Multilingual Question-Driven Approach and Software System to Obtaining Information from Texts

Olena Chebanyuk [1,2]

[1]*National Aviation University, Kyiv, Ukraine*
[2]*V.N. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv, Ukraine*

Abstract

An approach, allowing obtaining information from a text that is relevant to users' question, is proposed in this paper. The advantage of the proposed approach is using the same algorithms for analysis of texts from different groups of languages, namely for Latin and Cyrillic alphabets. The most relevant sentences from the text that match to concrete types of questions are found by means of matching questions and attributes of answers in texts. Such attributes are specific for the concrete type of the question. An ontology aimed to consider key information about questions and attributes of answers also is proposed. This ontology allows adding new languages, types of questions, and attributes of answers. Architectural solution of the software system for operations of ontology processing and searching of answers, is represented in the paper. A case study illustrates the process of text processing. As an example of the text, a Bulgarian legend was chosen, because the style of preparing fairytales is one of the informal styles of text representation.

Keywords

Text Processing on Different Languages, Ontology, Semantic Analysis, Software System for Text Analysis, Software Architecture,

1. Introduction

Nowadays values of information in digital world are increased. In order to perform effective search user must proceed a large amount of data. In order to take this process more effective different software systems for searching and processing of information are used. Existing systems have the next difficulties to use: systems are not flexible (based on artificial networks, based on ontologies), expensive, difficult to changes, difficult to be adopted to different languages or environments. From the other side the designing, development, and supporting of such systems requires many efforts to adopt software system for different purposes.

Solution may be in the area – flexible systems, open for other languages, allowing filtering user requests and reducing the number of searched information.

Paper is organized by the following way: Chapter 2 represents literature review and summarizing of drawbacks of existing approach. Chapter 3 represents main steps of the proposed approach. Chapter 4 illustrates a part of designed ontology for different types of questions. As examples, Bulgarian and English languages are considered. Chapter 5 represents the software architecture of the software system for realizing of the proposed approach. Chapter 6 contains a practical example of preparing XML files, as external attributes for two languages. Detailed information for domain analytics about testing the proposed approach is represented in Chapter 7. Chapter 8 illustrates the results of Case Study for asking questions about Bulgarian legend. It is known, that in fairytales and legends the freest style of representing ideas is admitted. That is why, such complex texts are selected for testing. Chapter 8 contains description of main class for corresponding software system realization. Conclusions represents quantity estimation of the proposed approach and ideas for the further research.

[1]13th International Scientific and Practical Conference from Programming UkrPROGP'2022, October 11-12, 2022, Kyiv, Ukraine
EMAIL: chebanyuk.olena@gmail.com (A. 1)
ORCID:0000-0002-9873-6010 (A. 1)

2. Literature review

Paper [1] represents a software system is used to determine the degree of semantic similarity of two short texts written in Serbian. An approach allowing performing Semantic Similarity of Short Texts in Languages. This approach is consists from the next steps:

Corpus acquisition deals with finding a sufficiently large set of texts that could be used to generate a semantic space.

Corpus parsing is used to remove any superfluous information from further consideration. (For example specific XML tags or other, irrelevant data.)

Corpus preprocessing serves to reduce the amount of different words in the corpus, effectively reducing the context vector dimension:

1. Text cleaning – this includes the deletion of all text characters not belonging to the native script of the language in question, the removal of numbers and words that contain numbers, the elimination of punctuation marks and the shifting of all capital letters into lower case [1].
2. Stop-words removal – stop-words are auxiliary words like prepositions, pronouns, interjections and conjunctions, which carry negligible semantic information, but which are often encountered due to their language function. By removing those words, we decrease the total number of different words in the corpus [1].
3. The result is that the semantic space is reduced and the accuracy of the semantic algorithms is increased, since the links between semantically important words become more emphasized. The stop-word was formed by gathering the most frequent words from the text corpus. (General knowledge were taken from an encyclopedia). The information about word frequencies in the corpus which is gathered in this step is saved for later use in calculating various Term Frequencies (TFs) for each word [1].
4. Stemming – (solving coding problems, for example comparing UTF-8 format and is written partially in Cyrillic and partially in Latin alphabet or ASCII coding system. In order to preserve compatibility with the stemmer module, special coding system was designed [1].

