UDC 004.415.2.045 (076.5) https://doi.org/10.15407/pp2022.03-04.281

AN APPROACH OF INTELLIGENT SEARCHING OF INFORMATION IN TEXTS

Olena Chebanyuk

Paper proposes an approach aimed at question oriented searching of information in texts. Texts are parsed, keywords and extra features of questions are marked, and sentences in text with the most relevant information to question are defined. Proposed approach is applied to Cyrillic and Latin languages.
Case study illustrates how to obtain answers to questions about Bulgarian fairytale that is represented on different languages (Bulgarian and English). Evaluation of the proposed approach is introduced. Description of the software architecture and source code of the corresponding software system are represented. Data structures and examples of *.xml files for storing information about question and answers are outlined.

Keywords: Text Processing on Different Languages, Grammar Rules, Semantic Analysis, Software System for Text Analysis, Software Architecture,

У роботі представлено методику пошуку інформації в текстах на запит. Тексти аналізуються, визначаються ключові слова та додаткові характеристики запитань, потім, у тексті шукаються речення з найбільш релевантною інформацією щодо запитання. Запропонована матодика застосовується до мов, що використовують як кирилицю так і латиницю. У прикладі показано, як отримати відповіді на запитання про болгарську казку, яка представлена різними мовами (болгарською та англійською). У роботі представлено оцінювання запропонованого підходу. Представлено опис архітектури програмного забезпечення та вихідного коду відповідної програмної системи. Описано структури даних та приклади *.xml файлів для зберігання інформації про запитання та відповіді.

Ключові слова: обробка текстів різними мовами, правила граматики, семантичний аналіз, програмна система аналізу тексстів, аріхтектура програмного забезпечення.

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 a lot of efforts to adopt software system for different purposes.

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

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 to perform 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].

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].

3. 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].

*C*hoosing 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].

Conclusion from the review and challenges for the approach

After analyzing the existing solutions for processing of texts, the following criteria for modern expert system aimed to process texts were defined:

1. Support of the functionality of processing information from various sources and preparing reports.
1.1. Speech to text input;
1.2. Processing text in different formats and encodings;
1.3. Recognizing texts from images.
2. Search the exact answer to the question.
2.1. Processing text with some grammar mistakes;
2.2. Processing texts containing smiles or special symbols;
2. Usability the system should be easy to use and find the answer to the question.
4. Convenient representation of answers.
4.1. Processing answers in text to speech modules
4.2. Representing answers in different languages

Proposed approach

Parse the question
Find keywords in questions
Define metainformation related to answers in text.
Search sentences that correspond to defined metainformation and represent them to user.

Metainformation about keywords for different types of questions

The list of the proposed characteristics for questions is given below. Aim of this classification is to take a marks in text according to types of questions.

Table 1. List of metainformation that is extracted from questions

Metainformation for answers in questions	Question words in Bulgarian	Question words in English
(1) alive entities	кой?	who?
(2) entities of place from domain. They 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) all entities that have numeric precision	колко?	how many(much)?)
(6) all event entities	как?	How?

Metainformation about features of answers extracted from text

It is proposed to classify domain entities according to defined characteristics. One entity may correspond to several classes.

Table 2. List of metainformation that is extracted from texts

Description of metainformation	Examples for English language	Examples for Bulgarian language
(1) All alive entities from domain description	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, 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	15 cats	15 котки
(6) Event entities	Quick, easy, hard, in the middle	Бързо, лестно, трудно, посреда

Case study

Consider a part of Bulgarian fairytale about three apples.

