Ontological Approach to the development of natural language interface for intelligent computer systems

Longwei Qian
Department of Intelligent Information Technology
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
qianlw1226@gmail.com

***Abstract*—Natural language interfaces are oriented towards achievement of interaction between human users and computer systems in the most natural way. In the process of artificial intelligence development, natural language interface has always been the main research direction. The paper is dedicated to the description of the ontological approach to the development of the natural language interface for intelligent computer system (mainly knowledge-based system) based on the OSTIS Technology. In the field of the development of natural language interface the existing approaches to solve the natural language texts processing (natural language texts analysis and natural language texts generation) are considered. Moreover, the ontological model that possibly integrate different kinds of linguistic knowledge in detail and various problem solving models oriented on solving natural language texts processing in a semantically compatible are described.**

***Keywords*—OSTIS, ontology, knowledge-based system, natural language interface, knowledge-driven, knowledge base**

I. Introduction

In the field of artificial intelligence research, recently knowledge-based intelligent systems are widely applied in various fields (e.g. in medical, in education, etc.). As one of the key components of this type of intelligent systems, intelligent user interfaces aim to enable more intelligent interactions between human users and intelligent systems.

In traditional computer systems, the user interfaces in general must have database access capabilities. In accordance with the various means of interaction used by human users, in traditional computer systems, user interfaces usually can be divided into the following types:

- user interfaces with graphical query;
- user interfaces with formal languages query;
- user interfaces based on filling in request forms.

For user interfaces with graphical query, information is exchanged between computer systems and human users through the graphical components (e.g. menus, controls) in the interfaces. For user interface with formal languages query, information is exchanged between computer systems and human users by writing commands in formal languages. Graphical components or commands in formal languages are generally complex and unnatural for the human users. Users must understand the functions of each graphical component in interface or learn the grammar of a certain formal language in advance. For user interfaces based on filling in request forms information is exchanged between computer systems and human users through the completion of specific forms. For form-based interfaces, users need to fill out the forms correctly according to specific requirements. For user interfaces of traditional computer systems, users require additional training (understanding graphical components, learning specific formal language) to interact with computer systems. Therefore human users expect to use more convenient interactive ways without learning complex operations. The natural language interfaces are closest to the natural form of human communication, and can provide human users with more natural interactive experience.

The natural language interfaces of knowledge-based intelligent systems are oriented to achievement of information exchange between natural language texts (especially declarative sentences) provided by human users and knowledge base of intelligent systems in which store specific knowledge. The modular scheme (Figure 1) shows automated information interaction process between human users and intelligent systems based on the classical architecture of modern question-answer systems based on the knowledge base.

As can be seen from the Figure 1, the development of the natural language interfaces of the knowledge-based intelligent systems mainly involves the realization of the following two components:

- conversion of input natural language handwritten texts into knowledge base fragments of intelligent systems;
- generation of natural language handwritten texts from the knowledge base fragments of intelligent systems.

Both conversion of natural language handwritten texts

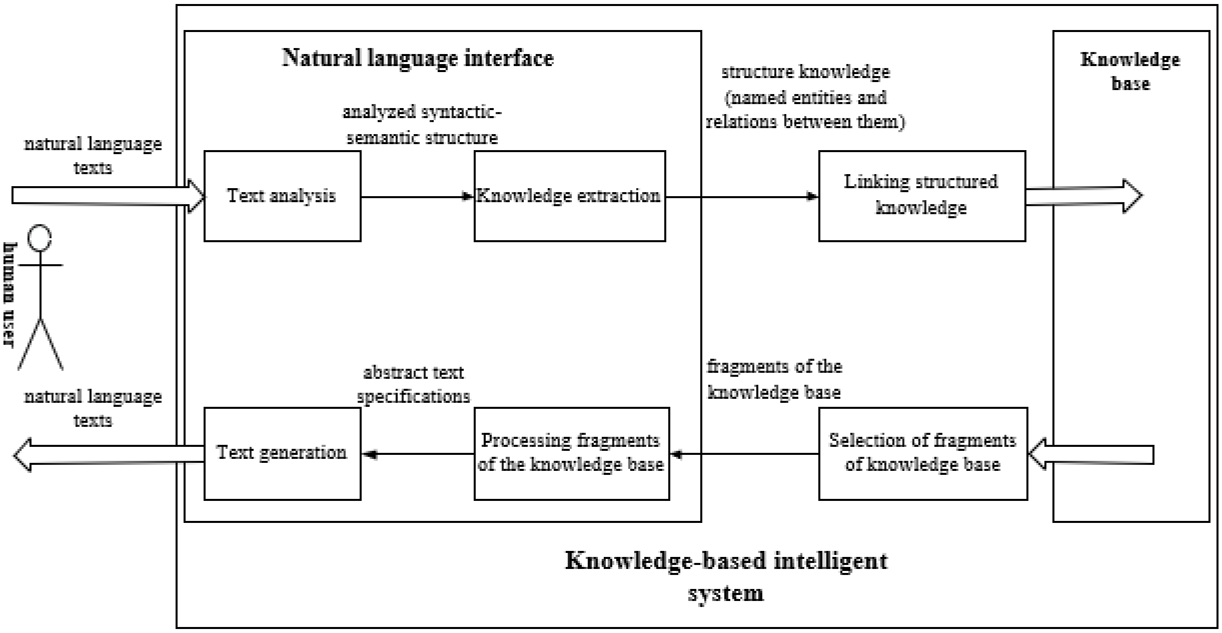

Natural language interface
Knowledge base
natural language texts
analyzed syntactic-semantic structure
structure knowledge (named entities and relations between them)
Text analysis
Knowledge extraction
Linking structured knowledge
human user
abstract text specifications
fragments of the knowledge base
natural language texts
Text generation
Processing fragments of the knowledge base
Selection of fragments of knowledge base
Knowledge-based intelligent system

Figure 1: Modular scheme of natural language interface of knowledge-based intelligent systems

and generation of natural language handwritten texts are considered subtasks of natural language processing. From the point of view of intelligent systems, these subtasks are kinds of so-called complex problems, the solution of which are still hot research topics. In general, solution of these subtasks requires a combination of various types of knowledge about linguistics and the use of various problem solving models for natural language processing. However in modern intelligent systems, there is a lack of a unified basis to integrate various types of knowledge about linguistic and problem solving models in the development process of natural language interfaces.

II. CLASSIFICATION OF NATURAL LANGUAGE INTERFACE

The development of natural language interfaces for knowledge-based intelligent systems generally requires consideration of following two aspects:

- types of processed natural language, i.e. characteristics of different types of natural language;
- the range of the knowledge base of intelligent systems, i.e. the breadth of knowledge in the knowledge base of intelligent systems.

According to statistics, on the Internet there are only dozens of natural languages that are widely used by human users for communication, such as Russian, English, Chinese, etc. Each natural language has its own unique features, for example, the writing habits of Chinese language are almost completely different from European languages. Therefore for processing various natural language texts, it is necessary to take into account the corresponding characteristics of different natural languages. In addition, the range of the knowledge base also affects the complexity of developing natural language interfaces of knowledge-based intelligent systems. The world-wide knowledge bases [1], [2], in which knowledge sources are oriented to the whole Internet, i.e., commonsense knowledge [3] is stored in the knowledge bases of intelligent systems. The knowledge bases about specialized domain [3], [4], in which knowledge sources are oriented to various encyclopedias (for example, an automobile encyclopedia, an encyclopedia about films, an encyclopedia about various disciplines (such as discrete mathematics, history, and others)), i.e. in the knowledge bases of intelligent systems stores specific industry knowledge (i.e., knowledge about a limited domain). The commonsense knowledge base draws attention to the breadth of stored knowledge, emphasizes the integration of more concepts, named entities across different domains and the relationships between them. from another aspect, the commonsense knowledge base will not provide high accuracy of the knowledge stored in it. With the process of integrating more concepts, named entities and relationships between them in the knowledge base it will lead to higher complexity of searching and processing fragments of the knowledge base. Unlike the commonsense knowledge base, knowledge bases about specialized domain draws more attention to the depth and accuracy of the stored knowledge that satisfies the solution of various specific tasks in intelligent systems with lower complexity.

In accordance with the two aspects listed above, we divide the natural language interfaces of knowledge-based intelligent systems into the following classes:

- natural language interface that is independent of the specific natural language and the specific domain. In this case, this kind of natural language interface can process texts of any kind of natural languages, for example, Russian, Arabic, English, Chinese, and so on, as well as process any complex knowledge structure, i.e. knowledge is stored in the world-wide knowledge bases;

- natural language interface that is dependent on the particular natural language, but is independent of the particular domain. In this case, the natural language interface can only analysis texts of certain natural language and process any complex knowledge structure;
- natural language interface that is independent of the particular natural language, but is dependent on the particular domain. In this case, the natural language interface can analyze texts of any kind of natural languages, which describe facts about the particular domain. The knowledge bases about specialized domains are the basis of the knowledge base of the intellectual systems. It should be noted that the description of facts in different domains may special different types of sentences. Therefore the processing of natural language texts describing facts in different domains (for example, the historical texts, the legal texts, the texts in different disciplines) has the different degree of complexity;
- natural language interface that is dependent on the particular natural language, and as well as is dependent on the particular domains. This kind of interface is considered the most basic natural language interface in knowledge-based intelligent systems. In this case, the natural language interface only needs to analysis texts of certain natural language and process the simpler knowledge structure in knowledge base about the specialized domain.

The development of the natural language interfaces for knowledge-based intelligent systems in a multilingual field requires the study of the structure (in particular, in the syntactic level and the semantic level) of the various natural language texts. Everyone knows that development of world-wide knowledge base is a huge undertaking. As well as world-wide knowledge base is mainly used in Internet-oriented searches, recommendations and question-answer tasks. The application scenario is relatively simple. However, the knowledge density in the knowledge base about the specialized domain is higher, and more complex reasoning and application scenarios can be performed in this kind of knowledge base. The application scenarios of knowledge bases about the specialized domains are more extensive.

The main goal of this article is to use ontological approach as a basis to develop a unified semantic model for natural language interface of knowledge-based intelligent systems. This kind of natural language interface can work in a multilingual field, as well as the knowledge base about the specialized domain are the main components of these intelligent systems, i.e., the natural language interface of intelligent systems that is independent of the particular natural language, but is dependent on the particular domain. The proposed mechanism allows, in principle, to potentially implement information transformation between texts of various natural languages and the knowledge base of intelligent systems about specialized domain, it's just that the complexity of construction of knowledge base about the specific natural language that provides linguistic knowledge of the specific natural language for natural language text processing will be different depending on features of a particular natural language, as well as specific additional problem solvers will be required for solving tasks in the specific natural language text processing.

III. Review of existing approaches

In the development of the natural language interfaces of knowledge-based intelligent systems, currently two main historically established directions [5] can be distinguished to solve the problem of converting natural language texts into fragments of the knowledge base of intelligent systems, and the problem of generating natural language texts from fragments of the knowledge base of intelligent systems.

The first direction involves developing the systems of logical reasoning based on rules that correspond to the grammar of a certain natural language, formulated by computational linguists and other linguists.

From the point of view of natural language processing, the problem of converting natural language texts into fragments of the knowledge base of intelligent systems is considered as the knowledge extraction. The knowledge extraction refers to obtaining factual information such as concepts, named entities, relations between them, as well as a certain type of events from natural language texts, with the extracted results formalized in the form of an internal knowledge representation language of the intelligent systems (such RDF, SC-code and others). In fact, from the point of view of construction of knowledge base of intelligent systems, the relations between concepts or named entities are considered as the special entities that are extracted from natural language texts.

Nowadays knowledge extraction from natural language texts can be divided into two directions according to the scope of the extracted domain [6]:

- knowledge extraction from closed domains;
- knowledge extraction from open domains.

