Abstract: The paper describes the design of the closed-domain question answering system. Key properties of the system are: (i) ability to answer questions in natural (Slovene) language, (ii) ability to employ structured as well as unstructured data sources, (iii) ability to automatically form questions and answers, (iv) ability to conduct a simple dialog with the purpose of retrieving user-specific information, (v) employability in new domains, and (vi) ability to answer questions referring to its own functioning as well as to react on the aggressive questions asked by its users.

Key-Words: question answering, closed-domain system, Slovenian language, morphological dictionary of Slovenian language, web services, ontology, XML

1 Introduction

The basic idea of question answering systems is to be able to provide answers to questions written in natural language. The answers can be retrieved from different sources, e.g. web pages, plain texts, knowledge bases, web services etc. Unlike the information retrieval applications like web search engines that flood their users with documents or best-matching passages, the goal of question answering systems is to find a specific answer. There are many ways of looking at question answering, and they depend on the approaches towards various dimensions [6]. The aforementioned dimensions are: question, response, technique, information source, domain, and evaluation. Currently the most extensive source of information, which is also used by question answering systems, is the World Wide Web. The World Wide Web was not designed for mere communication purposes but to contain information, therefore the idea that computers should also be capable of collaboration and automatic task management arose. However, the majority of information available on the web is suitable only for human use. Even the template based documents, e.g. the documents in different data bases, of which structure and meaning are defined, do not alleviate the work for programme agents [1]. There was a need for a change, and the idea of the Semantic Web was born [2], [9]. The Semantic Web is an extension of the World Wide Web. In the Semantic Web information is defined with unambiguous computer-understandable metadata, thus enabling computers and people to work in cooperation [2]. The next generation of natural language question answering systems will use the structure and elements of the Semantic Web in order to improve the web searches [10].

Students at our faculty are faced with various questions during their studies. In their search for answers, they browse the web portal, or answers are provided by the employees of the faculty via e-mail, forum, or telephone. In order to improve the retrieval of information for students and alleviate the burden on the employees, a closed-domain question answering system had to be developed. Since no solutions for our target domain existed, our system was the first step towards the introduction of question answering systems in education. It was one of the first question answering systems in the Slovene language.

The article is divided into four chapters. The following chapter describes the definition of the approaches towards the dimensions of the system. The third chapter describes the implementation of the system and the conclusion the summary and some suggestions for our future work.

2 Definition of the approaches towards the dimension of the system

The approach towards the dimensions of the system was determined based on the analysis of the questions which were asked on our faculty’s forum by its users. The dimensions are the following:

- Evaluation: The system will be evaluated according to the percentage of accurate answers.
• Domain: It will be a closed-domain system. The system’s structure will be general, so it will be employable also in other domains.

• Information source: The system will retrieve answers from the following sources: local database, databases of the faculty’s information system, MS Excel files, and web services.

• Technique: The system’s search for answers will be based on the principle of full matching or partial matching (at least in the minimum number of required keywords) of a question.

• Answer: According to the research described in [3] and our own judgment, we decided that the answers will not be exact (a word or a word phrase), but they will be in the form of a sentence or a paragraph. The answer in the form of a sentence will not contain any additional information that could be of interest to the user in his/her future questions, while the answer in the form of a paragraph will. If required, the answers will be accompanied by hyperlinks to web pages or documents containing additional information.

• Question: The system will provide answers to the following types of questions: insulting questions, ex. “You are stupid!”, special questions and phrases, ex. “What is your name?”, questions referring to the employees of the faculty, ex. “Where is professor Johnson’s office?”, questions which consist of the data stored in local databases and MS Excel files, ex. “What is the amount of the description fee?”, and other questions, ex. “When are the office hours of the student office?”. According to the analysis, the system should also have the following properties:

• The system will be capable of forming a simple dialog with the user in order to retrieve information that could improve its answers.

• The system will save user-specific information and information on the activity of users in order to deliver a user-specific answer.

• The system will provide answers to the questions which are written in standard Slovene.

• The system will carry a female name Sara in order to stimulate the use of the natural language.

We developed two applications: one for users and one for the administrators of the system. With the help of the latter application the administrators expand the database of questions and answers, import the data from local databases and MS Excel files, integrate web services into the system, and supervise the functioning and use of entire system. The former application is actually a system that answers the questions.

3 Implementation of the system

3.1 Application for administrators

As previously mentioned, one of the searching techniques is based on the principle of partial matching. The question in the database and the user’s question are similar when they contain the highest possible number of the same keywords. A keyword is not necessarily just one word, but it can also be a series of words. Keywords must be in their lemmatised form in order to be successfully compared.

