Implementation and assessment of the automatic question generation module

László Bednarik
Department of Information Technology
University of Miskolc
Miskolc-Egyetemváros, Hungary
laszlobednarik@gmail.com

László Kovács
Department of Information Technology
University of Miskolc
Miskolc-Egyetemváros, Hungary
kovacs@iit.uni-miskolc.hu

***Abstract*— Automated Question Generation is a key component of intelligent tutorial systems. This system combines several distinct tools from very different areas of information technology, among other clustering and classification units. The developed prototype system was tested in a real application with high-school students. The performed tests show that the automatically generated questions are as good as the manually constructed tests.**

Keywords—automated question generation; classification; e-learning;clustering

I. INTRODUCTION

Natural language processing is a more and more popular area of artificial intelligence and it has a wide application area like robot controlling. The combination of these two areas improves significantly the functionality of robotic assistants which became more common in environments such as offices, houses and industries. Robots often need to communicate with users much easier and natural ways like via spoken dialogs.

The natural language interface module usually includes three components: interpretation of human commands (speech to text and text to command) ; the generation of response sentences (text to speech) and a dialog module. The key role of the dialog module is to manage the sequence of question - answer pairs. The task includes among others the generation of new questions for the current context.

An important subtask of the dialog system is the generation of questions. Automatic question generation (AQG) could be used in different natural language processing systems and in many text content management systems. Thus, also e-learning frameworks can host an AQG module to increase the flexibility of the assessment module. In important component within the AQG is the clustering and classification engine [1] to gain all the information for selection of the best suitable expressions for question generation. Through the developed and implemented methods, the words of source text are clustered into partitions based on the algorithm parameters. As a result of clustering words within a cluster satisfy a distance criterion defined in advance. That means that for each cluster, the distance of any two words belonging to the same cluster is less than a threshold value given as an external parameter.

The word to be distracted from the sentence for a question generation is determined by the classification method [2]. The corresponding output cell of the applied neural net is called 'whether to be distracted for a question' has a binary value-set. If a word from the target document is selected for question by the classification model, then this subjective coordinate gets a true value. In this case, the rest words of the sentence yield a false value for this output item. The applied neural network classification can be used to rank several words judged to be distracted and to fitting the best for distraction according to the ranking.

The paper describes principles and methods applied to determine automatically the word to be distracted form a chosen sentence for the question generation according to information gained by methods mentioned above. In the paper we analyse how the automatic question generation pattern works in practice and we assess the results.

II. THE AUTOMATIC QUESTION GENERATION

In the developed system the following main parameters influencing the operation of the automatic question generation were used as key external parameters [9][10]:

ε: the maximum distance of words inside a cluster,
d: the function describing the distance of two words,
N: the number of words to be clustered,
A: the annotation of sentences to be analysed,
M: the annotated sentences of the document,
W_e: the set of words to be analysed,
W_t: the set of words used for teaching the neural net ($W_t \subseteq W_e$),
K_o: the objective coordinates of the words to be analysed,
K_{sz}: the subjective coordinates of words to be analysed,
α: the weigh modification rate performed in one step at the edges of the neural net,
m: the number of neurons in the input layer,
n: the number of neurons in the hidden layer,

978-1-4673-5188-1/12/$31.00 ©2012 IEEE

k: the number of neurons in the output layer,
g: the function describing the criterion for stopping the learning process of the neural net,
p: the function describing the criterion for stopping the learning process of the neural net,
H: concept hierarchy dictionary.

Through generating questions automatically our aim was to create algorithms which are able to carry out clustering or classification tasks according to set parameters in the case of documents containing several thousands of sentences. According to our experience the learning material we wish to ask questions about can mostly be characterised by this amount of sentences. However, neither the time taken to answer the questions nor the character of the document allow of generating a question for each sentence of the document. Therefore it was necessary to restrict the number of questions to be asked according to rational principles.