The knowledge extraction from closed domains is focused on the processing of natural language texts that describe a specific subject domain, for example, football, weather, and so on. When solving the problem of knowledge extraction from closed domains, researchers believe that in a certain subject domain there are limited types of concepts, named entities and relations between them. Therefore knowledge extraction from closed domains often requires predefined types of concepts, named entities and relations between them. However, the goal of knowledge extraction from open domains is to extract various sets of concepts, named entities and relations between them from massive and heterogeneous natural language texts corpora without requiring a predetermined vocabulary to define the types of concepts, named entities and relations between them. In other words, for massive and heterogeneous corpora of natural language texts, the types of concepts, named entities and the relations between them can be infinite, unknown, or even constantly changing.

In the early stages of the development of technology for knowledge extraction from closed domains, researchers study the rules that exist in natural language texts and automatically implement these rules when solving the knowledge extraction. The knowledge extraction from natural language texts in a standardized format has been achieved through artificial inductive and generalized patterns and rules. Therefore, rule-based approaches [7] are closely related to specific natural language texts. As soon as minor changes to the format specification of natural language texts occur, the established rules may not apply. The most obvious advantages of knowledge extraction based on handwritten rules are that they do not require a large of training corpus to train the model and high accuracy. However the construction of rules manually is inefficient, and the constructed rules lack portability in other intelligent systems.

In order to solve the knowledge extraction from closed domains, ontology-based approaches use different models for knowledge extraction. Ontology construction is the basis of knowledge extraction, in the processing of ontology construction the different models are used . The ontology can more accurately describe concepts and relationships between them in the specific domains. In the ontology-based approaches knowledge extraction is often implemented by organically combining multiple ontologies. Alexander Schutz et al. organically combined ontologies such as DOLCE, SUMO, SpotEventOntology to describe related concepts in the field of football and the relationships between them, and then established the RelExt system [8], which can automatically identify correlated triples (a pair of concepts and a relationship between them). However in the knowledge extraction systems based on ontologies the the lack of general principles for the implementation of automatic text processing and knowledge extraction based on ontologies in a unified framework leads to the fact that it cannot ensure the compatibility of various components (different ontologies, problem solving models for natural language text processing and knowledge extraction).

With the development of the Internet, the volume of linked electronic natural language texts is rapidly increasing. In heterogeneous massive texts, the types of named entities and the relationships between them are not defined and limited. In order to meet the requirements of practical application in this situation, more and more researchers are beginning to study technologies for knowledge extraction from open domains. The knowledge extraction from open domains directly determines relative words or phrases in the natural language texts by analyzing natural language texts (particularly, natural language sentences) to realize the extraction of named entities and relationships between them without the need to predetermine types of named entities and relationships between them.

Until now some OpenIE systems (knowledge extraction from open domains) have been developed to extract structured knowledge (i.e. named entities and relationships between them) represented in the internal knowledge representation language from English sentences. As the first generation of the OpenIE systems, TextRunner [9], WOE [10] formulated an mode of knowledge extraction from open domains: first, heuristic or supervised distance learning models are used for automatic sentence labeling; then sequence labeling models (for example, sequence to sequence models) are used in order to learn how to automatically extract knowledge from the labeled sentences; finally, the sentences are entered into the systems, and the systems identify the named entities and the relationships between them in the sentences. With the development of knowledge extraction technologies, researchers have proposed the second-generation mode of knowledge extraction system . Fader et al. developed the ReVerb system [11], which uses a common syntax and lexical constraints for knowledge extraction. This system performs a comprehensive morphological and syntactic analysis of randomly selected English sentences, recognizes verb phrases to express as the relationships between named entities in sentences, and the noun phrases or other phrases in the sentences related to the verb phrases are expressed as named entities. In order to extract more than just verbs or verb phrases from sentences as relations, Mausam et al. proposed the OLLIE system [12] for extracting phrases other than verb phrases as relations with the help of dependency parsing trees of input sentences.

As seen, all the OpenIE systems presented above are focused on knowledge extraction from English sentences. For other natural languages that are very different from English, such as Chinese, the structure of the sentences and the grammatical features between the English and Chinese have the huge difference. It is not known whether these technologies and approaches work for English and also for other natural languages. In order to extract knowledge from those natural languages that differ from English, when the development of the knowledge extraction system it's necessary to take into account the characteristics of the specific natural languages. Tseng et al. proposed a classical architecture for knowledge extraction from open domains for Chinese language. Their CORE system [13] is the first attempt to acquire knowledge from open domains for Chinese texts. The system uses a number of Chinese language processing technologies, such as word segmentation, automatic morphological markup (POS tagging) adjusted and appropriate for Chinese texts, syntactic parsing, semantic analysis, and extraction rules for Chinese sentences.

According to the above, when solving the problem of knowledge extraction from open domains, it is necessary to solve a number of subtasks. However, in the existing knowledge extraction systems there is no unified framework that ensures the consistent use of various models to solve these subtasks. In addition, existing knowledge extraction systems cannot provide a single semantic framework that uses different knowledge bases of linguistics to solve the problem of knowledge extraction. In the unified knowledge base describes the syntactic and semantic knowledge, as well as the logical rules that are used to solve the subtasks in the process of knowledge extraction.

For natural language generation there is a classic pipeline architecture. The following three main modules are the basis of the natural language generation systems developed based on the pipeline architecture [14]:

- content planning: specifying what information from the input data will be included in the generated texts, and how it will be organized;
- micro-planning: deciding how the selected information will be implemented in the form of natural language sentences;
- implementation in natural language: producing grammatically correct natural language sentences.

At an early stage in the applied field, many natural language generation systems were successfully developed in the field of journalism and media research, for example, soccer reports [15], weather or financial reports [16], [17], [**?**] and others. These systems widely use templates and grammar rules constructed by researchers to generate natural language texts. When specific application fields are small and changes to the specific data structures are minimal, template-based approaches can achieve good results for natural language generation. However, the developed templates are highly dependent on the specific application fields and specific natural languages, i.e. the templates do not generalize well to different application fields and different natural languages. In addition, manually developing templates is a laborious task (templates construction requires a significant overhead of labor and time), and the re-usability of templates is low.

The NaturalOWL system [19] is a classical natural language generation system that reasonably applies templates and linguistic rules to generate natural language texts that describe fragments of OWL ontologies. NaturalOWL is focused on generating natural language texts (mainly the English texts) from fragments of OWL ontologies. Currently, the so-called semantic reasoners that perform logical inference on ontologies presented in the OWL format can be developed, but most of the developed reasoners are capable of performing only direct logical inference based on the relations described in the ontology. However in the NaturalOWL system, the construction of templates and rules for generating natural language texts does not use the OWL standard. In other words, the semantic reasoners based on description logic cannot be used to construct templates and linguistic rules. It can be said that this system does not provide a model, which allows representing various types of knowledge in a single knowledge base, including linguistic knowledge, inference on the ontologies of linguistics, even the templates developed by ontologies.

The second direction, widely used in modern intelligent systems, involves the use of various machine learning models based on mathematical statistics and information theory, which are aimed at modeling natural language texts.

At present, approaches based on mathematical models are mainly used in knowledge extraction from closed domains. In recent years, some deep learning models have shown significant advantages in natural language processing. These models have also been applied in the field of knowledge extraction, for example, pre-training models series BERT, GPT achieve better results [20], [**?**]. These approaches encode the input natural language texts into the input information (embedding information), and then decode the input information into predefined markup for each word of the input natural language texts. Based on the marked-up natural language texts, the corresponding named entities and the relationships between them can be extracted. Statistical-based approaches reduce the dependence of knowledge extraction systems on humans. However, using the machine learning models or the deep learning models always requires very large corpora that are manually annotated by the human, hence human intervention is required. Moreover the performance of these models highly depends on the quality of the annotating training samples.

For natural language generation, pre-training models series BERT, GPT also widely applied in the natural language generation systems [22]. For these statistics-based approaches obtaining or constructing the high quality datasets is one of the main key tasks of these approaches, as well as one of the key difficulties. In order to generate natural language texts from fragments of knowledge base, for training these models, training data is usually presented as a pair of inputs (fragments of knowledge base) and outputs (natural language texts).

In recent years, in the WebNLG project [23], the main task is to develop neural generators using deep learning models. For the development of neural generators, the project team provides training data, which consists of Data/Text pairs, where the data are the fragments of the knowledge base in the form of RDF extracted from DBpedia [24], and the texts are the specific natural language texts corresponding to the fragments of the knowledge base. Due to the difficulty of constructing high quality consistent training datasets, only English and Russian datasets are currently provided in the WebNLG project. For other natural languages, such as Chinese, there are no high quality consistent training datasets, although there are several well-known knowledge bases in Chinese in the field of Chinese language processing, for example, CN-DBpedia [25], zhishi.me [26] and others, which can provide the fragments of the knowledge base in the form of RDF [27], [28]. However at the same time, high-quality consistent datasets with Data/Text pairs for generating Chinese language texts are difficult to obtain or construct.

In order to reasonably use different problem solving models to solve the corresponding subtasks in the process of generating natural language texts from the fragments of the knowledge base, Amit Moryossef et al. systematically proposed a text generator developed with a modular architecture that separates the generation process into a symbolic text planning focusing on processing of the fragments of knowledge base [29] and a neural generator focusing on realisation of resulted natural language texts. As seen, the proposed hybrid approaches can improve the performance of text generators with the help of modular architecture. However, in the development of the text generators, it is obvious that there is no unified formal basis for integrating various problem solving models when solving the complex task of generating natural language texts from the fragments of the knowledge base.

IV. PROBLEMS, NEED TO BE SOLVED

During the development of natural language interfaces of knowledge-based intelligent system, Whether for converting natural language texts into fragments of the knowledge base of intelligent systems or generating natural language texts from fragments of the knowledge base of intelligent systems, existing approaches only partially solve the problems. As mentioned above the developed knowledge extraction systems for the open domain and the developed various generators has been successfully applied in various fields. Nevertheless, the following problems still need to be considered:

- In modern intelligent systems, the lack of unified basis in the process of developing the natural language interfaces and the integrating various developed components by different developers leads to the impossibility of parallelization of the components development;
- Due to the diversity of natural languages, when solving the knowledge extraction from open domains and natural language generation, there are their own additional problems for the characteristics of different natural languages, for example, according to the writing habits of Chinese texts, Chinese characters are written one after the other, without natural spaces between them. Moreover the syntactic structures and parts of speech used for English or Russian language processing are not fully applicable to Chinese language processing. Therefore in modern intelligent systems the existing component or modular approaches cannot guarantee the compatibility and flexibility of the developed components in natural language interfaces for multilingual field;
- For converting natural language texts into the fragments of knowledge base, existing ontology-based approaches are only applicable to the knowledge extraction from closed domains. In ontology-based approaches, the lack of a unified principle for the organic combination of various ontologies leads to a large number of incompatible components in the knowledge base. In the development of modern extract knowledge systems from open domains, the lack of a unified framework to integrate automatic text processing and the logical models for knowledge extraction leads to low compatibility of the developed components;
- For generating natural language texts from the fragments of knowledge base, developed generators on the basis of the modular architectures can use various problem solving models, but the lack of unified principles for using various problem solving models leads to duplication of similar solutions in different generators. In turn, the overhead costs for the development of generators increase;
- For knowledge extraction from open domains or natural language generation, it is necessary to use various levels of linguistic knowledge, including syntactic knowledge, semantic knowledge, logical rules for knowledge extraction, and also other rules to solve specific subtasks. In the natural language interfaces, generally the knowledge base of linguistics provides the necessary knowledge to solve the corresponding tasks. Among the existing various knowledge bases related to linguistics, the lack of unified principles of combining various types of knowledge and knowledge representation models in a single knowledge base, i.e. the possibility of representing syntactic knowledge, semantic knowledge, logical rules in a single knowledge base is difficult. Therefore, the compatibility of developed components of knowledge base by different developers

cannot be guaranteed. The low re-usability of the developed components leads to a high labor intensity in the development of the knowledge base.