Table 3. Text for case study

Fairytale in Bulgarian	Bulgarian Fairytale in English
Една жена имала трима синове. В градината на къщата им растяло чудно красиво ябълково дърво. Всяка година то раждало само по една ябълка, но не каква да е, а златна. Ала в нощта, когато ябълката узрявала и така заблестявала между клоните, че цялата градина грейвала, долитала една хала и откъсвала златната ябълка. Една година, щом дошло време ябълката да узрее, най-големият син рекъл на майка си: — Мале, ще отида да вардя ябълката. Дай ми един нож и орехи, та да не заспя. Седнал най-големият син под ябълката и започнал да троши орехи. Изведнъж задухал силен вятър и дърветата се превили чак до земята. Тъмен облак закрил луната и звездите, а халата слязла от небето, грабнала златната ябълка и докато големият син се усети, отлетяла. На другата година средният син казал на майка си: — Мале, отивам да вардя ябълката. Дай ми един нож и орехи, че тръгвам. Седнал той под ябълката, ала се улисал да троши и да яде орехи и така и не разбрал как халата откъснала златната ябълка и изчезнала с нея.	A woman had three sons. In the garden of their house grew a wonderfully beautiful apple tree. Every year it bore only one apple, but not just any apple, but a golden one. But on the night when the apple ripened and shone so brightly between the branches that the whole garden glowed, a challah flew by and plucked the golden apple. One year, when the time came for the apple to ripen, the eldest son said to his mother: «Mom, I'm going to take the apple.» Give me a knife and walnuts so I don't fall asleep. The eldest son sat under the apple and began to crush walnuts. Suddenly, a strong wind blew and the trees were bent all the way to the ground. A dark cloud covered the moon and the stars, came down from heaven, grabbed the golden apple and before the eldest son felt it, flew away. The next year the middle son said to his mother: «Mom, I'm going to boil the apple.» Give me a knife and walnuts, that I'm leaving. He sat under the apple, but he enjoyed crushing and eating walnuts and never understood how the robe tore off the golden apple and disappeared with it.
Text is taken from https://bulgarianhistory.org/trimata-bratq-i-zlatnata-qbalka/	

Table 4. Analysis of text for case study

Domain entities	
Ябълка, хала, орехи, средният син, най-големият син, майка, нож.	Apple, challah, walnuts, middle son, eldest son, mother, knife
Statistical characteristics of the text	
Words – 173 Total value of the text - 990	Words – 205 Total value of the text - 1066

The next activity – is to take question about text. Matched metainformation about questions and answers are marked by different colors (the same color for the same type of question).

Table with answers is given below.

Table 5. Questions and answers (entities of time)

Question related to entities of time	
Кога е узряла ябълката?	When the apple was riped?
Possible variants of answers	
Ала в нощта На другата година средният син казал на майка си:	But on the night The next year the middle son said to his mother:
Statistical values	
Words – 12 Total value of text - 60	Words – 14 Total value of text - 64
Кога най-големият син е чакал ябълката?	When did the eldest son wait for the apple?
Possible variants of answers	
Ала в нощта На другата година средният син казал на майка си:	But on the night The next year the middle son said to his mother:
Statistical values	
Words – 12 Total value of text - 60	Words – 14 Total value of text - 64

Table 6. Questions and answers (alive entities)

Question related to alive entities	
Кой е откраднал ябълката?	Who stole the apple?
Possible variants of answers (all sentences with alive entities are considered)	
долитала една хала и откъсвала златната ябълка най-големият син рекъл на майка си: Седнал най-големият син под ябълката и започнал да троши орехи. На другата година средният син казал на майка си: ала се улисал да троши и да яде орехи и така и не разбрал как халата откъснала златната ябълка и изчезнала с нея.	a challah flew by and plucked the golden apple the eldest son said to his mother: The eldest son sat under the apple and began to crush walnuts. The next year the middle son said to his mother: but he delighted in crushing and eating walnuts and never understood how the robe tore off the golden apple and disappeared with it.
Statistical values	
Words – 55 Total value of text - 306	Words – 61 Total value of text - 266
Other Question	
Какво е ял най-големият син?	What the eldest son eat?
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

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.

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 2). After matching questions and information extracted from texts answers are proposed to user.