In order to find lexical forms, words must undergo the lemmatization procedure which is highly complicated when dealing with the inflectional languages like Slovene [4], [7]. The morphological dictionary of the Slovene language, which contains approximately 8000000 declination and conjugation forms of words and 300000 lemmas, was our source of the lemmatised forms of words. Despite the large number of words, the morphological dictionary does not contain all words that occur in the domain of the faculty, therefore it was decided that a closed-domain dictionary has to be built.

The domain-specific dictionary that was built does not contain only the words that are missing in the morphological dictionary, but it contains all words that occur in the above mentioned domain. All words in the domain-specific dictionary were indicated as domain-specific words that manifest themselves as keywords in questions. Words that are not in the morphological dictionary can be added to the domain-specific dictionary in whatever form with the purpose to of expanding it. Regardless of their form, series of words can be added to the dictionary. The connection between the morphological and the closed-domain dictionary enables the retrieval of all forms of series of words. This is possible for those series of words of which words occur in the morphological dictionary. The domain-specific dictionary also contains the information on the meaning of words or series of words as well as the information on their relationship. The meaning of words tells the difference between different persons, organizations, objects, etc. The synonyms, e.g. degree dissertation – diploma thesis, and acronyms, e.g. ECTS - European Credit Transfer System are very frequent. The relationships between words enable the
transformation of questions, e.g. “What is ECTS?” into “What is the European Credit Transfer System?” and vice versa. The keywords of questions that are in the database or have been imported from the columns of the tables in local databases or MS Excel files provide the source of words for the domain-specific dictionary.

### 3.1 Application for users

Prior to the implementation of the application for users, we had to determine the most important properties of the system. The most important properties are:

- **Reliability.** The system must provide an answer to the exact question. In order to provide the correct answer, it has to be capable of assessing whether or not there is enough information available to answer a particular question. If there is not enough information, the system must react in the most appropriate way. For example, the system is obliged to inform the user about the lack of information, or that it is not capable of providing an answer.

- **Response time.** If the system is able to provide the answer faster than a user can find it by clicking on the web portal or with the help of a search engine, its existence is justifiable. The response time, however, does not depend merely on the system itself but also on the user. It is of the utmost importance that a user is capable of requesting certain information in the appropriate way.

The goal of the project was to develop a fast and reliable system. Unfortunately, these two properties are contradictory. In order to improve the accuracy of the answers, a more complicated search algorithm which prolongs the response time of the system is required. On the other hand, in order to shorten the response time of the system, the search algorithm must be simpler, thus leading to the inaccuracy of answers. Higher priority was given to the accuracy of answers. The response time, however, should still be as short as possible; therefore we sought to achieve the optimal combination of both properties.

As already mentioned, the system responds to certain types of questions: insulting questions, special questions or phrases, questions referring to the employees of the faculty, questions which consist of the data stored in local databases and MS Excel files, and other questions.

Such classification was chosen due to the information sources and approaches that are essential for the processing of questions.

In order to reach the optimal accuracy and speed of the system, the algorithm processes questions according to the above sequence of question types, starting with the insulting questions. This enables extremely short response times in case of any occurrence of insulting questions or phrases. One part of the question processing procedure is the same for all questions. First of all, all scripts and HTML elements are eliminated in order to disable any security breaches via input form of the application for users [8]. Punctuation marks, special characters and emoticons are eliminated as well, since these elements of language do not influence the quality of answers at this point of the development of the system.

Before the system starts to examine if the question that was asked belongs to a certain type of questions written above, it determines if the question has already been asked in the past and if the correct answer was provided. If the system retrieves the answer in the described way, the further search is terminated, otherwise it continues.

First of all, the system determines whether the question is reasonable or not; therefore a simple algorithm is required. The pseudo code of the algorithm can be seen in picture 1.

If the question is reasonable, the system determines to which category of questions it belongs. The system starts with the category of insulting questions. Every word in a question is examined and compared to the list of the insulting words which was built in order to be used by the application. As soon as an insulting word is found, the system terminates the processing of the question and informs the user of its inappropriateness.

If the system determines that a question does not belong to the category of insulting questions, it tries to determine if it belongs to the category of special questions. Answers to special questions are retrieved with the help of full matching (the question matches all characters). If a question meets the criteria of a special question, it must be determined whether there are more possible answers to that question. If there are, the system chooses one of the possible answers and finishes the search. The variety of possible answers should increase the interest of users and improve their confidence in the system.

If a question does not belong to the category of special questions, the system continues to search for an answer. In the next step, the system determines which words in the question are keywords. A special algorithm is required in order to determine them. Before the description of the algorithm one must bear in mind that a keyword is not necessarily one word, but it can also be a series of words.
The representation of the algorithm is based on the question “When is the exam period?” The words in the question are: when, is, the, exam, period. In the beginning, the system processes the whole series of words, as it would be one keyword. If the series of words represent a keyword, the system memorizes it and continues with the search. In the next step, the system eliminates the last word in the sequence and determines whether the new series of words represent a keyword. If it does, the system memorizes it and continues with the search. The system repeats the procedure until there are no more words left in the series. In the described example all the keywords are written in bold.