In this paper in order to solve the above problems, it is proposed ontological approach to develop a unified semantic model that provides the unified basis to effectively combine the knowledge base of linguistics, including syntactic knowledge, semantic knowledge and other various types knowledge, and various problem solving models for developing the natural language interfaces, solving the problem of converting natural language texts into fragments of the knowledge base and the problem of generating natural language texts from fragments of the knowledge base. The principles underlying the ontological approach to develop a unified semantic model of natural language interfaces will be discussed below.

V. THE PROPOSED APPROACH

For the implementation of converting natural language texts into fragments of knowledge base and generating natural language texts from fragments of knowledge base in the natural language interfaces of knowledge-based intelligent systems, we propose to use ontological approach to represent various linguistic ontologies in a single knowledge base, including syntactic, semantic knowledge and logical rules for texts processing and to develop problem solvers having ability to integrate various approaches for solving related problems in the natural language interfaces within OSTIS Technology framework [30]. The OSTIS technology is aimed at developing a class of systems called knowledge-driven computer systems. Within OSTIS Technology framework this kind of knowledge-driven computer systems developed using the OSTIS Technology are called ostis-systems.

Within the OSTIS Technology, as an internal formal language for the semantic representation of knowledge in the memory of ostis-systems, the SC-code is used, which provides a unified version of information encoding and a formal basis for developing model of ostis-systems. This kind of knowledge representation model is a unified semantic network with a set-theoretic interpretation. The alphabet of SC-code consists of five main sc-elements, on the basis of which it is possible to build SC-code constructions of any complexity. Information is stored in the memory of ostis-systems (sc-memory) as SC-code constructions. Several universal variants of visualization of SC-code [30], such as SCg-code (graphic variant), SCn-code (nonlinear hypertext variant), SCs-code (linear string variant) will be shown below.

Within OSTIS Technology framework, each ostis-system consists of a complete model of such a system, described using a platform-independent SC-code (sc-model of a computer system), and an interpretation platform for such an sc-model, which can generally be implemented in software and hardware. From the point of view of components, the sc-model of the computer systems (ostis-systems) can conditionally be decomposed into the sc-model of the knowledge base [31], [32], the sc-model of the problem solver [33], [34], and the sc-model of the interface [35], [36], as well as the model of abstract semantic memory (sc-memory) [37], which stores the constructions of the SC-code (Figure 2). The principles of the development of knowledge bases and problem solvers are discussed in more detail in [31] and [33], respectively. Moreover the principles of the development of sc-model of the user interface, which is included in the sc-model of the interface, were desecribed in the article [35], [36].

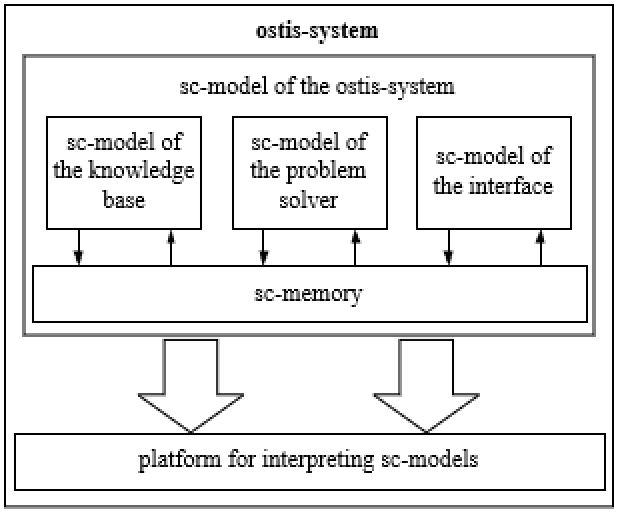

ostis-system
sc-model of the ostis-system
sc-model of the knowledge base
sc-model of the problem solver
sc-model of the interface
sc-memory
platform for interpreting sc-models

Figure 2: The architecture of the ostis-system

Within OSTIS Technology framework, according to the type of interaction between the human users and the computer systems, natural language interfaces of ostis-systems are defined as a subclass of user interfaces [38]. Therefore the development of natural language interfaces of ostis-system in fact is the development of sc-model of natural language interface, which is capable of solving related problems about natural language text processing in the natural language interfaces.

The OSTIS Technology is used as a formal basis to develop the sc-model of natural language interfaces of ostis-systems, the following advantages and characteristics are mainly considered:

- The SC-code is aimed at the sense knowledge representation, which provides a unified semantic model and means for generalizing various types of knowledge and knowledge representation models (for example, the semantic network, the frame, the production rule representation, the logical model) in a single knowledge base;
- The ontological and modular approaches are used to develop the components of the intelligent systems,

which allow integrate various types of linguistic knowledge and various problem solving models for natural language processing in a semantically unified manner. Thence these approaches ensure semantic compatibility and re-usability of the developed components;

- The structure of the knowledge base is developed as a hierarchical system of selected subject domains and their corresponding ontologies. This hierarchical structure allows describing various levels of linguistic knowledge in detail, including syntactic level, semantic level and others. Therefore according to the developed sc-model of knowledge base of the linguistics, for the specific natural language, the structure of the knowledge base is relatively easy to adjust to construct the knowledge base on specific natural language processing (in general, the specification of the syntax and semantics of the specific natural language is only need);
- The multi-agent approach is used to develop the sc-model of problem solver in the intelligent system. This approach allows to consider the interpretation of problem solvers as a hierarchical system of agents (sc-agents) working in a unified semantic memory (sc-memory). The sc-agents are a certain kind of subjects that perform actions in sc-memory. In order to solve problems at different levels on natural language processing this approach provides a unified basis for integrating various problem solving models (for example, logical models and artificial neural network models);
- The multi-agent model for development of problem solver provides developed component oriented on natural language processing modifiability, i.e. the ability to extend functionality of each components or to improve their performance.

Within OSTIS Technology framework, natural language interfaces of the ostis-systems, as a subclass of user interfaces, can be considered as the specialized ostis-systems focusing on solving the problems related to natural language interfaces (concretely, conversion of natural language texts into fragments of knowledge base and generation of natural language texts from fragments of knowledge base). According to the principles to develop the ostis-systems, the development of natural language interface generally includes the development of a knowledge base (mainly the knowledge base of linguistic) and a problem solver of the natural language interface [40]. The knowledge base of natural language interface mainly requires the presence of an sc-model of knowledge base of linguistic, actions for analysis of natural language texts, as well as actions for generation of natural language texts, as shown in Figure 3.

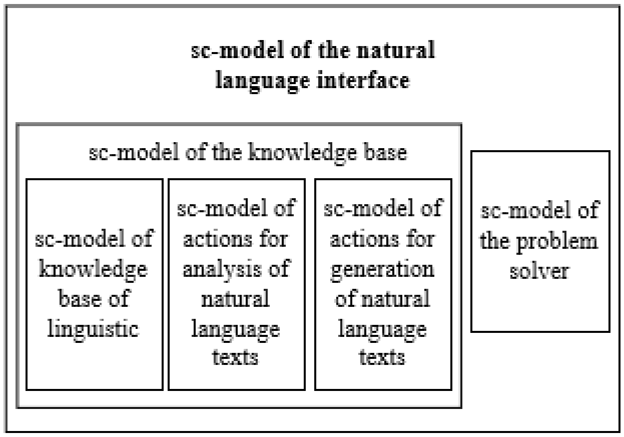

sc-model of the natural language interface
sc-model of the knowledge base
sc-model of knowledge base of linguistic
sc-model of actions for analysis of natural language texts
sc-model of actions for generation of natural language texts
sc-model of the problem solver

Figure 3: The general structure of the sc-model of the natural language interface

As can be seen from the Figure 3, due to the use of component approach, the development of the entire natural language interface comes down to development and improvement of separate specified components. Within OSTIS Technology framework, it is assumed that each problem solving model corresponds to some sc-agent. The knowledge base of linguistic just provides static linguistic knowledge to problem solving model for problem solution. The natural language interface should have the dynamic ability to perform some actions aimed at solving particular related problems in the natural language interfaces.

VI. SC-MODEL OF KNOWLEDGE BASE

According to the basic principles to construct the knowledge base using OSTIS Technology, the development of knowledge base is to consider structure of knowledge base as a hierarchical system of subject domains and their corresponding ontologies. From the point of view of OSTIS Technology, an ontology is interpreted as a specification of the system of concepts of the corresponding subject domain. The knowledge base of linguistic contains a formal description of necessary linguistic knowledge for analysis of natural language texts and generation of natural language texts [5]. The corresponding ontology provides a formal description of the concepts used for the representation of such linguistic knowledge in the different levels from basic word to syntactic and semantic structure of natural language texts. Linguistic theories offered by linguists provide a theoretical basis for building knowledge base of linguistics. Therefore the main hierarchy of knowledge base of linguistics used in the natural language interfaces to solve the particular problems related directly to conversion of natural language texts into fragment of knowledge base of ostis-systems and generation of natural language texts from fragment of knowledge base of ostis-systems is shown below in the SCn-code.

knowledge base of Linguistics
⇐ *section decomposition**:
{● *Section. Subject domain of lexical analysis*
● *Section. Subject domain of syntactic analysis*
● *Section. Subject domain of semantic analysis*
}

A. *subject domain of lexical analysis*

The subject domain of lexical analysis includes a series of ontology for lexical analysis, which describes characteristics of words and syntactic function of words, parts of speech, etc. A fragment of the ontology of the subject domain is presented below:

natural language text
⊂ *file*

word
⊂ *file*

nominative unit
⊂ *file*

The *word* is the smallest of the natural language units, carrying semantics, which serves to name objects, their qualities, characteristics and interactions, as well as for service purposes. A *nominative unit* is a stable string of word combinations.

natural language text
⇒ *decomposition**:
{● *part of speech*
    ⇒ *decomposition**:
    {● *content word*
    ● *function word*
    ● *modal word*
    }
● *sign of syntax alphabet*
}

The *part of speech* is a category of natural language words determined by morphological, syntactic and semantic features. The *signs of the syntax alphabet* are auxiliary syntactic means (at the macro level - prepositions, postpositions, conjunctions, particles, etc., as part of service words, at the micro level - inflections, prefixes, postfixes, infixes, punctuation, etc.) [**?**].

It is important to note that these ontologies in the respective subject domain are generally recognized in most natural languages. In other words, due to the diversity of natural language, the linguistic theory of a certain natural language is not suitable for other natural languages. Thus, when developing the knowledge base of linguistics for a particular natural language, the specification of subject domains must take into account the specific characteristics of a particular natural language.

In Russian, English or other European languages, word structure is studied within the framework of morphology. It's necessary to consider the relation *morphological paradigm**:

lexeme
⊂ *file*
⇒ *explanation**:
[The lexeme is a unit of lexical meaning that underlies a set of words that are related through inflection.]

morphological paradigm*
∈ *quasi-binary relation*
⇒ *first domain**:
lexeme
⇒ *second domain**:
word form

As you can see, the terminology lexeme is studied only to describe the features of the words in European languages that have the feature of morphological paradigm.

B. *subject domain of syntactic analysis*

The subject domain of syntactic analysis describes the characteristics of natural language syntax, the functional characteristics of syntactic components (such as, word, phrase, sentence, etc.). Among them, in the filed of natural language processing sentences have always been considered as the smallest research units. The analysis of sentence is an important intermediate stage, connecting the analysis of the entire text and the analysis of individual words. A fragment of the ontology of the subject domain is presented below in the SCn-code:

sentence
⇒ *decomposition**:
{● *simple sentence*
    ⇒ *decomposition**:
    {● *sentence with subject and predicate*
    ● *sentence without subject and predicate*
    ● *special sentence*
    }
● *complex sentence*
}

A *sentence* is considered simple if it contains one predicative unit, if more - complex.

part of the sentence'
⇒ *decomposition**:
{● *principal part of the sentence'*
    ⇒ *decomposition**:
    {● *grammatical subject'*
    ● *grammatical predicate'*
    ● *grammatical direct object'*

}
- *subordinate part of the sentence'*
  ⇒ *decomposition**:
  {• *grammatical indirect object'*
  • *grammatical attribute'*
  • *grammatical circumstance'*
  }

}

A *part of the sentence* is a role relation that indicates a certain of syntactic relation between the sentence and its components (words, phrase etc.) [39].

relation of dependency*
∈ *non-role relation*
⇒ *decomposition**:
{• *subject**
• *object**
• *complement of verb**
• *coordinate**
• *attribute**
• *adverbial**
}

On the basis of theory of dependency grammar, a *relation of dependency** is a relation that indicates directed link within decomposed two components of sentence, the components are separately called the *head*, which plays the role *grammatical predicate'* in the sentence, and the *depedent*. Each *depedent* has a syntactic function depending on the *head* in sentence. Hence in the syntactic analysis of sentences, the *predicate* of sentence is generally considered as the structural center of the syntactic structure of sentence. All other syntactic components (words or phrases) are directly or indirectly connected with the *predicate* through directed links (i.e., the direction with the *predicate* to other syntactic components).

In the logical ontology of subject domain of the sentence, it is possible to describe a series of logical definitions and logical statements for the sentences. In the form of logical statements, templates or rules about the sentence analysis can be constructed, as well as can be used by problem solvers to solve specific problems in the natural language interfaces.

In the Figure 4 the logical statement in SCg indicates that a sentence with the subject and the predicate can consist of the noun phrase and the verb (or verb phrase), which serve as subject and predicate in the sentence, respectively. As shown Figure 4, all used concepts and relations, such as “sentence with subject and the predicate”, "grammatical subject'”, “grammatical predicate'”, and others, are contained in the subject domain of the sentence. Various type of knowledge built in the subject domain can be used to process natural language texts.

This fragment of the knowledge base describes that the *sentence with the subject and the predicate* can be considered as a sequence of the noun phrase and the verb phrase (or verb) with the corresponding attributes. Moreover, in the process of text generation, a sentence with the subject and the predicate can be generated according to this given structure.

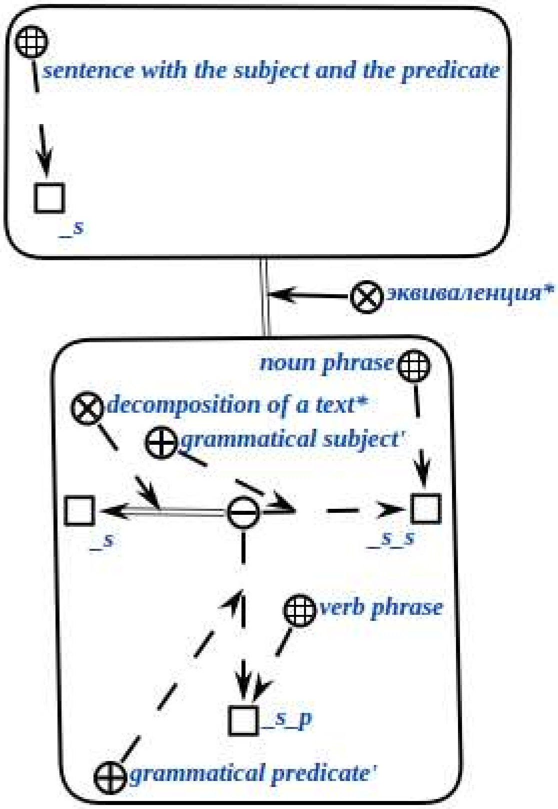


Figure 4: The logical statement about sentence with the subject and the predicate

C. subject domain of semantic analysis

The subject domain of semantic analysis describes the semantic characteristics of the words and the semantic structure of natural language sentences, the functional characteristics of semantic of components (words, phrase and sentence, etc.), the semantic role, the rules for semantic analysis, and so on. The general structure the subject domain is presented below in the SCn-code:

The subject domain of semantic analysis at the level of lexeme, the subject domain of semantic analysis at the level of sentence, and the subject domain of semantic analysis at the level of paragraph describe different linguistic knowledge for semantic analysis at different aspects, respectively.

The subject domain of semantic analysis at the level of lexeme describes the semantic classifications of common basic concepts expressed by the lexeme or the binding of the lexeme to individual entities. The subject domain was constructed based on various types of knowledge bases of semantic about natural language used for semantic analysis at the level of lexeme, for example, WordNet, ConcetpNet, The Semantic Knowledge base of Modern Chinese [41], TAPAZ-2.

semantic of lexeme
⇒ *decomposition**:
{• *semantic of verb*
• *semantic of noun*
• *semantic of adjective*
• *semantic of adverb*
}

participant of the exposure*
:= [participant of the action*]
∈ *non-role relation*
⇒ *first domain**:
individ
⇒ *second domain**:
action
⇒ *decomposition**:
{• *subject**
⇒ *decomposition**:
{• *initiator**
• *inspirer**
• *spreader**
• *creator**
}
}

An *individ* is a kind of the pattern as a separate entity, which is an instance of the specific concepts.

The *participant of the action** is a non-role relation that connects the action with the individ participating in it, to a certain extent, it can be considered as the semantic role of the action, which generally is expressed by the verb or the verb phrase in the sentence.

In turn, similarly the subject domain of semantic analysis at the level of sentence describes the semantic structure of sentences, the semantic relations between the components (words, phrase and so on) in a sentence, and the semantic relations between the sentences in the text. Some open sources, for example, TAPAZ-2, PropBank and others, served as the basis for building this subject domain.

D. subject domain of actions for natural language interface

Previously we consider the subject domains and the corresponding ontology in the knowledge base of linguistic, within which the various type of linguistic knowledge is described. However, the formalization of the linguistic knowledge is not enough for the natural language interfaces to solve the particular problems, the natural language interfaces should perform some *actions* to implement the conversion of natural language texts into fragments of knowledge base and the generation of natural language texts from fragments of knowledge base.

Let us consider the hierarchy of classes of actions for natural language interface in the SCn-code:

action for natural language interface
:= [action for processing the natural language texts]
⊂ *action*
⇒ *decomposition**:
{• *action for converting natural language texts into fragments of knowledge base*
• *action for generating natural language texts from fragments of knowledge base*
}

action for converting natural language texts into fragments of knowledge base
⊂ *action for processing the natural language texts*
⇒ *decomposition**:
{• *action for decomposing texts into separate units*
• *action for marking up separate units*
• *action for syntactic analysis*
• *action for semantic analysis*
• *action for determining the knowledge structure*
• *action for linking knowledge structures in the knowledge base*
• *action for determining contradictions*
• *action for resolving contradictions*
}

action for generating natural language texts from fragments of knowledge base
⊂ *action for processing the natural language texts*
⇒ *decomposition**:
{• *action for determining the converted knowledge structures*
• *action for converting knowledge structures into standard basic sc-constructions*
• *action for determining candidate basic sc-constructions to be generated to texts*
• *action for converting sc-constructions into message triples (subject-action-object)*
• *action for generating resulted texts from the message triples*
}

From the perspective of text generation for fragments of knowledge base of ostis-system, it is worth noting that message triple is a kind of ordered triple represented in the form of <semantic subject, action, semantic object>, where semantic subject is always the identifier of the set of sc-node denoting concepts that are not relations or the element of this set of sc-node; relation is the identifier of the set of sc-node denoting role relation or no-role relation, which points the linking (bundle) that connects the semantic subject and object; semantic object is the identifier of the set of sc-node denoting concepts that are not relations or the element of this set of sc-node.

The message triple is considered as intermediate conversion between the sc-structure and resulted text, mainly because this kind of triple is easier to express as natural language text (mostly sentence) using either rule-template models or modern statistical models. Moreover the message triples are easily to be converted to RDF triples, in turn, can be generated to natural language text using modern deep learning models. Within the framework of OSTIS Technology, the relations defined in the IMS-system are consider as the domain-independent relations, in addition, there are many relations that are dependent on the specific ostis-systems in various subject domain. According to the category of relations, it is divided into different set of message triples corresponding to various types of relations in different sc-constructions.

Since as a result of the actions for converting natural language texts into fragments of knowledge base and the actions for generating natural language texts from fragments of knowledge base, with the help of the natural language interface, the mutual conversion between the natural language texts and the fragments of knowledge base of the ostis-systems is realized, in turn the information interaction between the human users and the ostis-systems is completed.

VII. SC-MODEL OF PROBLEM SOLVER

Within the OSTIS Technology, in order to realize the natural language interfaces of the ostis-systems it need to have ability to perform the above-mentioned actions. Within the OSTIS Technology framework, an action is defined as a purposeful process carried out by one or more subjects. In this case, the sc-agent is considered as a certain subject capable of performing actions in the sc-memory to solve problems. The multi-agent approach is used as a basis for development of the problem solvers. The interaction of agents will be performed exclusively in the semantic memory (sc-memory), which stores the SC-code constructions. Such approach provides the flexibility and modularity of developed sc-agents, as well as provides the ability to integrate various problem solving models corresponding to these developed sc-agents. In the term of implementation, the programs of each sc-agent can implement logical reasoning based on a hierarchy of statements comprised in the logical ontology, as well as the data-driven algorithms (machine learning, deep learning and so on) in various programming language.

Let us consider the general structure for problem solver of natural language interface in the SCn-code:

Problem solver for natural language interface
⇐ *decomposition**:
{
- *Problem solver for converting natural language texts into fragments of knowledge base*
- *Problem solver for generating natural language texts from fragments of knowledge base*

}

From the point of view of OSTIS Technology, the sc-agents (the copies of the same sc-agent or functionally equivalent sc-agents) may work in different ostis-system, and can be considered physically different sc-agents. Rather than considering properties and typology of sc-agents, it's more rational to focus on classes of functionally equivalent sc-agents, which are called abstract sc-agents. The abstract sc-agents is a certain class of functionally equivalent sc-agents, various items of which can be implemented in different ways to specific problems [33].

Problem solver for converting natural language texts into fragments of knowledge base
⇐ *decomposition of an abstract sc-agent**:
{
- *Abstract sc-agent of lexical analysis*
  ⇐ *decomposition of an abstract sc-agent**:
  {
  - *abstract sc-agent of decomposing texts into separate units*
  - *abstract sc-agent of marking up separate units*

  }
- *Abstract sc-agent of syntactic analysis*
- *Abstract sc-agent of semantic analysis*
- *Abstract sc-agent of extracting knowledge structures into the knowledge base*
  ⇐ *decomposition of an abstract sc-agent**:
  {
  - *abstract sc-agent of determining knowledge structure*
  - *abstract sc-agent of linking knowledge structure into knowledge base abstract sc-agent of determining contradictions*
  - *abstract sc-agent of resolving contradictions*

  }
- *Abstract sc-agent of logical inference*

}

Problem solver for converting natural language texts into fragments of knowledge base is a group of sc-agents that implement the mechanisms of extracting fragments of knowledge base from natural language texts. In principle, this problem solver can potentially extract structured knowledge (generally, the named entities, relations, concepts and so on) from various types of natural language texts into the knowledge base of the ostis-system about a specific subject domain, but the construction of the corresponding knowledge base on the specific natural language processing, as well as within which the rules for processing the corresponding natural language texts and the rules for knowledge extraction will become more complex, and also overhead costs will increase.

