goel A. Neo4J Cookbook / A. Goel // Birmingham: Pact Publishing Ltd. – 2015. – 206 p.
12. Sun C. A Natural Language Interface for Querying Graph Databases: master's thesis … master in computer science and engineering / C. Sun. – USA: Massachusetts Institute of Technology, 2018. – 69 p.
13. GIT-hub Convert English sentences to Cypher queries documentation [Електронний ресурс]. Режим доступу: https://github.com/gsssrao/english2cypher
14. Litvin A. Synthesis of chat-bot responses in the inflecting natural language based on the results of queries to ontology and analysis of the chat previous phrase // A. Litvin, V. Velychko, V. Kaverinsky // Information Theories and Application. – Vol. 27, No. 2. – P. 152 – 199.

References

1. A. Litvin, V. Velychko, & V. Kaverinsky (2022) A new approach to automatic ontology creation from untagged text on natural language of inflective type. ***International conference on software engineering "Soft Engine 2022"***. Kyiv, NAU. Available from: http://pp.isofts.kiev.ua/ojs1/ article/view/145 [Accessed 6/06/2017].
2. A. Litvin, V. Velychko, & V. Kaverinsky (2021) Development of natural language dialogue software systems. ***Information Theories and Applications***. 28. p. 233 – 270.
3. Quamar Abdul, Fatma Ozcan, Dorian Miller , Robert J. Moore , Rebecca Niehus & Jeffrey Kreulen Conversational BI: An Ontology-Driven Conversation System for Business Intelligence Applications. ***Proceedings of the VLDB Endowment***. 13. p. 3369 – 3381.

4. Abdul Quamar, Chuan Lei, Dorian Miller , Fatma Özcan1, Jeffrey Kreulen, Robert J. Moore, Vasilis Efthymiou (2020) An Ontology-Based Conversation System for Knowledge Bases ***Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data***. p. 361 – 375. DOI:10.1145/3318464.3386139.
5. P. Ochieng PAROT: Translating natural language to SPARQL. (2020) ***Expert Systems with Applications.*** 5. p. 1–16.
6. D. Damljanovic, M. Agatonovic & H. Cunningham (2011) FREyA: an interactive way of querying linked data using natural language. ***The Semantic Web: ESWC 2011 Workshops***. p. 125–138.
7. GIT-hub: ***FREyA documentation*** [Online] Available from: https://github.com/nmvijay/freya
8. S. Shaik, P. Kanakam, S.M. Hussain & D. Suryanarayana (2016) Transforming natural language query to SPARQL for semantic information retrieval. ***International Journal of Engineering Trends and Technology***. 7. p. 347–350.
9. Duygu Altinok (2018) An Ontology-Based Dialogue Management System for Banking and Finance Dialogue Systems. ***1st Financial Narrative Processing Workshop***. Japan, Miyazaki. DOI: https://doi.org/10.48550/arXiv.1804.04838
10. H. Jung, W. Kim (2020) Automated conversion from natural language query to SPARQL query. ***Journal of Intelligent Information Systems.*** 55. p. 501–520.
11. A. Goel (2015) ***Neo4J Cookbook.*** Birmingham: Pact Publishing Ltd.
12. C. Sun A. (2018) ***Natural Language Interface for Querying Graph Databases: master's thesis master in computer science and engineering***. USA: Massachusetts Institute of Technology.
13. GIT-hub ***Convert English sentences to Cypher queries documentation*** [Online] Available from: https://github.com/gsssrao/english2cypher
14. A. Litvin, V. Velychko, & V. Kaverinsky (2020) Synthesis of chat-bot responses in the inflecting natural language based on the results of queries to ontology and analysis of the chat previous phrase. ***Information Theories and Application***. 27. p. 152 – 199.

Received 16.07.2022

About the authors:

Litvin Anna Andreevna,
Post-graduate student in the V.M.
Glushkov Institute of Cybernetics NAS of Ukraine.
The number of scientific publications in Ukrainian journals is 7.
The number of scientific publications in foreign journals is 3.
http://orcid.org/0000-0002-5648-9074

Velychko Vitalii Yuriiovych,
Doctor of Sciences, assistant professor,
Senior researcher in the V.M. Glushkov Institute of Cybernetics NAS of Ukraine,
visitor leading researcher of the department of creation and use
of intellectual network tools of the National Center
"Junior Academy of Sciences of Ukraine"
The number of scientific publications in Ukrainian journals is 80.
The number of scientific publications in foreign journals is 31.
H-index: Google Scholar – 11,
Scopus – 2,
http://orcid.org/0000-0002-7155-9202.

Kaverynskyi Vladislav Vladimirovich,
Ph.D. in technical sciences,
Senior Researcher of Department of Wear-Resistant
and Corrosion-Resistant Powder Construction Materials
in the I. N. Frantsevich Institute for Problems
of Materials Science NAS of Ukraine.
The number of scientific publications in Ukrainian journals is 88.
The number of scientific publications in foreign journals is 22.
H-index: Google Scholar – 6
Scopus – 2,
http://orcid.org/0000-0002-6940-579X

Place of work:

Litvin Anna Andreevna:
V.M. Glushkov Institute of Cybernetics NAS of Ukraine.
03187, Kiev-187, Academician Glushkov Avenue, 40.
Phone: (097) 570-99-84, E-mail: litvin_any@ukr.net

Velychko Vitaliy Yurievich:
V.M. Glushkov Institute of Cybernetics NAS of Ukraine.
03187, Kiev-187, Academician Glushkov Avenue, 40.
Phone: (096) 139-96-28, E-mail: aduisukr@gmail.com

Kaverynskyi Vladislav Vladimirovich:
I. N. Frantsevich Institute for Problems of Materials Science NAS of Ukraine.
03142, Kiev, Academician Krzhizhanovsky st., 3.
Phone: (050) 212-17-24, E-mail: insamhlaithe@gmail.com

Прізвища та ініціали авторів і назва доповіді українською мовою:
А.А. Літвін, В.Ю. Величко, В.В. Каверинський
Природномовна діалогова система на основі онтології,
що побудована на базі автоматизованого семантичного аналізу тексту

Прізвища та ініціали авторів і назва доповіді англійською мовою:
Litvin A.A., Velychko V.Yu., Kaverynskyi V.V.
A dialogue system based on ontology automatically built through
a natural language text analysis

Контакти для редактора: Каверинський Владислав Володимирович,
старший науковий співробітник
Інституту проблем матеріалознавства НАН України,
e-mail: insamhlaithe@gamil.com, тел.: (38)(050) 212-17-24