Choosing an algorithm for the creation of the semantic space and supplying it with the preprocessed corpus text.

The reduction of context vector dimension. Each algorithm has its own post-processing routine which is encapsulated within the algorithm, as defined in the S-Space package [1].

Paper [2] presents the approach of defining entities in court decisions. It is proposed to prepare courts' decisions in the special structure of documents in order to simplify search procedure. Also classification and advantages and drawbacks of different software systems for semantic analysis is represented. Preparing unified structure of a document simplifies search procedure, but requires additional efforts for preparing of document in a specific view.

Approaches devoted to analysis of object state in decision support systems allows to analyze different states of objects (and as a conclusion – characteristics of entities) with the aim to extract metainformation about entities and get the answers to questions in the text. Such approaches may be implemented if there is an information that state of object or domain entities may be changed during story [3].

Other implementation of question-oriented analysis of texts is to get questions essential for specific group of users to predict their relation, emotions, and opinions about texts. Typical question may be applied to many different texts with the aim to select the best test considering some conditions of chosen social group. For example searching the most appropriate texts fro children [4].

2.1 Conclusion from the review and requirement specification for the approach

After analyzing the existing solutions for processing of texts, requirement specification with the essential software requirements was designed. Points that has the highest level of priorities are considered in this paper.

Table 1
Challenges for the approach allowing to process text.

RQ code	RQ description	Priority
F0	Flexible extension of the system allowing to add new types of questions	10
F1	Convenient representation of answers	10
F2	Searching answers in different languages.	10
F3	Usability: the system should be easy to use and find the answer to the question.	10
F4	Processing text with some grammar mistakes	6
F5	Processing text in different formats and encodings	6
F6	Process an information from various sources of texts	6
F7	Search the exact answer to the question	3
F8	Processing texts containing smiles or special symbols	3
F9	Recognizing texts from images	3
F10	Processing answers in text to speech modules;	3

3. Proposed approach

High priority requirements are challenges for

1. Define type of the question and its key attributes
2. Define attributes related to answers for concrete type of question
3. Define all most relevant sentences according to attributes of concrete type of question and represent them to user.

4 Ontology for different types of questions

It is proposed to classify domain entities according to defined characteristics. One entity may correspond to several classes. The list of the proposed characteristics for questions is given below (Table 2 and Table 3). Aim of this classification is to take marks in text according to types of questions.

Table 2
List of attributes that are **extracted from questions**

Attributes for answers in questions	Question words in Bulgarian	Question words in English
(1) alive entities	кой?	who?
(2) domain entities of place, which are defined by special prepositions in text	къде?	where?
(3) entities of time	кога?	when?
(4) not alive entities	какво е това?	what is this?
(5) numeric entities	колко?	how many(much)?)
(6) event entities	как?	How?

Table 3
List of attributes **extracted from texts**

Description of attributes	Examples for English language	Examples for Bulgarian language
(1) alive domain entities	Defined by domain experts	
(2) Place entities	At the in the, near, far from, under, above	на, в(във), близо, далеко, под, над
(3) Time entities	parts of the day (day, night, evening, morning, at __ o'clock) parts of year (mounts)	Части на денонощието (ден, нощ, вечер, сутрин, в __ часа) части на годината (месеци)
(4) Not alive entities	All domain entities that are leaved after performing alive entities	
(5) number entities all entities that have numerical definition before them	a cat 15 cats Many cats	котка (котарак) 15 котки (15 котарака) Много котки (много котараци)
(6) Event entities	Quick, easy, hard, in the middle	Бързо, лестно, трудно, посреда

5. Designing of software architecture

Component diagram of the proposed software system is represented in the Figure 1. It illustrates the structure of the system and interaction between its components.