1. complete series: when is the exam period
   1. keyword: when is the exam period
   2. keyword: when is the
   3. keyword: when
   4. keyword: when

Because none of the above words can be found in the domain-specific dictionary, none is written in bold.

As the process continues, the first word in the original series of words is eliminated. The new series of words is then processed in the same way as the original.

2. complete series without the first word: is the exam period
   1. keyword: is the exam period
   2. keyword: is the
   3. keyword: is

Because Slovene proper names that end in the same way are declined according to the same declination pattern, they could be divided into three categories: feminine names, masculine names, and last names. Later, the declination rules for individual groups of names were determined and implemented into the system. All possible declination forms of the names of the faculty’s employees are currently stored in two tables. The first one contains the standard (lemmatized), the second all the declination forms of names. The number of employees, however, can change on a daily basis, therefore all three aforementioned tables are daily updated.

An example will illustrate the functioning of the system. If a question contains the name of a faculty employee and this name is not one of the names Rok, Samo, Rado and Avgust, the system starts searching for an additional keyword in order to determine whether the user asked for specific information on a certain employee. If such a keyword is found, the specific information on a certain employee is presented to the user in the first sentence of the answer, and the following...
sentences contain general information on the employee. If such a keyword cannot be found, the answer contains only general information on the employee which is represented to the user in the default sequence.

Slovene proper names Rok, Samo, Avgust, and Rado must be processed separately. The aforementioned Slovene words can carry two meanings. They can be understood either as proper names, or they can carry a general meaning. As words that carry a general meaning they are also a part of the morphological dictionary. Consequently, the system has to determine whether a certain question refers to a person or not. The following example serves as an illustration. The word avgust in Slovene can refer either to a person (a proper name) or to a month. The questions “Where can I find Avgust?” and “When are the office hours of the student office in August?” (Kdaj so avgusta uradne ure referata?) contain the same word which in the first question refers to a person and in the second to a month. Based on the context people are able to distinguish between the two referring objects instantly. For computers, however, such deduction is a complicated task which they have to learn. Because the system in its current phase of development was not able to deduct in such a way the necessary mechanism was not planned even in the beginning a temporary solution had to be found. If one takes a closer look at the above examples, one might think that the solution lies in the capital letter. If the word avgust refers to a person, it must capitalized, however, it is not realistic to expect that users will use the capital letter according to the rules of grammar, therefore it is not an appropriate solution. An appropriate solution is to require additional information. For example, if a user asks “Where is Avgust?” or “Where can I find Avgust?”, the system inquires whose Avgust the user might have in mind. Otherwise, the system considers the word to be a general expression (it refers to a month) and continues to search for an answer among the general questions. The same procedure applies for the other above mentioned proper names.

The following section of the article describes the retrieval of answers which use data from local databases including the databases of which content was transferred from MS Excel files. The questions are described with the help of question templates that cover the conceptual model of the database and describe the concepts, their attributes, and the relationships [5]. A question template is actually a question with empty slots for data instances that represent the main concepts of the question. For example, “Where is <place>?” is a question template where <place> represents an empty slot that belongs to the concept of place. If this slot is filled up with data instances that belong to the concept, an ordinary question, e.g. “Where is the Gama lecture hall?”, is formed. The templates are written in the form of an XML file [11]. An example can be seen in picture 2.

```xml
<?xml version="1.0" encoding="utf-8"?>
<listOfQuestionsAndAnswers>
  <questionAnswer>
    <question>"Where [synonym] [Places_Place]"
    <answer>"[Places_Location]"
    <synonym>"it is located, is, can be found"
    <case>"r;"
  </questionAnswer>
  ...
</listOfQuestionsAndAnswers>
```

Fig. 2: An example of the XML template for questions that can be answered

Question template in the XML file is written as an attribute of the element questionAnswer. Empty slots that can be filled up with the examples of concepts are marked with square brackets “[]”. In order to use the same template also for questions that refer to the same specific concept but have a different form, round brackets were introduced (“”). Round brackets represent an empty slot that can be filled up with the attribute values of a synonym. In this way the capability of a template is increased. Each template can have more empty slots for concepts and attribute values of synonyms. In this case the values of individual synonym slots are separated with the semi colon “;”, and the capability of a template is increased even further. The number of questions that can be formed on the basis of one template is calculated with the help of the following equation:

\[
N = \prod_{k=1}^{n} k \times \prod_{s=1}^{m} s
\]

in which \(k\) is the number of all instances of an individual concept, and \(s\) is the number of all instances that can fill up individual slots of synonyms.

