CONCEPTUAL FRAME MODEL FOR THE PRESENTATION OF THE CONCEPTS AND RULES IN NATURAL LANGUAGE INTERFACE FOR DATABASE

M Zhekova¹ and G Totkov²

¹ PHD student, Department of Computer Science, University of Plovdiv „Paisii Hilendarski“, Tzar Asen str.24, Plovdiv, Bulgaria
² Prof., Department of Computer Science, University of Plovdiv „Paisii Hilendarski“, Tzar Asen str.24, Plovdiv, Bulgaria

m_jekova@uft-plovdiv.bg , totkov@uni-plovdiv.bg

Abstract: The article presents a frame model for conceptual description of a subject area by presenting a detailed structure of multiple objects and concepts, knowledge about them, hierarchies of types, rules, specifying classes and relations between them. Include all possible words and word combinations that formalize a subject area. The model itself represents knowledge and reflects the conceptual basis of the organization of the memory of a person, as well as its flexibility and visibility. When creating an appropriate database of accumulating frame models (AFM), frame questions, and at a later stage a frame model for automatic question generation, it is necessary to build an appropriate structure that allows for both flexibility and easier building of the databases of data, as well as efficient retrieval of different types of information when summarizing the results. The publication is part of a study devoted to the application of frame models for the presentation of knowledge and processes in e-learning.

1. Introduction

With native language processing, interfaces can be built to allow users to interact with data more intuitively using their native language. This paper provides a methodology for creating a question-response information system using a natural-language interface to gain knowledge for the benefit of users through a web-based system.

Natural Language Interfaces (NLI) are a kind of computer-human interfaces, where language forms such as verbs, phrases, and clauses act as parameters in the formation of queries to a data mining database. Retrieving database information requires in-depth knowledge of technical languages such as Structured Query Language [10]. By contrast, database NLI would allow questions to be asked in the native language of the user. Therefore, natural language interface information systems would be more suited to casual users as the user would not need to learn a language through which to communicate with the system.

This paper aims to compile a database query model by manipulating the queries introduced by the user, then differentiating and parameterizing the parts of the trait and reaching the rules and relationships that lead them. Steps in my work includes create frame models, define parameters for
these models and then we use models and semantic defined parameters to produce user outputs. NLI is not just a question-response dialog based on a given knowledge base but also access to structured data modeled and designed syntactically and semantically classified. For this purpose, it is necessary to precisely define the scope of the subject area, correctly examine the peculiarities of the lexical units and phrases, to construct a system for their linguistic description and classification and to create a methodology for their recognition and tagging with grammatical and syntactic information. A major challenge will be to solve the problem of natural language comprehension, how to interpret the query data, and how the user's information is transformed into a question.

The purpose of the development is to construct a conceptual map in the form of a linguistically correct and argued theoretical frame model for presenting a database of concepts, relationships and relationships by describing the individual language units and rules, and then able to derive this information and identify all SQL elements included in frames, slots, attributes, aggregates, values from the input entry queried using pre-created rules for the purpose. In this way, a new database is being built - a computer model full of content extracted from various sources. And once we have the knowledge base from the limited subject area, we can derive the data from all these sources.

2. Related work

One of the tasks of computer linguistics is the creation of software programs and applications that, with the help of the knowledge formalizations realized so far, facilitate the access to the accumulated resources and the huge amount of information and work with it. Like the systems that generate text from documents, question-response systems rely on complex techniques for natural-linguistic analysis.

Here are some systems that are oriented to a multi-contextual knowledge base that describes both concepts and roles that correspond to objects in the subject area, as well as rules for constructing a query to the database and for constructing concepts.

- An information system that processes the user's question determines the type and focus of the question, the type of response, and then uses that information to create the query is a system based on retrieving the information. The PLINIUS system is designed to automatically extract knowledge from texts in chemistry [12]. Several sets of concepts such as a chemical element, integer, group, etc. are defined in it and rules for constructing the other concepts. About 150 concepts and 6 rules are described in the ontology describing the system [13]. It described using frame constructions.

- Knowledge-based systems generate a semantic description of the question that is then used to search in structured resources. They are suitable for limited domains such as tourism, medicine and transport, where there are no large volumes of pre-recorded documents. Knowledge-based approaches are based on the fact that no new rules need to be added for each new domain or language of the application. Apple's SIRI is such a system. It contains information, messages are sent, events are added to a calendar instructions are received. The application uses machine learning technology to make it "smarter" and more capable of predicting and understanding the questions and requests in a natural language by the user.