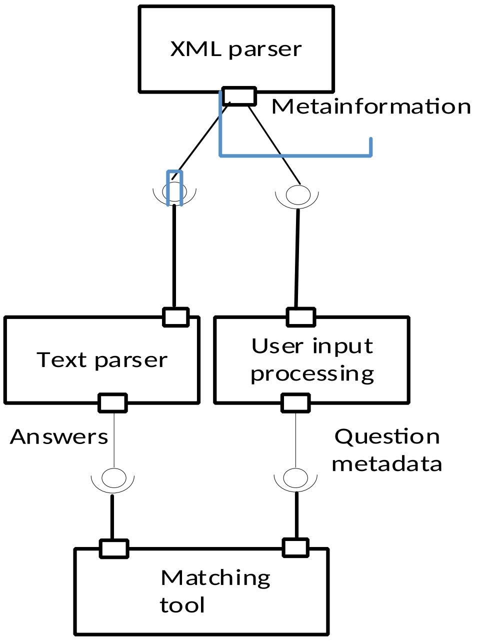

Figure 1: Component diagram of the software system "Intelligent search in texts' "

Metainformation about questions and answers for every language is stored in special XML file. Then, it is necessary to parse texts and questions. Information, extracted from XML parser, is transmitted to text parser and module for the questions processing (Use Input processing module in the Figure 1.).

After parsing texts answers according to the type of question are obtained. After parsing questions keywords are extracted (Table 3). After matching questions and information extracted from texts answers are proposed to user.

6. Preparing of attributes for text analysis

In order to represent connection between metainformation in texts and questions part of XML file is proposed. Structure of the file shows that in order to add new features of defining new answers in text it is necessary to modify items of list <MetaList> (See Table 2 and Table 3). In addition, it is easy to add, modify or remove new types of question.

```
<Bulgarian>
  <TypeQ> къде? </TypeQ>
     <MetaList>
            <Attribute>  на </Attribute>
            <Attribute> в </Attribute>
            <Attribute>  във </Attribute>
            <Attribute>  близо </Attribute>
            <Attribute>  далеко </Attribute>
            <Attribute>  под </Attribute>
            <Attribute>  над </Attribute>
     </MetaList>

         <TypeQ> кога? </TypeQ>
           <MetaList>
                <Attribute> в нощта </Attribute>
                <Attribute> нощ </Attribute>
                <Attribute>  деня </Attribute>
                <Attribute>  денят </Attribute>
                <Attribute>  сутрин </Attribute>
                <Attribute>  сутринта </Attribute>
                <Attribute>  днес </Attribute>
                <Attribute>  сега </Attribute>
                <Attribute>  никога </Attribute>
            </MetaList>
 </Bulgarian>
```

7. General description of the proposed approach

In order to use the proposed approach it is necessary to perform the next steps:

1. Domain analytics designs a domain ontology (message 1 in UML Sequence diagram on figure 2). Ontology may looks like a UML class diagram.
2. Using this ontology XML files with questions' and answers' attributes are prepared. In order to do this common XML files for specific language may be précised (see examples from chapter 6). Domain analytic gathers specific question attributes for ontology (message 2 on UML Sequence Diagram on figure 2), and then prepares attributes of answers (message 3 on UML Sequence Diagram on figure 2). Then XML files for different languages are prepared for the further using (messages 4.1, 4.2, and 4.3 on UML Sequence Diagram on figure 2).
3. Then domain analytics starts to test prepared XML files for the chosen text. For every question the next activities are performed

3.1 Using XML files attributes of question are extracted (messages 5.1 and 5.2 on UML Sequence Diagram on figure 2).

3.2 Using XML file for concrete language and concrete question and considering text a set of answers are founded (messages 6 on UML Sequence Diagram on figure 2).