At first for this task defining a parameter with which the number of words to be randomly selected from a complete input document can be given provided a solution. Afterwards question were generated only for these randomly selected sentences. Through refining the first approximation method we achieved that the expert controlling the generation of questions can determine the type of sentences to be put to the question generation module in a separate parameter. In almost each area of the application of the developed algorithm such a selectiveness of documents of mixed type is required.

The required solution was gained by annotating sentences in the document. According to this each sentence in the document which can be suitable to distract for question generation was annotated by a special character set describing the type of a sentence [3]. As a further possibility, it was worked out that during setting the type of sentences to be distracted for questions not only one specific type of sentence can be selected. The algorithm is suitable for defining filter conditions containing several annotations and searching according to this. Hence it can be achieved to select all the sentences annotated in the document for question generation. By applying annotations sentences of the documents which are not annotated can be completely excluded from the questions.

After annotating sentences of the document the document file is created which can be processed by the algorithm worked out during our research work. This yields one of the inputs of the question generation algorithm. For the expected running the file defining the kind of words and the stem of words for each word of the document – in the succession of their appearance - is needed. This information is needed when generating automatically the possible answers to be given to words distracted from sentences for question generation. This information was gained with the help of a freely available stemming application [4]. Results of this stemming application are available in a text file separated by semicolons that gives the other input of the developed algorithm. This represents the stemmed file.

After determining the stemmed file the algorithm analyses words of the stemmed and source files one by one and when it finds identical words it links them. As in the stemmed file each word is in a separate line therefore separating the text into words is easy to do. In the case of the document file, however, for this task the program searches the symbols separating words in the text. Words of the document are separated from each other according to these symbols. After separating the document file into words both files can be considered as a succession of words. Proceeding parallel in the two lists the algorithm has to find the pair of the word in the document file in the stemmed file [12].

Each sentence of the source document is processed, even those, which not satisfies the given annotation condition. In the next phase of the algorithm, the input sentences are analysed one by one to check if they satisfy the filtering criterion defined by annotations or not. Sentences which satisfy one of the annotations containing the filtering criterion are linked to the list of sentences used for the question generation. Afterwards it deletes the annotation from them so this information will not be present on the test page on the screen. From the list of sentences obtained in this way – satisfying the filtering criterion – some are randomly selected within the selection module of the proposed algorithm. Questions on the test page are generated from these sentences. The word within the selected sentences to be distracted for a question is determined by the classification algorithm. The classification algorithm determines the subjective coordinates of words according to the objective coordinates of words of the selected sentences to be distracted for the question generation [13].

The subjective coordinate called ‘whether to be distracted for a question’ is used to determine whether the given word of a sentence can be distracted for the question generation or not. In the case of negative decision the algorithm randomly selects another sentence from sentences satisfying the annotation condition. However, if this subjective coordinate comes up with several words of a given sentence to be eligible then the word to be distracted will be determined according to the input function of the output neuron defined by this subjective coordinate of the neural net carrying out the classification.

On the abstract structure level, the proposed engine consists of the following modules:

- text preprocessing
  - annotations
  - stemming
- - domain preprocessing:
  - topic based clustering of words (semi-synonyms)
  - generation of term hierarchy
- - selection of key word: classification of words
- - selection of the candidate words using the word clusters
  - similar words
  - un-similar words
- - grammatical alignment of the candidate words
- - generation of test
- - evaluation module.

The system uses a semi-automated selection method for determining the base sentences of the questions. Initially, an annotation framework is used to denote by human experts which sentences are suitable and optimal for question generation. Having a large set of training examples, a classification method is applied to find out the hidden relationship between the sentence parameters and the experts' decisions. Like in the selection of candidate words for the question, also here a CPN (Counter Propagation Network) variant was invoked. Within the text prepossessing phase, another important step is the grammatical prepossessing of the words. As no powerful commercial language morpheme analyzer could be involved, a simplified free version of the Szószablya framework [21] is applied here. This module can determine the word class and stem of the words, but the dictionary of the free version can't cover all words of the incoming documents