- In hybrid systems, like IBM's WATSON, while the issue is semantically analyzed, responses are generated using information retrieval methods.

3. Frame model description

In artificial intelligence systems, the presentation of the database in the form of a conceptual model is close to the way the concepts are stored in the human brain.

In the development we present a method for creating a natural-language interface used for the purpose of semantic search in structured data and content through a system of questions and answers.

The description of the proposed model is presented in figure 1, which is a University database created with MySQL for a university.
Classifying, maintaining and storing information as well as defining relations between data aims to facilitate access to the database as well as to facilitate communication between the user and the computer.

Several correctly normalized tables have been created in the University database - Branch, Faculty, Specialty, Discipline, Student,... Obtaining a response from this database, however, requires SQL knowledge to be learned and assimilated.

If we look at the example: User wants to view information such as - the names and codes of the courses studied by a student which faculty number equals 16810235411, whose specialty is "Business Information Technologies". From the Specialty table of the University database we use the following SQL query:

```sql
SELECT discipline_name, discipline_code
FROM discipline
WHERE specialty_name = "Бизнес Информационни технологии" and student_facNumber = "16810235411"
```

Native language processing and SQL query to a Specialty Table in University database will result in the same output, and the question is not a SQL query, but the user's natural language.
What are the name of the discipline and code of the discipline which specialty name equals "Business IT" and faculty number of student equals "16810235411"?

In this database we can ask following questions:
- Who are the lectors working in the Department of Computer Science? and
- Who are the lectors reading lectures in the Department of Computer Science?

4. Scope of the Frame model

Scope of the proposed frame model is:
- For the purpose of testing, a University database (figure 1) has been created. For a relational database management, knowledge of software for working with MySQL databases is required.
- Language for working with the system for facilitating a wider range of users has been chosen to be Bulgarian.
- User input is in the form of questions - Who? What? What? Where? When?
- Is used a detailed structure of multiple objects and concepts, knowledge about them, hierarchies of types, rules, specifying classes and relationships between them to describe the corresponding element. Is included all possible words and word combinations that formalize a subject area. The list is updated with words and phrases specific to the specific domain area.
- The question divides into lexical units or phrases from which it is composed, giving each unit a unique number (Tokenization).
- To remove unnecessary words from the user's question. In the table 3 are specific domain area words not bearing important information, which will be updated periodically.
- Identify all SQL items included in tables, attributes, aggregates, values from the input entry queries, using pre-created rules for the purpose.
- Create an SQL template using follow template:

```
ID of slot,
ID of table name,
ID of and, ID of aggregate,
ID of interval, ID of value
```

- Create a SQL query using SQL Elements.
- For the avoidance of ambiguity, frames have been assigned a frame ID field, frame name, attribute ID, and attribute name to store all the attributes belonging to a frame.

5. Frame as a concept

The frame, according to [1], is used to represent the minimum structured information that defines unity phenomena, events, situations or processes. The term was introduced in the artificial intelligence by M. Minsky [2] as a means of presenting knowledge. According to [7], the frame model is used to represent relatively independent, logically distinct units of knowledge with the possibility of multiple use in different situations (including retrieval and aggregation of data).

Each frame has a unique name and structure consisting of stacked elements called slots. Each slot has a unique name in the frame, it contains certain information, and it carries classification information about the syntactic function and the belonging of the language unit located in it to part of the speech. The content of each slot can be a numeric or text value, predicate dependence, function, or other frame. A group of frames connected to each other forms a system [3]. The structure of the frame is dynamic, ie. allows adding new slots at any time [6].

6. Frame model description

The article presents a frame model for conceptual description of a subject area by presenting concepts and relations with their concrete representatives - prototypes and specimens. The model itself
represents knowledge and reflects the conceptual basis of the organization of the memory of a person, as well as its flexibility and visibility. When creating an appropriate database of accumulating frame models (AFM), frame questions, and at a later stage a frame model for automatic question generation, it is necessary to build an appropriate structure that allows for both flexibility and easier building of the databases, as well as efficient retrieval of different types of information when summarizing the results. In the presented computer model, the task is to synthesize text to answer a question asked by a system user.

The accumulation frame model (the set of interconnected frames) is both a form, method and means of learning, logical and consecutively ordered knowledge structures composed of generalized structural models (blank frame-prototype) and their specific, filled-in representatives (frame-specimen). Both the descriptive and basic knowledge for each domain area are found in the model described above.

Table 1. A common structure of the frame model with the n-slot.

| Frame name: | slot 1 name (content of slot1); |
| slot 2 name (content of slot2); |
| slot 3 name (content of slot3); |
|…… |
| slot N name (content of slotN); |

Frame models, composed of slots describing concepts and surations of reality, are presented within the article. Each slot is described by one or more attributes that represent knowledge describing the element. Attributes or so-called veneers serve to describe both static knowledge - values and dynamic knowledge through so-called "attached procedures" (methods) or "demons" [6]. Demons are a special type of procedures that start automatically when an event occurs. The value of a slot is of a certain type of data, which may be elementary or compiled.

The purpose of classifying data in frames is to maintain and store the data, to define the links between them and to facilitate access to them.

6.1. Description of lexical unit

According to [4], a lexical unit is the ordered three - lemma, lexical meaning, examples. In Table 2 is presented a frame-prototype - Concept with a description of the slots, defined exhaustively with the terms of computer linguistics and their content.

The frame model shown in Table 2 describes a separate lexical unit, the so-frame-prototype with a detailed description of its slots and content. In this way, the terms used in the restricted domain area are described.

In the conceptual model, elements of the system under consideration involve different concepts with their connections described by facts and rules from the respective domain area. Structurally, it is a systematically bound concept with a different level of abstraction and a prototype of the future database.

Table 2. Frame-prototype – Concept.

| Name | (name / full name of frame); |
| EN  | (name of concept in English) |
| Description | (lexical description, meaning) |
| Sunonims | (synonymous order of the frame name); |
| Domain | (limited subject area); |
| Question Word | (questions that can be asked over the name of the frame) |
| Part Of Speech | (grammar category) |
| Semantic class | (belonging to a semantic class) |
| Semantic role | (semantic function / role) |
| Sub Frame Of | (inherit from) |
| Is Argument In | (participates as an argument in a predicate structure) |
We will present the database as a set of frame models. From the frame-prototype, by filling the slots with specific values, a frame-specimen is obtained.

7. Frame model architecture
The natural-language interface for database deals with translating user queries from its native language to the corresponding SQL query to retrieve the relational database data via a web based information system. The result is displayed on the user screen. This web-based system is initially developed for Bulgarian, and other languages will be discussed later. An algorithm for mapping a natural language, implemented in Bulgarian, has been developed effectively to convert an SQL query to create appropriate responses. The algorithm is conceived with C#, Apache, MySQL and tested successfully. The system architecture is presented in figure 1, which figure the process layout involved in transforming the user query into the NL query into a syntactic SQL query that is created by relational database management system and receives database responses.

| Table 3. List of words that are not involved in the SQL query. |
|-----------------------|-------------------|
| are | is | which | that | for | in front of | this | that | like | a | in | those | on | towards | throw | will | under | another | over | must | however | anyway | yet | though | still | else | some | even | only | indeed |

| Table 4. Frame – prototype – Person. |
|-----------------------|-------------------|
| Person: Name | (text value – first name, last name ) | Who? |
| ID_card | (integer) | What? |
| City | (text value) | Where? |
| Address | (text value) | Where? |
| Post code | (integer value) | What? |
| Sex | (text value) | What? |
| Nationality | (text value) | What? |
| Sub Frame Of (Person) Is Argument In | live_in {Person, Address}, live_in_city {Person, City}, have_ID {Person, ID}, sex {Person, Sex}, nationality {Person, Nationality} ) |

The term Student is described by the criteria that we can outline a student.

| Table 5. Frame – prototype – Student. |
|-----------------------|-------------------|
| Student: Frame name | (Student) | Who? |
| Student name | (Person.Name) | What? |
| Faculty number | (text value) | What? |
| Semester | (integer value) | Which? |
| Evaluate | (function) | How much? |
| Edu form | (редовно/задочно/платен) | What education form? |
| Edu degree | (bachelor/master) | What education degree? |
| University | (text value) | Which university? |
| Specialty | (text value) | What specialty? |
| Diploma number | (text value) | What diploma number? |
| Personal skills | (text value) | What personal skills? |
| Role | (text value) | What role? |
| Internship | (text value) | What internships? |
| Sub Frame Of (Person) Is Argument In | study {Student, Specialty}, study {Student, Discipline}, is_in {Student, Degree}, live_in {Student, City}, ID {Student, ID_card} ) |

...
All values from Person frame - Person.Name, Person.Address, Person.Nationality .... are inherited from Student frame because Student is a Person type. Student is a Person and all attributes of Person are attributes of the Student.