4. Answers are returned to domain analytics for analysis (messages 7 on UML Sequence Diagram on figure 2)

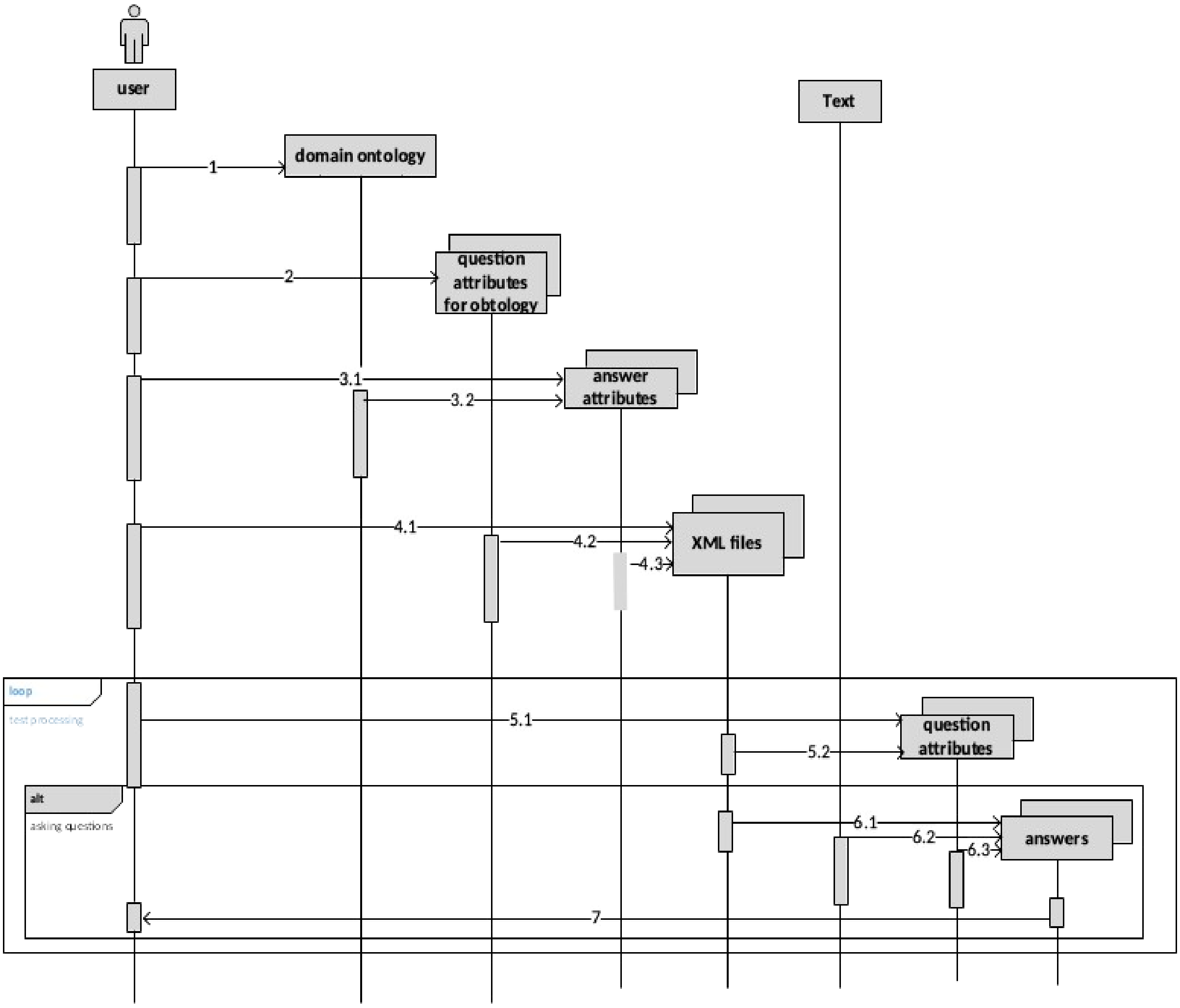

Figure 2: UML Sequence Diagram of the proposed approach for Multilanguage searching of information in texts

7. Case study

Consider a part of Bulgarian legend in two languages.

Bulgarian Сред живописните възвишения на Рила, редом до църковните храмове, жилищни комплекси и костници на Рилския манастир, своята снага издига и една сурова на вид сграда от масивни ломени камъни, наречена Хрельовата кула. Нейните дебели зидове не изглеждат никак на място в този храм, посветен на Бога и по-скоро оставят усещането за средновековно укрепление. А всъщност става въпрос именно за това.

Най-старата сграда от целия съвременен комплекс на Рилския манастир е построена от протосеваст Стефан Хрельо Драговол – благородник, който притежавал владения в региона.

От надпис, запазил се в самата кула, разбираме, че отбранителното съоръжение било издигнато от влиятелния властел през 1335 г., заедно с църквата „Св. Богородица Осеновица“

и било отдадено на защитата на богослужебните сгради, които тогава оформяли комплекса посветен на Св. Иван Рилски.

Малко се знае за вида, който е имал тогава Рилския манастир, тъй като пораженията нанесени от османското завоевание са довели както до загубването на изворови данни, така и до щети върху самите паметници на българската култура [5].