Abstract sc-agent of lexical analysis – the agents that implement the mechanisms of decomposition of input natural language texts into separate lexical units, and each separate unit belongs to a specific markup according to a certain of markup standard for the specific natural language. Components (words, phrases, sentences, and others) of natural language input texts can be defined according to a certain standard of a particular natural language.

Abstract sc-agent of syntactic analysis – the agents that implement the mechanisms of constructing the syntactic structure of input natural language texts.

Abstract sc-agent of semantic analysis – the agents that implement the mechanisms of constructing the semantic structure of input natural language texts.

Abstract sc-agent of extracting knowledge structures into the knowledge base – the agents that implement the mechanisms that determine knowledge structures in natural language input texts (i.e., the definition of named entities and relationships between them), and construct the knowledge structures in the form of SC-code into knowledge base of the ostis-system about specific subject domain.

In the process of constructing knowledge structures into the knowledge base, when comparing the knowledge structures extracted as a result of the analysis of natural language input texts with the knowledge stored in the knowledge base of a particular ostis-system, the ***abstract sc-agent for determining contradictions*** implement mechanisms for determining contradictions, for example , for a certain named entity in the input natural language texts, there may be several corresponding instance of concept in the knowledge base.

Abstract sc-agent of resolving contradictions - agents that implement mechanisms that resolves the contradictions in case of detection of contradictions.

Abstract sc-agent of logical inference - agents that implement mechanisms that use logical rules written by means of SC-code for syntactic, semantic analysis and also extracting knowledge structures for natural language texts. Thus, this agent can semantically interact with the agents of syntactic and semantic analysis and the agent of extracting knowledge structures.

In order that the natural language interface can generate fluent natural language texts from fragments of knowledge base of ostis-system about specific subject domain, the problem solver for generating natural language texts from fragments of knowledge base is developed based on the classical pipeline for solving the problem of natural language generation. Although the development of this abstract agent is based on the classical pipeline, the development of composed sc-agents in the problem solver can be flexible by using a multi-agent approach. For the problem of generating texts for a particular natural language, the compositions of the problem solver can be easily modified accordingly.

Let us consider the general structure for problem solver for generating natural language texts from the fragments of knowledge base represented in SCn-language:

Problem solver for converting natural language texts into fragments of knowledge base
⇐ *decomposition of an abstract sc-agent**:
{• *Abstract sc-agent for content selection*
  ⇐ *decomposition of an abstract sc-agent**:
  {• *Abstract sc-agent determining sc-structure*
   • *Abstract sc-agent dividing determined sc-structure into basic sc-structure*
   • *Abstract sc-agent determining the candidate sc-structures*
   • *Abstract sc-agent transferring candidate sc-structures into message triples*
  }
 • *Abstract sc-agent text planning*
  ⇐ *decomposition of an abstract sc-agent**:
  {• *Abstract sc-agent ordering message triples*
   • *Abstract sc-agent ordering entities of a message triple*
  }
 • *Abstract sc-agent for micro-planning*
 • *Abstract sc-agent for surface realization*
}

Abstract sc-agent for determining sc-structure - the groups of agents that provide the specific fragments of knowledge base (i.e. content) in the form of sc-structures, from which the interface will generate the natural language texts.

Abstract sc-agent dividing determined sc-structures into basic sc-structures – the agents that implement the mechanisms of decomposition of determined sc-structures into basic structures, which can be transferred into message triples.

Sometime not all transferred message triples are useful to the end user. According to specific requirements the interface will finally select among the basic sc-structures the ones to generate natural language texts. ***Abstract sc-agent determining the candidate sc-structures*** implements the mechanism of determining the appropriate candidate sc-structures according to requirements.

Abstract sc-agent transferring candidate sc-structures

into message triples – the agents that implement the mechanism of converting candidate sc-structures into message triples.

The order of transferred message triples, as well as the order of elements in each message triple will completely influence on the resulted generated natural language texts. ***Abstract sc-agent for text planning*** implements the function of the ordering of elements in each message triple and the ordering of message triples themselves.

Abstract sc-agent for miro-planning - the agents that implement the mechanism of transferring message triples to abstract sentence specifications that are varied according to selected methods in the natural language generation systems, for example, the simple text templates with slots, syntactic structures and so on.

Abstract sc-agent for surface realization - the group sc-agents that implement the mechanisms of concatenating the sc-links to generate the resulted texts, i.e. filling the sc-links corresponding to elements of message triples with the appropriate forms (generally in European languages there is morphology for lexical units) of lexical units according to rules, and concatenating them.

It is worth noting that developed abstract sc-agents in the above listed problem solvers are general according to the general process of natural language processing, therefore they are flexible and changeable. For development of natural language interface, which can process the specific natural language, it is possible to adjust and make extensions of the already developed abstract sc-agents. Moreover, in some cases for processing of some specific natural languages it is even necessary to add a new agent or remove (deactivation) of one or more existing agents. With the help of multi-agent approach based on OSTIS Technology to develop the sc-agents, the one of advantages is that the development of each agent is independent of each other, i.e. the adjustment of agents, addition of a new agent, even removal of one or more existing agents usually does not lead to changes in other agents.

VIII. Implementation of Chinese Language interface

With the help of the previously proposed sc-models to development of the natural language interfaces of the ostis-systems, the specific natural language interface of intelligent help systems for various subject domains can be implemented. However, due to the laboriousness and complexity of developing an intelligent help system for a specific subject domain, within the framework of this section, the greatest attention will be paid to the implementation of the prototype of the Chinese language interface of the intelligent help system for discrete mathematics.

According to the above-mentioned sc-model of the natural language interfaces of the ostis-systems, on the basis of the general structure of knowledge base of linguistic and problem solver of natural language interface, the development of Chinese language interface requires the development of knowledge base on Chinese language processing, within which constructed various types of linguistic knowledge about Chinese language processing, as well as the development of problem solver for Chinese language processing, consisting of a group of sc-agents, which has possibility to integrate various problem solving models.

In order to implement the Chinese language interface of intelligent help system for discrete mathematics, it's necessary to implement the conversion of Chinese language texts into fragments of knowledge base of intelligent help system for discrete mathematics and generation of Chinese language texts from fragments of knowledge base of intelligent help system for discrete mathematics. The following examples show the processing stage to implement the two main functions in the Chinese language interface.

A. knowledge extraction from Chinese language texts

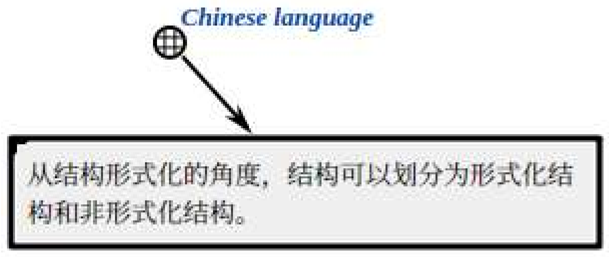
The conversion of natural language texts into fragments of knowledge base is considered as the problem of knowledge extraction. In this case the conversion of Chinese language texts into fragments of knowledge base of intelligent help system for discrete mathematics is mainly to extract named entities and relations between them from Chinese language texts, and the extracted results are stored in the form of SC-code in the intelligent help system for discrete mathematics.

In this example several requirements are defined for processed Chinese language texts:

- The input of Chinese language interface is a standard written Chinese declarative sentence;
- The input Chinese declarative sentence has completed sense and fact information;
- The input doesn't need a predefined vocabulary that includes predefined types of named entities and relations between them;
- The output of Chinese language interface is the sc-structure formally represented in the knowledge base.

From the point of view of OSTIS technology, any natural language text is a file (generally represented as sc-node with content or so-called sc-link). Our example will show the knowledge extraction process on analyzing a Chinese sentence, which is represented in such a node in Fig 5. The content of such a node demonstrates a Chinese sentence that describes: «从结构形式化构角度 (in terms of formalization of the structure) ，(comma) 结构(structure) 可以 (can be) 划分 (divided) 为 (into) 形式化构结构 (formal structure) 和 (and) 非形式化构结构 (informal structure) 。(full stop)».

Step 1 The Chinese sentence is decomposed into separate segmentation units, as well as mark these seg-



从结构形式化的角度，结构可以划分为形式化结构和非形式化结构。

Figure 5: The representation of natural language text in the ostis-system

mentation units with standard part-of-speech categories in Chinese language by agents for lexical analysis (Figure 6).

According to the features of Chinese texts, the Chinese texts consist of a stream of hieroglyphs without natural blanks on the written form. Moreover in Chinese language there are no clear indicators of the categories of number, case and gender, such as in Russian and other European languages, the function of a word in Chinese language becomes clear not on the basis of morphology of a word, but due to its connection with other words. Therefore in the process of analyzing the Chinese texts it is first necessary to perform a lexical analysis, decomposing the stream of hieroglyphs into separate meaningful segmentation units.

Because of the features of Chinese texts, lexical analysis for Chinese texts is more complicated. From point of view of linguists, the definition of Chinese words has always been an important theoretical problem. In order to realize the Chinese language processing by computer, in 1992, the project "The Principle of Segmentation in the Processing of Modern Chinese Information" was released [42], in which the term "segmentation unit" was introduced and a clear criterion of segmentation was established. Moreover in the field of Chinese language processing, the "Modern Chinese word segmentation standard used for information processing" was proposed. In this standard, a word in Chinese language is represented as a "segmentation unit". Its precise definition is "a basic unit for Chinese language processing with certain semantic or grammatical functions. In this case speaking of the Chinese word in this section, we will mean the segmentation unit of the Chinese language.

These decomposed separate units in the lexical analysis meet the principles to be "segmentation units" of Chinese language. Meanwhile the information describing (marking up) these segmentation units satisfying the standards is considered as the linguistic knowledge (e.g. part-of-speech categories) in the lexical analysis of the Chinese texts and is constructed as a part of knowledge base on the Chinese language processing.

Step 2 The agents for syntactic and semantic analysis performs the transition from the lexically marked structure to its syntactic structure or semantic structure, then to the semantically equivalent fragment of knowledge base based on the rules described in the corresponding subject domain.

The main task in this stage is to analyse the relations (or dependency relations) between the input Chinese sentence and separate segmentation units and between these separate segmentation units of that input Chinese sentence to reveal syntactic semantic structure of input sentence. The result of syntactic and semantic analysis is usually a constructed syntax semantic tree or graph, such as phrase structure, dependency structure and others. Based on the constructed syntactic and semantic structure, it is possible to develop a collective of logical reasoning for some applications in the field of Chinese language processing. Therefore the syntactic semantic analysis allows to establish a basis (i.e., a set of logical rules) for extracting structural knowledge from the Chinese sentences. However, the constructed syntactic semantic structure is not conveniently processed in sc-memory. After obtaining the result of syntactic semantic analysis, it is necessary to convert the constructed syntactic semantic structure into a semantically equivalent fragment of knowledge base in the form of SC-code.

From the perspective of the knowledge base, the relations between segmentation units and the Chinese sentence, or the relations between segmentation units of the Chinese sentence, are generally regarded as certain types of linguistic knowledge included in the knowledge base on Chinese language processing.

In the Figure 7 shows a semantically equivalent fragment to the given input Chinese sentences with the help of syntactic semantic structure of this sentence based on the constructed ontologies in knowledge base on Chinese language processing.

The advantage of implementation of transition from input Chinese sentences to semantically equivalent fragments allows to directly use logical reasoning or other knowledge processing operations to analyze the input Chinese sentences based on the dynamic graphical model considered within the OSTIS Technology.

Step 3 knowledge structure is extracted from the semantically equivalent fragment corresponding to the syntactic semantic structure of input Chinese sentence, and fusion of knowledge structure into the knowledge base of intelligent help system (i.e. the construction of fragment of knowledge base) is performed based on the logical rules described in the corresponding subject domain.