III. TEST RESULTS

The correctness of the model was tested by software implemented in Java language. The developed software outputs questions in electronic or printable format. In the case of electronic version both the generated questions and possible answers appear in computer environment in front of the user. In this version, users attending the test can have a local menu on the screen offering all the possible answers to questions by clicking on the blank part of the sentence containing the question to be answered. After giving the answer the chosen alternative is automatically substituted in the sentence. After finishing the test, the filled test sheet can be saved in a file format and can be forwarded to the reviewer of the test. For representing the test sheet in a printable format the software indicates the place (blank part) of the word taken out as a question with dots for the person who is filling it in and the possible answers appear listed next to each other below sentences

14. Database managing system is a program system whose task is to guarantee approach to the database and to perform internal maintenance functions <.....> .
1.) connect 2.) starter 3.) information 4.) task 5.) database

Figure 1: Sample sheet entry

In order to fill the test in a printed format the user has to underline the word judged to be appropriate or write it where the dots are. Figure 1. shows an example for the test sheet in the form it appears [5].

IV. THE ASSESSMENT OF THE QUESTION GENERATION PATTERN SYSTEM

The testing and the assessment of the results of the automatic question generation module took place in the second term of the academic year 2011/12 [3]. The aim of the test was to prove the adaptability of the developed algorithms in practice. The study covered the analysis of acceptability and interpretability of automatically generated questions created by the developed software through testers. Currently the question generation prototype system is capable to generate three types of sentences (concept, definition, declarative sentence) from which each of the three types appeared in the prepared test paper in the form of multi-choice questions.

The curriculum used in the test was the note of the subject Database systems from which computer engineers (B.Sc.) and web programmers (Adult Vocational Training) were trained and it can be downloaded from the website of the Faculty of Comenius College freely. 45 people took part in testing from who 40 were students and 5 were experts. Students were selected for the test such that 50% of them (20 students) were taught the subject mentioned above and the other half of students were not taught the content of the subject. Each of the professionals was an instructor of the subject. According to the criteria defined in this way people who took part in testing were categorized into the following three categories:

- Students not taught the subject,
- Students taught the subject,
- Experts

In the survey each person had to fill in two test papers. One of them contained questions generated by the developed question generation pattern system and the other consisted of manually generated questions by instructors teaching the subject. In both cases sentences used as multi-choice questions were randomly selected from the 1000 sentences annotated in the whole written curriculum of the subject. Both test papers contained 30 questions and 5 possible answers were assigned to each question by the instructor of the subject or the computer. The aim of using manually created tests too was to enable us to compare results gained after filling it in with the efficiency of questions generated automatically by the computer application. A similar approach can be read in the paper of Canella, Ciancimino and Campos [7] where the extracted concepts were ranked manually. Ranking of concepts was carried out using Likert scale 5 point [8].

Each participant in the survey filled in both test papers. The one created by the automatic question generation system and the other created by the instructors of the subject. Afterwards results were evaluated. On test papers created by the pattern system in 19 cases out of 30 a word belonging to the natural science was extracted for the question. Concerning the rest of the 11 sentences the extracted word belonged to the linguistic science. Contrarily on the test paper created manually in 25 cases out of 30 the extracted word belonged to the linguistic science and only in 5 cases it belonged to the natural science. The average of the correct answers given by the tested people belonging to the three groups is represented in Figure 2.

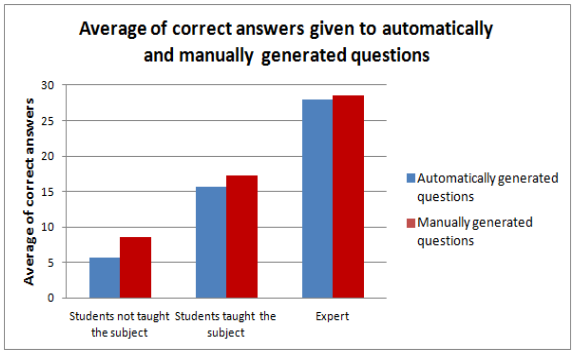

Figure 2: The ratio of correct answers in groups

According to Figure 2 it is noticeable that the ratio of correct answers given to automatically generated questions approaches very well the ratio of correct answers given to manually generated questions. Results also show that the tested people managed to give more correct answers to questions belonging to the linguistic science on the average. This difference was demonstrable at the least extent in the case of professionals of the subject who knew the exact content and importance of special concepts in sentences.