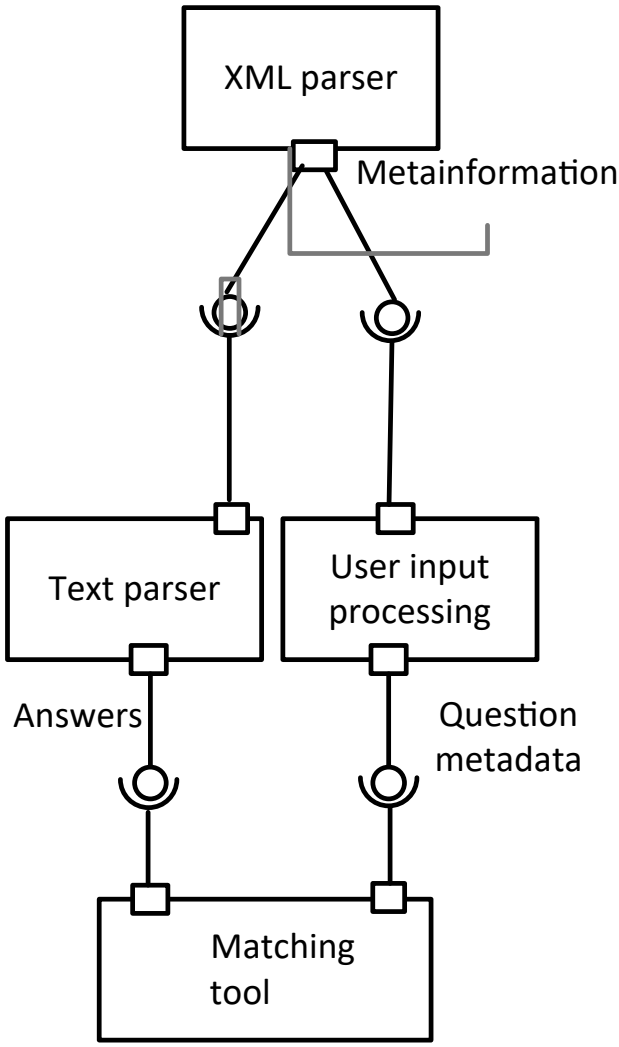

Figure 1. Component diagram of the software system “Intelligent search in texts’ ”

Preparing of metainformation 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 1 and Table 2). Also it is easy to add or remove new types of question.

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

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

Description of software system realization

In order to realize such a software system it is necessary to solve the next task

Select data structures fro storing <TypeQ> with <Metalist>, referencies to text? Questiona and possible answers. – Data structure Dictionary is selected.

Develop and test classes for serialization and deserialization of Dictionaries [6].

Develop and text classes for parsing text files.

Develop approaches of searching metainformation in text that corresponds to the type of question.

Prepare web-layer that visualizes results of searching.