When converting Chinese texts into fragments of knowledge base of intelligent help system for discrete mathematics, from the point of view of knowledge extraction from open domain, the main task is to extract named entities and relations between them from the Chinese sentences without requiring predefined types of

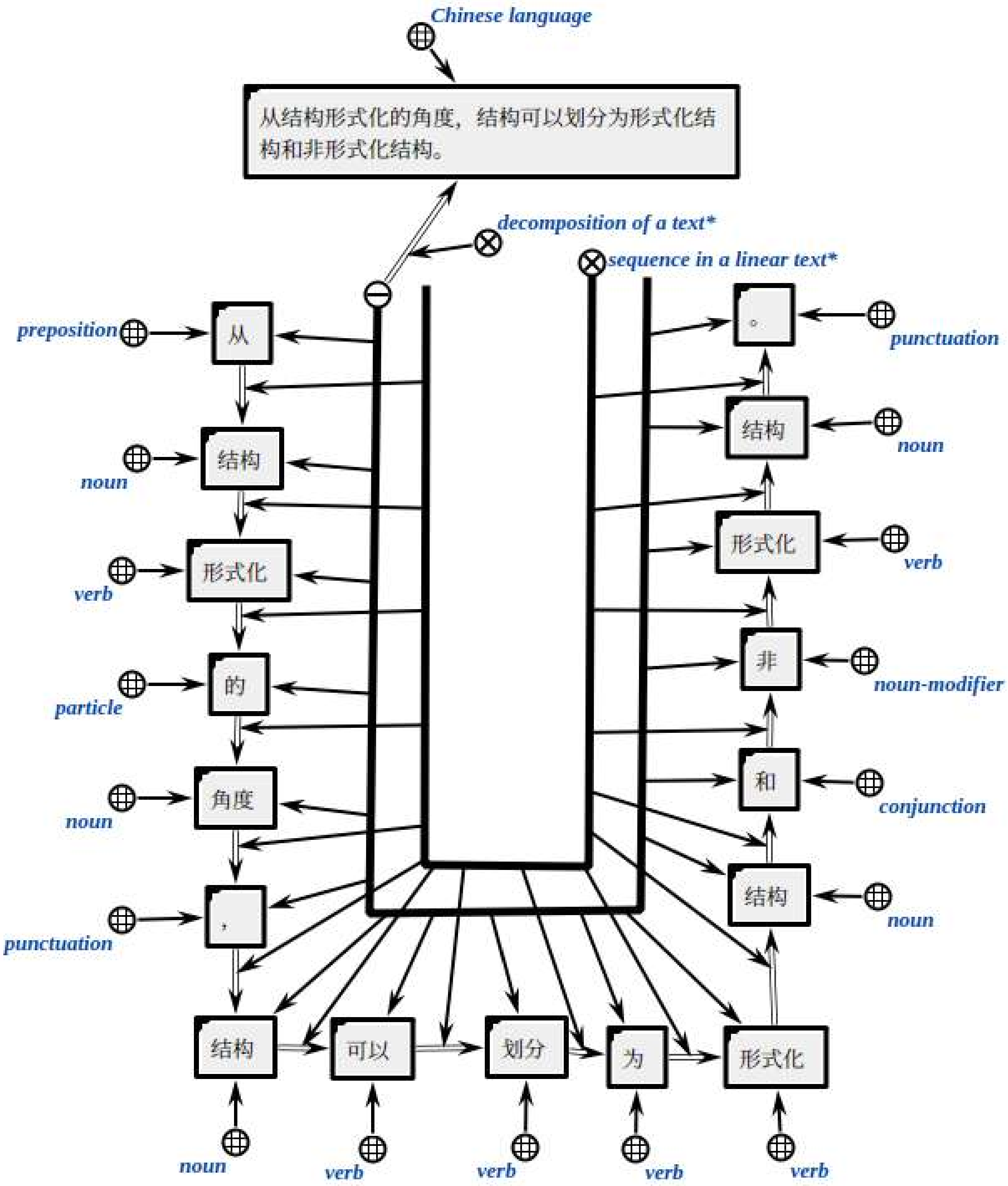


Figure 6: The result of lexical analysis for the input Chinese text

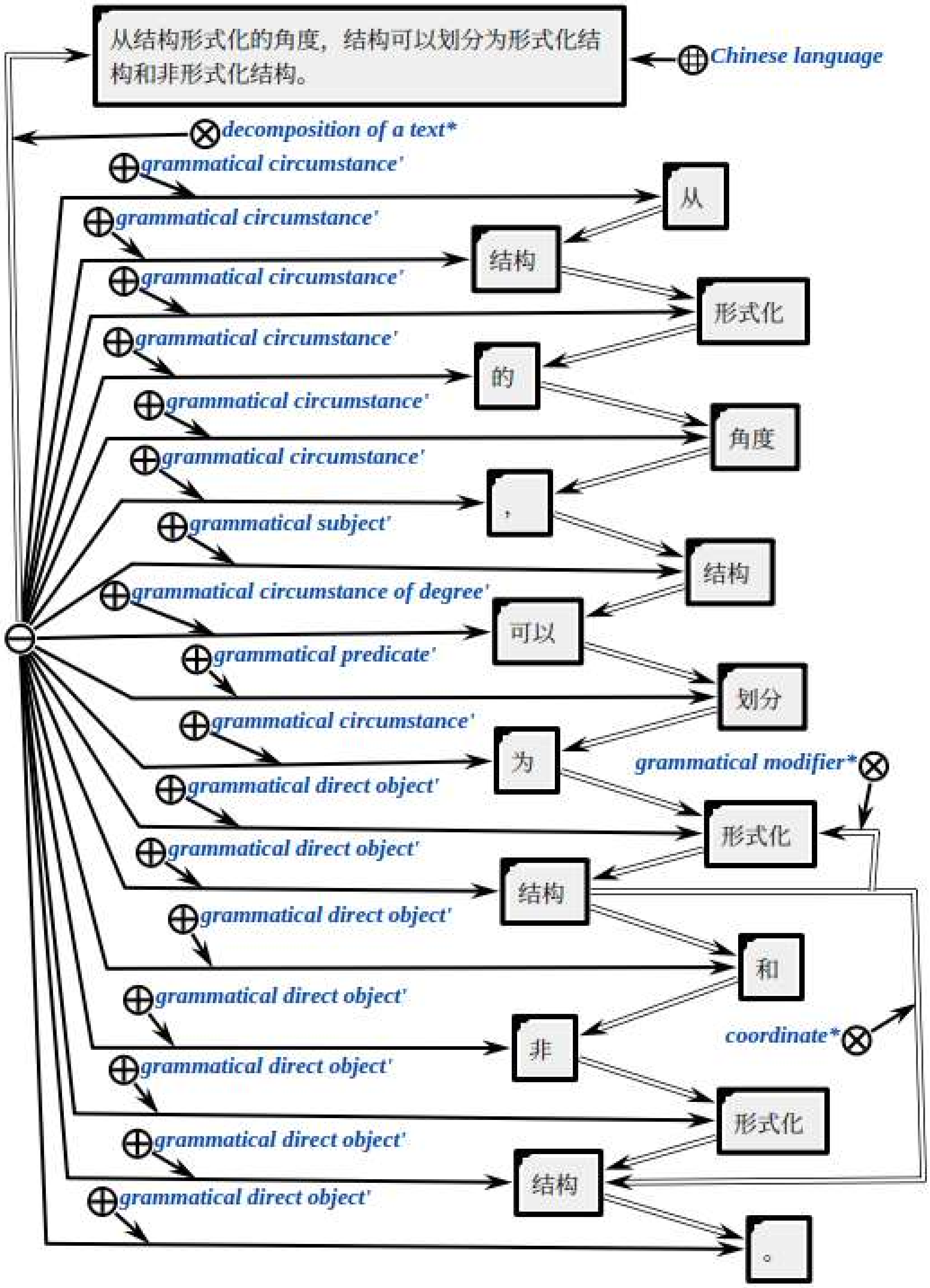

从结构形式化的角度，结构可以划分为形式化结构和非形式化结构。
Chinese language
decomposition of a text*
grammatical circumstance'
从
grammatical circumstance'
结构
grammatical circumstance'
形式化
grammatical circumstance'
的
grammatical circumstance'
角度
grammatical circumstance'
，
grammatical subject'
结构
grammatical circumstance of degree'
可以
grammatical predicate'
划分
grammatical circumstance'
为
grammatical modifier*
grammatical direct object'
形式化
grammatical direct object'
结构
grammatical direct object'
和
grammatical direct object'
非
coordinate*
grammatical direct object'
形式化
grammatical direct object'
结构
grammatical direct object'
。

Figure 7: The semantically equivalent fragment of the input Chinese text

named entities and relations, with the extracted results stored in the form of SC code. In fact, within OSTIS Technology, a relation is treated as a special entity that specifies a certain relation between pairs of independent named entities, which generally is represented respectively as the sc-node denoting the non-role relation or the sc-node denoting the role relation in the form of SCg-code.

In general, the pairs of independent named entities should appear in the analyzed syntactic semantic structure as noun phrases that belong to the nominative units. Afterwards the connections between these noun phrases and the input Chinese sentence, as well as the path linking the two noun phrases through other segmentation units, will reflect the corresponding relations between the pairs of named entities. More precisely, the noun phrases appearing in the analyzed syntactic semantic structure of input Chinese sentence are represented as identifiers of sc-nodes denoting the name of some named entities or some concepts stored in the knowledge base of intelligent help system for discrete mathematics.

In the Figure 8 shown the determination of named entities, the syntactic semantic relations between them and between named entities and the input sentence, that is, the determination of named entities , the syntactic semantic relations between them in the input Chinese sentence, as well as the syntactic semantic relations between named entities and the input Chinese sentence.

In the Figure 9 shown the resulted constructed fragment of the knowledge base from the input Chinese sentence in the intelligent help system for discrete mathematics without contradiction detection.

As can be seen from the Figure 9, in this case, a fragment of the knowledge base can be directly constructed without linking the named entities mentioned in the input source Chinese sentence with the corresponding exiting defined entities in the knowledge base of intelligent help system for discrete mathematics. In some cases, for a named entity in the knowledge base, there are different names in the natural language texts to describe this entity. In this case, it is necessary to perform contradiction elimination to relate different named entities (precisely identifiers of named entities) in natural language texts to the same named entities in the knowledge base of ostis-systems.