Coordinates \Values	Strong tally of the category	Tally of the category	Irrelevant from the point of view of the category	Represen-tative of the category	Strong representative of the category
Relevance	0	9	84	35	41
Speciality	0	9	64	37	27
Meaning-fulness	0	0	28	92	53
Difficulty	0	5	41	125	2

Figure 3: The relationship between the correct answers and the subjective coordinates

In the next step of the analysis our aim was to show if there is any correlation between the subjective coordinates defined during the classification of words and the correct answers. In order to do this first we analyzed the group of tested people who were not taught the subject.

It is clear from the data of the table that from the subjective coordinates the strongest correlation can be shown with the number of correct answers if the value of the Difficulty coordinates belongs to the category. Most of the correct answers were given to questions from which the extracted word was estimated more difficult than the average by the professional. Tested students who were taught the subject data gained show unambiguously that in the case of all four subjective coordinates' words fitting to a given category more than the average result in most of the correct answers. Contrary to a similar demonstration for the students who were not taught the subject the current data show correlation of almost double strong between the subjective coordinates representing the belonging to a given category and the number of correct answers. The main reason for this is that the group of students who were taught the subject has a more categorized knowledge in the field of the analyzed subject.

V. CONCLUSION

The developed AQG prototype system was tested in a real application with high-school students and it was compared with manually generated tests. The performed analysis shows that the automatically generated questions are as good as the manually constructed tests and could be used in future e-learning frameworks.

ACKNOWLEDGMENT

This research was carried out as part of the TAMOP-4.2.1.B-10/2/KONV-2010-0001 project with support by the European Union, co-financed by the European Social Fund.

REFERENCES

[1] Kovacs, L., Bednarik, L.,"Efficiency Analysis of Quality Threshold Clustering Algorithms", Production Systems and Information Engineering, University of Miskolc.2012

[2] B. Wyse and P. Piwek, "Generating Questions from OpenLearn study units", Proceedings of AIED 2009, pp.66-73.

[3] Kovacs, L., Gyöngyösi, E., Bednarik, L., "Development of classification module for automated question generation framework", Teaching Mathematics and Computer Science, 2012, in press.

[4] Porter, M.F. "Snowball: A Language for Stemming Algorithms.", http://www.snowball. tartarus.org/texts/introduction.html., 2001

[5] Kovacs, L. & Bednarik, L., "Automated EA-type Question Generation from Annotated Texts", Proc. of SACI 2012, Timisoara, Romania.

[6] Kovacs, L, Barabás P.: "Requirement analysis of the internal modules of natural language processing", Proc. of SAMI 2012, Herlany, Slovakia

[7] Canella, S., Ciancimino, E. & Campos, M.L "Mixed e-Assessment: an application of the student-generated question technique", Proc of IEEE International Conference EDUCON 2010, Madrid, Spain

[8] Earl R., "The practice of social research, California", USA: Wadsworth Publishing Company, 1998

[9] D. Coniam, "A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Tests", CALICO Journal, 14 (2-4), 1997, 15-33

[10] M. Collins and N. Duffy, "New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron", Proc. of 40th Annual Meeting of the Association for Computational Linguistics, 2007, 263-270

[11] E. Sumita, F. Sugaya, and S. Yamamoto, "Automatic Generation Method of a Fill-in-the-blank Question for Measuring English Proficiency", Technical report of IEICE, 104 (503), 2004, 17-22

[12] Y. C. Lin, L. C. Sung and M.C. Chen, "An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding", ICCE 2007 Workshop Proc. of Modeling, Management and Generation of Problems / Questions in eLearning, 2007, 137-142.