Translation into English Among the picturesque elevations of Rila, next to the churches, residential complexes and ossuaries of the Rila Monastery, a harsh-looking building made of massive rubble stones, called the Khrel tower, is also erected. Its thick walls do not look at all out of place in this temple dedicated to God and rather leave the feeling of a medieval fortification. And in fact, this is exactly what it is about.

The oldest building of the entire modern complex of the Rila Monastery was built by Protosevast Stefan Hrelio Dragovol - a nobleman who owned estates in the region.

From an inscription preserved in the tower itself, we understand that the defensive facility was erected by the influential ruler in 1335, together with the church "St. Bogoroditsa Osenovitsa" and was devoted to the protection of the religious buildings that then formed the complex dedicated to St. Ivan Rilski.

Little is known about the appearance that the Rila Monastery had then, since the defeats inflicted by the Ottoman conquest led to both the loss of source data and damage to the monuments of Bulgarian culture.

Table 4
Analysis of text for case study

Domain entities	
Бог,	God
Стефан Хрельо Драговол	Stefan Hrelio Dragovol
Иван Рилски	Ivan Rilski
Statistical characteristics of the text	
Words – 160	Words – 182
Total value of the text - 1076	Total value of the text - 1088

The next activity – is to take question about text. Tables with answers are given below.

Table 6
Question related to alive entities

Bulgarian	***English***
Кой е притежавал владения в региона?	Who was the owner of castle in region?
Possible variants of answers (all sentences with alive entities are considered)	
Стефан Хрельо Драговол – благородник, който притежавал владения в региона.	Stefan Hrelio Dragovol - a nobleman who owned estates in the region.
които тогава оформяли комплекса посветен на Св. Иван Рилски.	which then formed the complex dedicated to St. Ivan Rilski.
посветен на Бога и по-скоро оставят усещането за средновековно укрепление	dedicated to God and rather leave the impression of a medieval fortification
Statistical values	
Words – 28	Words – 61
Total value of text - 207	Total value of text - 266

Table 5
Question related to entities of time

Bulgarian	***English***
Кога отбранителното съоръжение е било издигнато?	When the defensive facility was built?
Possible variants of answers	
през 1335 г	in 1335
Statistical values	
Words – 3	Words – 2
Total value of text - 11	Total value of text - 7

Table 7
Other Questions

Bulgarian	***English***
Кои завоеватели са унищожили Рилския манастир?	What conquests destroyed the Rila Monastery?
Няма как да се намери отговор в текста по този метод	There is no way to find an answer in the text using this method

8. Description of software system realization

Detailed requirement specification, represented in the Table 8, describes main features of system described in this chapter.

Table 8
Challenges for the approach allowing to process text.

RQ code	RQ description	Priority
F0	Flexible extension of the system allowing to add new types of questions	10
F0.1	Develop and test classes for serialization and deserialization of Dictionaries [6].	
F1	Convenient representation of answers	10
F1.2	Prepare web-layer that visualizes results of searching.	
F2	Searching answers in different languages.	10
F2.1	Develop and text classes for parsing text files.	
F2.2	Develop approaches of searching metainformation in text that corresponds to the type of question.	
F3	Usability: the system should be easy to use and find the answer to the question.	10
NF3.1	Design data structures for storing Question and possible answers (<TypeQ> a <Metalist>)	
NF3.2	Use a data structure Dictionary for matching questions and answers in operating memory	

The development of project is started from the class allowing to save and restore XML file from hard disk. Storing and restoring is made by means of XML serialization. Class dataStore incapsulates the serialization and deserialization operations.

```
class dataStore {
    public string filename { get; set; }
public Dictionary<string,List<string>> quest_ans { get; set;}
    public void SerializeD() { }
    public void DeSerializeD() { }
}
```

The next class should support basic operations with datastore. (Something like CRUD operations when databases are processed.) Class ManageDataStore implements datastore processing operations.

```
  class ManageDataStore {
      dataStore ds { get; set; }
      public void DataStore_Create() { }
      public void DataStore_Edit() { }
      public void DataStore_View() { }
      public void DataStore_Delete() { }
  }
```

`Class Language is aimed to proceed operation of searching answers in texts.`

```
  class Language {
      dataStore ds { get; set; }
 public string text { get; set; }
 public List<string> questions { get; set; }
 public List<string> answers { get; set; }
      Language(string text, List<string> questions, string Lang) {
      }
      public void GetInformation() { }
      public void FindAnswewrs() { }
  }
```

9. Conclusion and estimation of the proposed approach

The paper proposes the multilingual approach for searching information in texts. Searching procedure matches attributes of the question and the text. The main idea of the proposed approach is to reduce amount of information needed to be processed by user in order to obtain answer to questions.