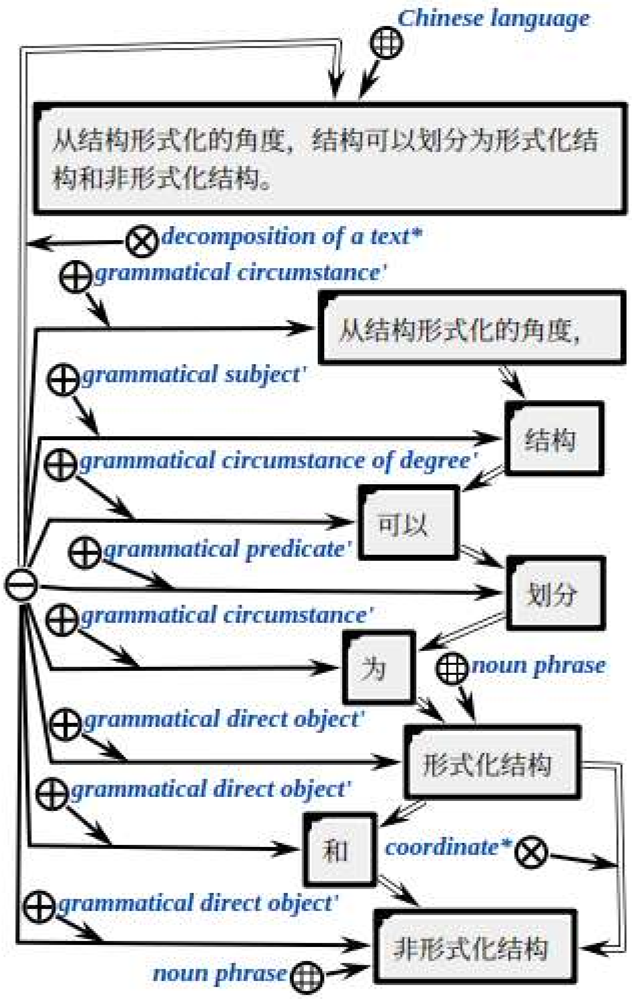

Chinese language
从结构形式化的角度，结构可以划分为形式化结构和非形式化结构。
decomposition of a text*
grammatical circumstance'
从结构形式化的角度，
grammatical subject'
结构
grammatical circumstance of degree'
可以
grammatical predicate'
划分
grammatical circumstance'
为
noun phrase
grammatical direct object'
形式化结构
grammatical direct object'
和
coordinate*
grammatical direct object'
非形式化结构
noun phrase

Figure 8: The determination of named entities and syntactic semantic relations in the input Chinese sentence

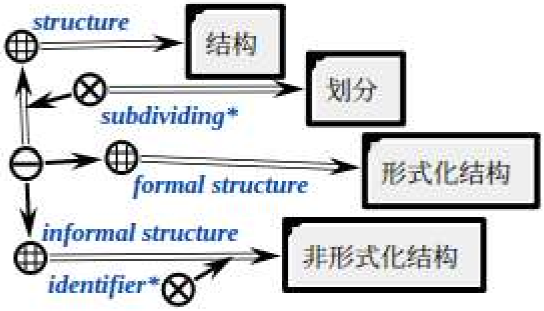


Figure 9: The constructed fragment of knowledge base from the input Chinese sentence

B. text generation from knowledge base

The current version of the component of Chinese texts generation from fragments of knowledge base for the intelligent help system for discrete mathematics is implemented in accordance with the corresponding general text generation architecture. In knowledge base on Chinese language processing, in addition to constructing linguistic knowledge (e.g., rules for text processing, extraction rules), linguistic knowledge for text generation (e.g., templates for text generation) can also be constructed in certain subject domains of knowledge base.

For fragments of knowledge base of ostis-system, the realization of natural language generation is roughly divided into two steps: rule-based symbolic generator converting fragments (sc-structure) of knowledge base into message triples; rule and template-based approached or statistical generator (when high quality aligned datasets is accessible.) translating message triples to resulted natural language texts. As mentioned above, unfortunately, the high quality aligned datasets is relatively difficult to access. The example about text generation is just to demonstrates the generation process of a simple narrative Chinese sentence using the rule template-based approach.

Due to the complexity and diversity of text generation tasks, in the example about text generation there are following several constrains:

- The input fragment of knowledge base is completed and has sense;
- The input fragment formally represented in the knowledge base for discrete mathematics;
- The output of Chinese language interface is a simple narrative Chinese sentence;
- The sc-nodes denoting concepts, named entities or relations within the input fragment of knowledge base have corresponding identifiers in Chinese language, which might be used in the resulted sentence.

Step 1 The Chinese language interface is provided with a given fragment (sc-structure) from knowledge base of intelligent help system for discrete mathematics represented in the form of the SC-code. For visual representation of given fragment of knowledge base, the fragment formally is represented in SCg (Figure 10).

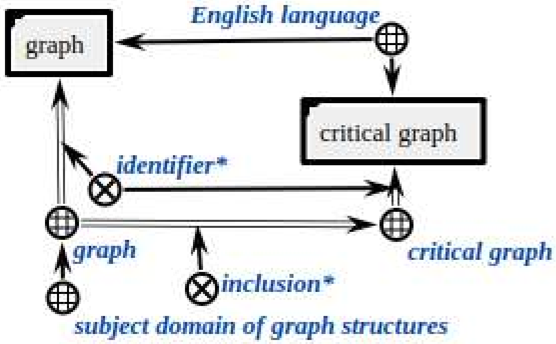


Figure 10: The fragment (sc-structure) of knowledge base for discrete mathematics

Step 2 The given fragment is divided into standard basic sc-constructions, afterwards from which the candidate sc-constructions are selected to be generated to the resulted Chinese texts. The candidate sc-construction (belong to standard basic sc-construction) is shown in SCg (Figure 11).

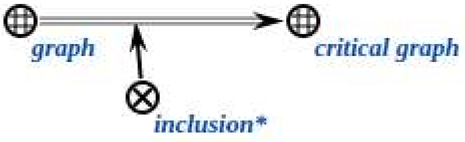


Figure 11: The determination of candidate sc-construction

Step 3 The candidate sc-construction is transferred to the message triple, each sc-element of message triples is a file (an sc-node with content) corresponds to a certain lexical unit (e.g., in the Figure 12 shown the sc-node "I_ch_graph") in the Chinese language. Specification of lexical units that is stored in the knowledge base on Chinese language processing. As seen in the Figure 12, the example just consider one type of the set of message triples.

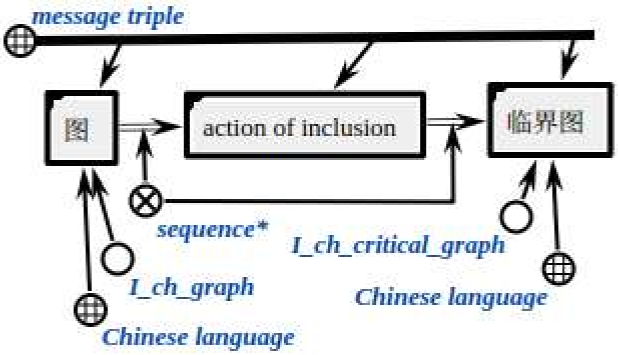


Figure 12: The message triple for candidate sc-construction

It is important to note that action of each message triple is the core that need to be replaced with corresponding text fragment (e.g., verb phrases or others) in order to generate fluent texts. Moreover, in general the subject and object of each message triple remain the same. Unless, in some cases of generating complex texts, they can be replaced by pronouns in the resulted texts.

Step 4 Based on the rule-template approach, for action of each message triple, a suitable rule or template constructed as logical ontologies in the knowledge base on Chinese language processing can be matched. For this example, the specific template shown in the Figure 14 will be applied.

Step 5 The agent fills the sc-links of corresponded sc-elements of that message triple, i.e. a certain sc-link needs to be filled with the result of a certain inflection form of a lexical unit. Finally the sc-links are concatenated to generate the resulted Chinese sentence according to valid ordering (Figure 13).

In the knowledge base there are different identifiers for each lexical unit in natural language. For Chinese language, there is not a certain inflection form for lexeme, therefore a lexical unit "graph" is expressed in resulted Chinese sentence is same as itself (Figure 13). Unlike European languages, in the resulted generated texts the lexical units in the sc-links require a specific inflection form (e.g., singular or plural and other inflection form) according to the syntax of a specific language. However

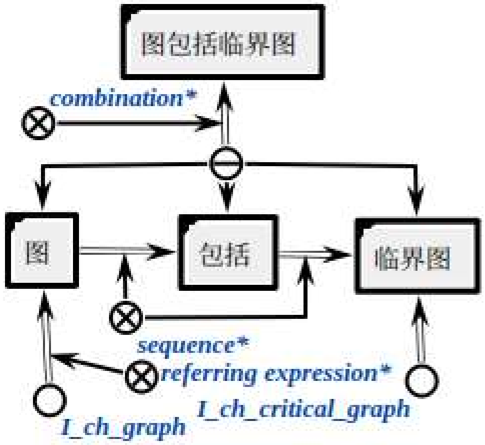


Figure 13: The resulted generated Chinese sentence corresponding to sc-construction

due to the features of Chinese language, the processing of this step is relative simpler. In this example we just consider the generation of the simple declaration sentence, the referring expression of the lexical unit means that the final form of this lexical unit in the resulted text. Based on the template, the referring expression to the lexical unit "graph" is served as the subject. The lexical unit "critical graph" is served as object in Chinese language. In some cases, the template is predefined with fixed phrase, it means that there are some sc-link in the template without corresponding lexical unit is the fixed part [41].

When these steps are implemented, a fragment sc-structure of knowledge base of Discrete Math domain can be transferred into Chinese sentence with specific sense. This natural language text is easier to access to ordinary end-users.

IX. Conclusion

This article has proposed an ontological approach based on the OSTIS Technology to develop a unified semantic model for natural language interface of knowledge-based intelligent system, which has ability of converting the natural language texts into the fragments of knowledge base and generating the natural language texts from the fragments of knowledge base. Within the framework of OSTIS Technology the development of semantic model for natural language interface mainly requires the development of sc-model of knowledge base of linguistics, which represented in the form of described subject domains and corresponding ontologies about linguistic knowledge, as well as sc-model of problem solver of natural language interface, which consists of sc-agents developed independently each other using various problem solving models to solve corresponding tasks in the natural language interface. The semantic model of natural language interface is universal and can be used for the implementation of various natural language interface of knowledge-based intelligent system in the specific subject domain. Within the developed semantic model of natural language interface, according to the features of the specific natural language, it becomes possible to develop knowledge base, within which includes linguistic knowledge in various level (lexical, syntactic, semantic and so on) about the specific natural language, and the corresponding problem solver on the specific natural language processing, in turn, to implement the specific natural language interface of knowledge-based intelligent system.

The developed semantic model of natural language interface using ontological approach mainly offers the following advantages:

- The component of converting the natural language texts into the fragments of knowledge base is applied to extract the knowledge structure represented in the form of SC-code from the natural language texts without predefined categories of entities and relations (i.e. from open domains);
- The component of generating the natural language texts from the fragments of knowledge base is applied to generate fluent, coherent, and multi-sentence natural language texts appropriating for end-users. For text generation, the semantic structures (fragments of knowledge base) that processed by this component is more complex than simple tabular or triples structure;
- The knowledge base of linguistics is constructed in the form of subject domains that are as independent of each other as possible and corresponding ontologies about linguistic knowledge. The hierarchical structure makes it possible to consider linguistic knowledge in various level (lexical, syntactic, semantic and so on) into a single knowledge base, as well as reduce development complexity of knowledge base and increase the development efficiency;
- The multi-agent model to develop the problem solver of each component in the natural language interface is possible to integrate different approaches to extend, modify the sc-agents to improve the performance of interface independently, without any change in other sc-agents;
- The modular and component approaches are considered to develop the natural language interface underlying a unified semantic basis. It makes sure the efficiency and semantic uniformity of each component development in the natural language interface of knowledge-based intelligent system.