When the user asks if there are more criteria in the question, if the search is more parameters, it is convenient to arrange in a sequence of parameters / parameters here are the names of the slots /.

**Table 7.** Rules for representing some different frame in University database.

| Students  | table Student | Rule for representing table Student |
|-----------|---------------|-------------------------------------|
| Faculties | table Faculties | Rule for representing table Faculties |
| Specialties | table Specialties | Rule for representing table Specialties |
| Disciplines | table Disciplines | Rule for representing table Disciplines |
| Teachers  | table Teachers  | Rule for representing table Teachers  |
| Internship | table Internship | Rule for representing table Internship |

**Table 7.** Rules presenting different attributes of frame Student.

| Student name | attribute Student_Name | Rule for representing attributes |
|--------------|------------------------|-----------------------------------|
| Student faculty | attribute Fac_Number | Rule for representing Fac_Number |
| Semester | attribute Semester | Rule for representing Semester |
| Education degree | attribute Edu_Degree | Rule for representing Edu_Degree |
| University | attribute Uni_Name | Rule for representing Uni_Name |
| Personal Data | attribute PersonalData | Rule for representing Personal Data |
| Specialty | attribute Specialty_Name | Rule for representing Specialty_Name |

**Table 8.** Rules for representing an ‘and’, ‘as well as’, ‘or’ and ‘nor’ concept.

| and     | s_and | rule for representing And |
|---------|-------|----------------------------|
| as well as | s_and | rule for representing And |
| also    | s_and | rule for representing And |
| in addition | s_and | rule for representing And |
| or      | s_or  | rule for representing OR |
| not     | s_not | rule for representing Not |
| not include | s_not | rule for representing Not |

**Table 9.** Rules for the aggregate functions.

| most     | aggregate_max | rule for representing max() |
|----------|---------------|-----------------------------|
| maximum  | aggregate_max | rule for representing max() |
| highest  | aggregate_max | rule for representing max() |
| maximum number of | aggregate_max | rule for representing max() |
| highest number of | aggregate_max | rule for representing max() |
| smallest | aggregate_min | rule for representing min() |
| minimum  | aggregate_min | rule for representing min() |
| smallest number of | aggregate_min | rule for representing min() |

**Table 10.** Rules for the interval ‘equal’ concept.

| equal    | interval = | rule for interval equal |
|----------|------------|-------------------------|
| exactly  | interval = | rule for interval equal |
| match with | interval === | rule for interval absolutely equal |
| less than | interval < | rule for interval less |
| bigger than | interval > | rule for interval greater |
| less than or equal to | interval <= | rule for interval less or equal |
| greater than or equal to | interval >= | rule for interval greater or equal |
| different from | interval <> | rule for interval different |
8. **Algorithm used in natural language system**

The following steps describe the algorithm for the proposed natural-language system.

- Checking for an input value and if it exists, the value is removed from questional string.
- Check all the words in the query string whether they exist in the resource dictionary / dictionary of the domain area.
- Tokenization
  - Split the questional query of tokens
  - The ID number of each identified token is given
- Remove unnecessary words / words that are not essential / from the string.
- Make a comparison by removing each last word from the question string. If compliance with an existing rule is found, the corresponding word / phrase / slogan is removed from the query string and the action is repeated until the string is run out until the query string becomes null.
- The arrays store the IDs / IDs of the identified SQL elements.

**Figure 2.** Architecture of a web system with a native-language interface to connect to a database.
An SQL template string is constructed considering the IDs of SQL elements available.

- Mapping the new SQL string with SQL templates.
- Generate SQL query using SQL Elements.
- The system returns an answer.
- Show the result in the user interface.

In the computer analysis of the user query first, it is divided into words and punctuation (tokenization) and then proceeds to semantic analysis at different levels where each token is defined by grammatical characteristics, belonging to syntactic categories, parts of speech, semantic classes and dictionary meaning.

In the abstract structure of the Bulgarian language, the sentence exists in the form of a scheme called a construction of sentence [9]. The sentence construct includes certain positions corresponding to given syntactic roles and arranged in a particular scheme.