Table 9 shows the ratio of information needed to be proceed to find answer to question. Time ratio is calculated in the next way: average speed for reading such types of texts is 160 words in minute, It is assumed that reading will take 1 minute for Bulgarian language (Table 4) and 182/160=1.13 minute for English language.

Table 9
Time ratio for obtaining answers for different types of questions

Type of the question	Ratio of information (words in answer/all words in the text)		Time ratio	
Average information for one minute reading text (from example)				
	Bulgarian	***English***	***Bulgarian***	***English***
Question related to alive entities	207/1076	266/1088	3/160	2/182
Other Questions	-	-	-	-
Question related to entities of time	11/1076	7/1088	3/160	2/182
Average information for one minute reading text (another one-minute texts from [5])				
	Bulgarian	***English***	***Bulgarian***	***English***
Question related to alive entities	0.15	0,14	0.015	0,018
Question related to entities of time	0,012	0.014	0.009	0.010
Average information for ten minute reading text (other texts are taken from [5])				
	Bulgarian	***English***	***Bulgarian***	***English***
Question related to alive entities	0,22	0,28	0.11	0,11
Question related to entities of time	0.11	0.10	0.12	0.11

The analysis of the experimental results proves that time needed to process the answer to question is reduced nearly to ten times in comparison with reading the initial one minute texts.

Possible factors that may influence to effectiveness of searching the exact answers are:
(i) number and quality of questions' attributes. They are set manually by means of XML files processing. Thus, the qualification of expert may improve or reduce the quality of search operation;
(ii) Grammar errors in text, if they change the key attributes of answers.

The drawback of the approach is illustrated in the table 6. All questions of type "Who?" will give the same answer.

The defined drawback is basic for representing of the approach of the further research.

Further research: Development of the approach of questions' normalization. This approach will search the answers according to the questions' type and other keywords of the question. It is planned to add to practices of defining useful information from texts plug-ins or environment for automated verification of extracted facts from texts [7] as well as analysis of emotions [8].

10. References.

[1] B. Furlan, V. Batanović, B. Nikolić, Semantic similarity of short texts in languages with a deficient natural language processing support. Decision Support Systems 55(3), (2013). 710-719.

[2] G. T. Ngompé, S. Harispe, G. Zambrano, J. Montmain, S. Mussard, Detecting sections and entities in court decisions using HMM and CRF graphical models. Advances in Knowledge Discovery and Management Volume 8, (2019). 61-86.

[3] Q. A. Mahdi, R. Zhyvotovskyi S. Kravchenko, S. Borysov, I., Orlov, O. Panchenko, & S. Boholii, Development of a Method of Structural-Parametric Assessment of the Object State. Eastern-European Journal of Enterprise Technologies 5(4), (2021). 113.

[4] Z. Hu, A. Gizun, et al. Method of informational and psychological influence evaluation in social networks based on fuzzy logic, in: Proceedings of International Workshop on Control, Optimization and Analytical Processing of Social Networks. CEUR Workshop Proceedings, 2019, p.-2392.

[5] Българска история, 2020. URL: https://bulgarianhistory.org/hreliova-kula/

[6] Stack overflow. Serialization Dictionary, 2021
URL:https://stackoverflow.com/questions/14304034/serialize-dictionarystring-liststring-into-xml

[7] F. Andonov, V. Slavova, G. Petrov. On the open text summarizer. Information Content and Processing 3 (2016). 278-287 URL: http://www.foibg.com/ijicp/vol03/ijicp03-03-p05.pdf

[8] V. Slavova, Emotional valence coded in the phonemic content–Statistical evidence based on corpus analysis. Cybernetics and Information Technologies 2, (2020), 3-21.