We discussed the general structure of knowledge base of linguistics, which provides various types of linguistic knowledge to knowledge extraction and text generation. In addition to general declarative knowledge, within the constructed knowledge base includes knowledge for logical reasoning represented in logical ontologies

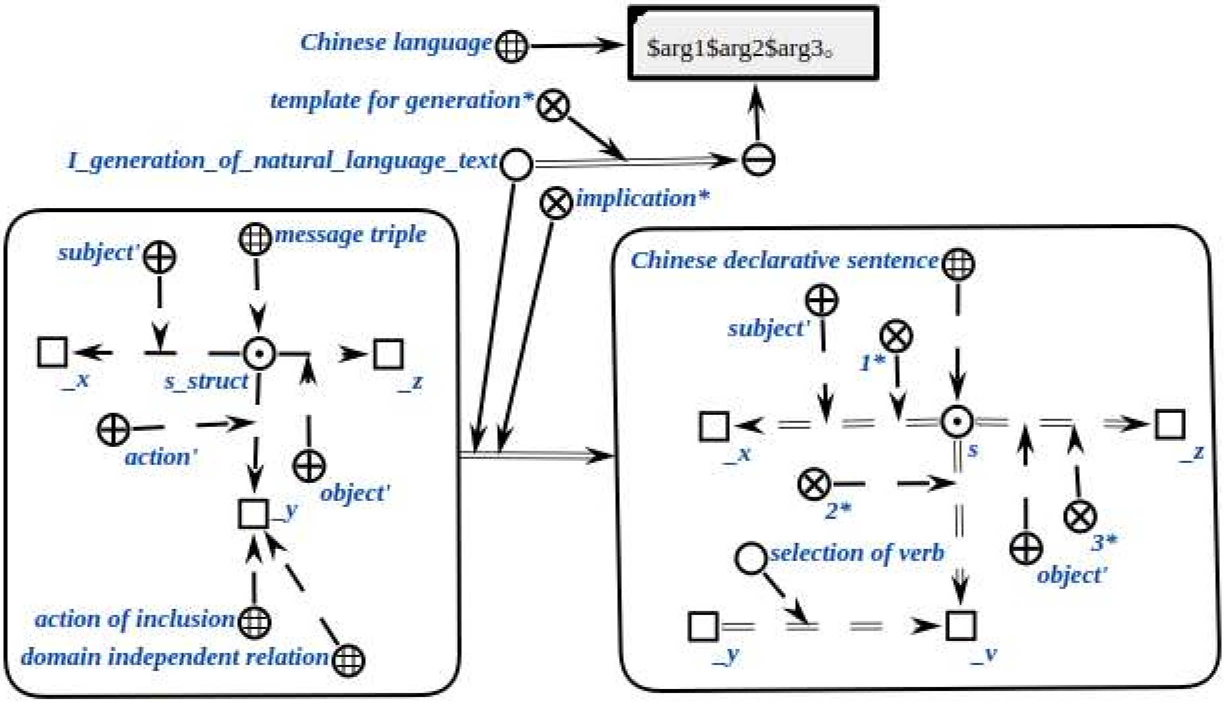


Figure 14: The template for generating natural language text

(e.g., the rules for knowledge extraction, the rules or templates for text generation). Relying on the sc-agents of each component to use constructed static linguistic knowledge to realize automatic knowledge extraction and text generation, it makes the intelligent system to obtain knowledge to help end-users, as well as the information of intelligent system easier accessible not only to computer programs, but also to end-users.

Finally the Chinese language interface of intelligent help system for discrete mathematics is implemented in the trial on the basis of the proposed semantic model of natural language interface to verify the practicality and generalization of the semantic model. In addition to Chinese language, it would be particularly interesting to explore the semantic model's possibility to support knowledge extraction and text generation for multiple language as long as the corresponding knowledge base on specific natural language processing (ontologies on specific natural language processing) and suitable sc-agents are developed.

References

[1] Liu Y. C.: Survey on Domain Knowledge Graph Research. Computer Systems Applications, 2020, vol. 26 No 06, pp. 1-12.

[2] Knowledge graph: the foundation for big data semantic link. Available at: http://www.cipsc.org.cn/kg2/. Date of access: 03.05.2022.

[3] Commonsense knowledge (artificial intelligence). Available at: https://en.wikipedia.org/wiki/Commonsense_knowledge_(artificial_intelligence)/. Date of access: 01.09.2022.

[4] Xu Z. L., Sheng Y. P., He L. R., Wang Y. F.: Review on Knowledge Graph Techniques. Journal of University of Electronic Science and Technology of China, 2016, vol. 45 No 04, pp. 589-606.

[5] Qian, L. W.: Ontological approach to Chinese text processing. Doklady BGUIR, 2020, vol. 18 No 06, pp. 49-56. (In Russian).

[6] Zhao H. X., Li L., Wu X. D., He J.: Knowledge Graph Oriented Information Extraction. Hans Journal of Data Mining, 2020, vol. 10 No 04, pp. 282-302.

[7] Li D. M., Zhang Y., Li D. Y., Lin D. Q.: Review of Entity Relation Extraction Methods. Journal of Computer Research and Development, 2020, vol. 57 No 07, pp. 1424-1448.

[8] Schutz A., Buitelaar P.: RelExt: A Tool for Relation Extraction from Text in Ontology Extension. International Semantic Web Conference, Springer, Berlin, Heidelberg, 2005, pp. 593 – 606.

[9] Banko M. J., Soderland S. Cafarella: Open Information Extraction from the Web. Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, 6-12 January 2007, pp. 2670-2676.

[10] Wu F., Weld D. S.: Open Information Extraction Using Wikipedia. Proceedings of Annual Meeting of the Association for Computational Linguistics, Uppsala, 11-16 July 2010, pp. 118-127.

[11] Fader A., Soderland S., Etzioni O.: Identifying Relations for Open Information Extraction. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, John McIntyre Conference Centre, Edinburgh, 27-31 July 2011, pp. 1535-1545.

[12] Schmitzm M., Bai R., Soderiand S.: Open Language Learning for Information Extraction. Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju, Island, 12-14 July 2012, pp. 523-534.

[13] Tseng Y. H., Lee L. H.: Chinese Open Relation Extraction for Knowledge Acquisition. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26-30 April 2014, pp. 12-16.

[14] Gatt A., Krahmer E.: Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research, 2018, vol. 61, pp. 65-170.

[15] Theune M., Klabbers E., Pijper de: From data to speech: a general approach. Natural Language Engineering, 2001, vol. 07 No 01, pp. 47-86.

[16] Yao T. F., Zhang D. M., Wang Q.: System Demonstration

Multilingual Weather Forecast Generation System. In Natural Language Generation, Niagara-on-the-Lake, Ontario, Canada, 1998, pp. 296-299.

[17] Reiter E. Sripada S., Hunter J. R.: Choosing words in computer-generated weather forecasts. Artificial Intelligence, 2005, vol. 167 No 01-02, pp. 137-169.

[18] Plachouras V., Smiley C., Bretz H.: Interacting with financial data using natural language. In Proc. SIGIR'16, Italy, Pisa, 2016, pp. 1121-1124.

[19] Androutsopoulos I., Lampouras G., Galanis D.: Generating Natural Language Descriptions from OWL Ontologies: the NaturalOWL System. Journal of Artificial Intelligence Research, 2013, vol. 48 No 01, pp. 671-715.

[20] Xia Z. T., Qu W. G., Gu Y. H., Zhou J. S., Li B.: Review of Entity Relation Extraction based on deep learning. In Proceedings of the 19th Chinese National Conference on Computational Linguistics, Haikou, China, Chinese Information Processing Society of China, 2020, pp. 349–362.

[21] Zhuang C. Z., Jin X. L., Zhu W. J., Liu J. W., Bai L, Cheng X. Q.: Deep Learning Based Relation Extraction: A Survey. Journal of Chinese Information Processing, 2019, vol. 33 No 12, pp. 1-18.

[22] Li J. Y., Tang T. Y., Zhao W. X.: Pre-trained Language Models for Text Generation: A Survey. Thirtieth International Joint Conference on Artificial Intelligence IJCAI-21, Montreal-themed virtual reality, 19th -26th August 2021, pp. 4492–4499.

[23] Claire G., Anastasia S., Shashi N.: The WebNLG Challenge: Generating Text from RDF Data. In Proceedings of the 10th International Conference on Natural Language Generation, Santiago de Compostela, Spain, 2017, pp. 124–133.

[24] Auer S., Bizer C., Kobilarov G.: Dbpedia: A nucleus for a web of open data. The Semantic Web, Springer, Berlin, Heidelberg, 2007, pp. 722-735.

[25] Xu B., Xu Y., Liang J.: CN-DBpedia: A never-ending Chinese knowledge extraction system. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, Cham, 2017, pp. 428-438.

[26] Niu X., Sun X. Wang H.: Zhishi.me – weaving Chinese linking open data. International Semantic Web Conference, Springer, Berlin, Heidelberg, 2011, pp. 205-220.

[27] Chinese General Encyclopedia Knowledge Graph (CN-DBpedia). Available at: http://www.openkg.cn/dataset/cndbpedia/. Date of access: 10.09.2022.

[28] Chinese Encyclopedia Knowledge Graph (Zhishi.me). Available at: http://openkg.cn/dataset/zhishi-me-dump/. Date of access: 10.09.2022.

[29] Amit M., Yoav G., Ido D.: Step-by-Step: Separating Planning from Realization in Neural Data-to-Text Generation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), USA, Minneapolis, 2019, pp. 2267-2277.

[30] Golenkov V.V. Gulyakina N.A.: Proekt otkrytoi semanticheskoi tekhnologii komponentnogo proektirovaniya intellektual'nykh sistem. Chast' 1 Printsipy sozdaniya [Project of open semantic technology of component designing of intelligent systems. Part 1 Principles of creation]. Ontologiya proektirovaniya [Ontology of designing], 2014, No 1, pp. 42-64. (in Russian).

[31] Davydenko, I. T.: Ontology-Based Knowledge Base Design, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Belarus Minsk, 2017, pp. 57-72.

[32] Davydenko, I. T.: Semantic Models, Method and Tools of Knowledge Bases Coordinated Development Based on Reusable Components, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Belarus Minsk, 2018, pp. 99–118.

[33] Shunkevich D.V.: Ontology-Based Knowledge Base Design, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Belarus Minsk, 2017, pp. 73-94.

[34] Shunkevich D.V.: Ontological approach to the development of hybrid problem solvers for intelligent computer systems, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Belarus Minsk, 2021, pp. 63–74.

[35] Boriskin A. S., Sadouski M. E., Koronchik D. N., Zhukau I. I., Khusainov A. F.: Ontology-Based Design of Intelligent Systems User Interface, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Belarus Minsk, 2017, pp. 95–106.

[36] Sadouski M. E.: Ontological approach to the building of semantic models of user interfaces, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Belarus Minsk, 2021, pp. 105–116.

[37] Shunkevich D.V.: Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Belarus Minsk, 2018, pp. 119–132.

[38] Qian L. W., Sadouski M. E., Li W. Z.: Ontological Approach for Chinese Language Interface Design, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Minsk, 2020, pp. 146–160.

[39] Hardzei ., Svyatoshchik ., Bobyor L., Nikiforov S.: Processing and understanding of the natural language by an intelligent system, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Belarus Minsk, 2021, pp. 123–140.

[40] Qian L. W., Li W. Z.: Ontological Approach for Generating Natural Language Texts from Knowledge Base, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Belarus Minsk, 2021, pp. 159–168.

[41] Wang H., Zhan W. D., Yu S. W.: Structure and Application of The Semantic Knowledge base of Modern Chinese. Applied Linguistics, 2006, No 01, pp. 134-141.

[42] Contemporary Chinese language word segmentation specification for information processing. Available at: http://std.samr.gov.cn/gb/search/gbDetailed?id=71F772D78FC6D3A7E05397BE0A0AB82A/. Date of access: 01.07.2022.

Онтологический подход к разработке естественно-языкового интерфейса

Цянь Лунвэй, Ли Вэньцзу

В статье рассматриваются существующие методы к разработке двух основных компонентов в естественно-языковом интерфейсе интеллектуальнных систем, т.е. преобразование текстов естественного языка во фрагменты базы знаний и генерация текстов естественного языка из фрагментов базы знаний. Был проведен анализ проблем, возникающих при преобразовании текстов естественного языка во фрагменты базы знаний и генерации текстов естественного языка из фрагментов базы знаний в настоящее время.

На основании различных рассмотренных методов был предложен онтологический подход к разработке естественно-языкового интерфейса, который позволяет интегрировать разные типы лингвистических знаний и методов в единой семантичесской модели для анализа текстов естественного языка и генерации текстов естетсвенного языка. Этапы реализации подхода были созданы лингвистические онтологии и решатели для анализа и генерации текстов естественного языка. Более того, в качестве китайского языка, функции каждых этапов анализа текстов и генерации текстов, а также назначение лингвистических онтологий в процессе генерации проиллюстрированы, чтобы проверять практичность модели.

Received 14.11.2022