In the proposed theoretical frame model, the criteria and features used to describe each slot of the frame provide information about:

- Semantic role / relationship - The purpose of semantic roles is to summarize similar semantic properties of classes of arguments. Semantic (theta) roles are reviewed by J. Penchev (1999), who distinguishes the following basic relations: agent, actor, experimental, elemental, object, patient, addressee, topic, locative, temporativ.
- Selective Restrictions - This classification divides the nouns into abstract and concrete, animate and inanimate, common and own.
- Part of the speech / Syntactic function - subject, verb, object, direct supplement.
- Part of a syntactic category - Latin alphabets are used to denote the parts of the speech to which the final elements of the sentence belong: Noun - noun, Adj - adjective, Verb - verb, P - preposition, Adv - adverb, Conj - conjunction, Pron - pronoun, Num - numerical. S is the category - sentence.

9. Conclusion

In computer systems for artificial intelligence, the presentation of the database in the form of a conceptual model is close to the way the concepts are stored in the human brain. Natural language generation is a task for generating understandable answers, natural language texts from a computer software application, such as a database or logical form, such as a frame. Frames are also the sources from which the data will be derived in the proposed theoretical model. In the study is presented model for classifying concepts, relations and rules in the domain area. The model will be explored experimentally and analyzed in subsequent publications.

The proposed methodology is used for a question-answer information system in order to leverage knowledge for the benefit of users in solving organizational communication problems. The dependencies between words or phrases in the user query are only identified with the links that exist between concepts and rules in the subject area knowledge base. The approach strongly depends on the qualitative selection of knowledge about objects, rules and relationships between them.

Future developments will add to the existing conceptual frame model and syntactic semantic analyzer. The natural language interface for dialogue with an information system has been developed as a start in Bulgarian, with the issue of developing other languages as a task for further research.

Acknowledgements

The paper is partly supported within the project MU17-FF-023"Accumulative frame models for extraction and aggregation of data for knowledge and processes in learning" of the Scientific Research Fund at the University of Plovdiv.


References

[1] Zhekova M, Totkov G and Panayotova T, Accumulative frame models for generating test units in e-learning systems, Scientific works of the Union of scientists in Bulgaria - Plovdiv, Series C. Technique and Technology, Volume X, 2017

[2] Minsky M, A Framework for Representing Knowledge, MIT, Cambridge, 1974
Yordanova S and Stefanova K, Retrieving knowledge from unstructured data through consumer opinion analysis, issue 1, 2017

[3] Shivacheva G and Totkov G, Accumulative frame models in programming training, Oktober 2017

[4] Koeva Sv, Bulgarian FrameNet, Semantic-syntactic dictionary of the Bulgarian language, print edition, Sofia 2008

[5] Nestorova P, Semantic and syntactic peculiarities of predicates in the Bulgarian language, print edition, Plovdiv 2008

[6] Doneva R, Gaftandzhieva S and Totkov G, Frame Models in Programming Training, XI National Conference "Education and Research in the Information Society", 2018

[7] Doneva R, Gaftandzhieva S and Totkov G, Frame representations in e-learning applications and new developments, International Journal on Information Technologies & Security, No 2 (vol. 10), 2018
Penchev Y, Construction of the Bulgarian Sentence, Sofia 1984

[8] Penchev Y, Bulgarian syntax. Manage and Connect, Plovdiv 1993

[9] Kolev K, Course project in linguistics

[10] Rukshan A, Prashanthi R and Sinnathamby M, Natural Language Web Interface for Database (NLWIDB), Proceedings of the Third International Symposium, SEUSL: 6-7 July 2013, Oluvil, Sri Lanka
Barkalova P, Bulgarian syntax - familiar and unknown, 1997 Sofia, Abagar
Stoyanova Iv, Automatic recognition and tagging of composite lexical units in the Bulgarian language, 2012 Sofia

[11] Popov D, Bulgarian vocabulary, ed. "Science and art", 1994
Fellbaum C, WordNet - An Electronic Lexical Database, MIT Press, 1998

[12] Van der Vet P, Speel P and Mars N, The PLINIUS Ontology of ceramic materials, April 1995

[13] Terzieva T and Totkov G, An ontological approach to the learning process Ruppenhofer J, Ellsworth M, Petruck M R L, Johnson C R, Scheffszyk J, FrameNet II: extended theory and practice, Technical Report, 2005