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() { }
}
```

Conclusion

Paper proposes the approach of searching information in texts. Searching procedure matches type of the question and the specific metainformation in text. Experimental results show that the proposed approach has the next advantages:

It allows reducing amount of text to be proceeded to find an answer to the question.

Approach works with the same effectiveness for different languages.

Key features of the approach:

It is quick and easy for realization of the software system.

It allows extending metainformation both about new types of questions and about specific metainformation in the texts without modification of code.

It is realized for easily adding of Latinic and Cyrillic languages. In order to add new language it is not necessary to modify code. Only new XML files with key features about questions and metainformation for answers in texts must be added.

Drawback of the approach: as it is the first version of the approach it allows to find the same answer for different questions of the same type (see case study). This paper starts the cycle ow works, representing results of intelligent texts' search.

The defined drawback is basic for representing the ***further research:***

Development of the approach of questions' normalization. This approach will search answers according to questions' type and other keywords in the text of the question. For example, questions "When did the eldest son wait for the apple?" and "When the apple was riped?" receive the same answers (see case study). Normalization approach will allow avoiding this drawback.

References

1. Furlan, B., Batanović, V., Nikolić, B. Semantic similarity of short texts in languages with a deficient natural language processing support. Decision Support Systems, 2013. 55(3), 710-719.
2. Ngompé, G. T., Harispe, S., Zambrano, G., Montmain, J., & Mussard, S. Detecting sections and entities in court decisions using HMM and CRF graphical models. In Advances in Knowledge Discovery and Management 2019.
3. Mahdi, Q. A., Zhyvotovskyi, R., Kravchenko, S., Borysov, I., Orlov, O., Panchenko, I., & Boholii, S. Development of a Method of Structural-Parametric Assessment of the Object State. Eastern-European Journal of Enterprise Technologies, 2021. 5(4), 113.1-86. Springer, Cham.
4. Gizun, A., Hriha, V., Roshchuk, M., Yevchenko, Y., & Hu, Z. Method of informational and psychological influence evaluation in social networks based on fuzzy logic. Paper presented at the CEUR Workshop Proceedings, 2019., 2392
5. Text for case study https://bulgarianhistory.org/trimata-bratq-i-zlatnata-qbalka/
6. Dictionary serialization https://stackoverflow.com/questions/14304034/serialize-dictionarystring-liststring-into-xml

Bibliography

1. Furlan, B., Batanović, V., Nikolić, B. Semantic similarity of short texts in languages with a deficient natural language processing support. Decision Support Systems, 2013. 55(3), 710-719.
2. Ngompé, G. T., Harispe, S., Zambrano, G., Montmain, J., & Mussard, S. Detecting sections and entities in court decisions using HMM and CRF graphical models. In Advances in Knowledge Discovery and Management 2019.
3. Mahdi, Q. A., Zhyvotovskyi, R., Kravchenko, S., Borysov, I., Orlov, O., Panchenko, I., & Boholii, S. Development of a Method of Structural-Parametric Assessment of the Object State. Eastern-European Journal of Enterprise Technologies, 2021. 5(4), 113.1-86. Springer, Cham.
4. Gizun, A., Hriha, V., Roshchuk, M., Yevchenko, Y., & Hu, Z. Method of informational and psychological influence evaluation in social networks based on fuzzy logic. Paper presented at the CEUR Workshop Proceedings, 2019., 2392
5. Текс для експерименту https://bulgarianhistory.org/trimata-bratq-i-zlatnata-qbalka/
6. Серіалізація словника https://stackoverflow.com/questions/14304034/serialize-dictionarystring-liststring-into-xml

Received 04.08.2022

About the authors:

Olena Chebanyuk
Doctor of Sciences, Assoc. professor
Glushkov Institute of Cybernetics, department 205,
senior researcher,
Approximately 25 Ukrainian publications,
Approximately 60 International publications,
H-index: Google Scholar – 6,
Scopus – 2,
0000-0002-9873-6010.

Place of work:

National Aviation University,
Software engineering Department,
Lubomir Guzar ave 1
Telegram @tocarelcielo

Glushkov Institute of Cybernetics of
National Academy of Sciences of Ukraine,
40 Glushkov ave., Kyiv, Ukraine, 03187,
tel.: (+38) (044) 526 3348
email: Chebanyuk.olena@gmail.com

Прізвище та ім'я автора і назва доповіді англійською мовою:
Chebanuyk O.V.
An approach of intelligent searching of information in texts

Прізвище та ім'я автора і назва доповіді українською мовою:
Чебанюк О. В.
Методика аналітичного пошуку інформації у текстах

Контакти для редактора: Чебанюк Олена Вікторівна,
старший науковий співробітник відділ 205 інститут кібернетики
ім В.М. Глушкова НАН України,
e-mail: Chebanyuk.olena@gmail.com
тел.: телеграм @tocarelcielo вайбер 091-619-54-24