In order to form an answer, the system also uses templates that are written in the attribute answer in the above XML file. In order to retrieve an answer, the system forms all possible questions based on the individual descriptions from the XML. The original question is then compared to the series of potential
questions. Because Slovene is an inflectional language, questions must be lemmatised before the comparison. If two lemmatised questions match, the answer is formed on the basis of the description in the XML and later presented to the user. The formation of an answer is again a difficult task, especially due to the inflectional Slovene language. Empty slots cannot be simply filled up with the data (series of words) from the database, but they must be in the correct form. The correct form of the series which is based on the value of the attribute “case” is provided by the morphological dictionary of the Slovene language. It has to be taken into consideration that not all forms of words in the series are changed. It is difficult to determine which word in the series changes. Answers that refer to people must also be written in the appropriate form considering the gender of the person, to which they refer. This means that the endings of certain words in a sentence have to change, and it is difficult to automatically determine which ones. The above mentioned difficulties represent a challenge for our future work. If the system finds the answer in such a way, the further search is terminated, otherwise it is continued.

In the last step, the system processes the remaining questions. The main idea of question processing is to gradually decrease the number of potential questions questions of which answers could be the answer to the user’s question in order to find a question that is most similar to the user’s question.

In our case those are the questions that contain at least one of the keywords in the user’s question. The system memorizes their unique identifiers. In order to speed up the system, the keywords of questions in the database are prepared in advance.

During the process, it is determined how many keywords are necessary in order to define a particular question from the database as a potential question. If the user’s question contains one or two keywords, a potential question must also contain one or two keywords. The questions must be a 100% match considering the keywords.

The reason the questions must be of the same number of keywords will be presented in the following example. In the question “When does the exam start?” there are two keywords: exam and start. Because the word exam is very frequent in the domain of the faculty, the system will find 150 potential questions containing the keyword exam and 45 questions containing the keyword start. It will have to process 195 questions or 194 questions if it is presumed that there is a question which contains both keywords. If it is also taken into consideration that all keywords in questions with a small number of keywords are of greater importance than the keywords in questions with a large number of keywords, both keywords are of the utmost importance. If in the case of two keywords we demand that a potential question must also contain both keywords, there is only one question left in the database for the system to process. In such a way, the number of potential questions and the search time are considerably diminished.

The importance of keywords diminishes with their number; therefore the matching percentage of keywords in questions that contain three or more keywords was lowered to 75%. For example, if the original question has four keywords, the potential questions from the database must have at least three matching keywords. In this case three keywords are enough to determine that two questions are similar. The one keyword that differs allows the user to form a question in different ways or include redundant information. Despite that fact, the system still answers correctly. From all potential questions the system selects only those that are at least a 75% match of the original question considering the keywords. The system then calculates the matching percentage for each appropriate question and determines the highest. All questions with the highest matching percentage are later processed by the system.

For each potential question, the difference between the number of keywords in a potential question and the user’s question is also calculated, and the minimum is determined. Questions that are below the minimum are eliminated. If there is only one potential question left, the answer to that question is presented to the user. If there are more potential questions, the question forms of the potential question and the user’s question are compared. For example, the questions “Who is the dean?” and “Where is the dean?” differ only in their question forms. If the question form of a potential question corresponds to the question form of the users question, the answer to the potential question in the database is delivered to the user. If the question forms do not match, the system selects another potential question and compares the question forms anew. This procedure is repeated until the last question but one. If also those question forms do not match, the answer to the last potential question is delivered to the user. The answers can contain also one or more hyperlinks to web pages or documents with additional information. The first hyperlink opens in a parent window. The parent window is the window in which the application for users was started; the user application itself runs as a popup window.
If an answer cannot be found, or if the question does not contain enough keywords, the system asks the user to paraphrase it and provide additional information; otherwise the system forwards the keywords to the search engine on the web portal which then tries to find the documents containing these keywords.

4 Conclusion
The article describes the Slovene language question answering system. The system retrieves the answers from structured and unstructured information sources. A small portion of answers is automatically formed on the basis of templates. The system is capable of forming a simple dialog with the user for purposes of information retrieval that enables better answers.

The system can also provide answers to the questions referring to the functioning of the system and respond to aggressive behaviour of users. The system is in use on the web portal of the faculty. The development of this system has enabled faster access to information for students and alleviated the burden on the employees that were answering students’ question in the past. A new method of information retrieval, question answering, has been presented in our target domain. The use of the Slovene language in the system is highly innovative. The fact that the Slovene language is an inflectional language is reflected in the difficulty of its processing. During the development of the system, some solutions that are necessary for the processing of Slovene language have been developed. Some solutions, however, were only foreseen due to their extensiveness and represent a challenge for future work.

References: