11/1 (2013), 89–102
tmcs@math.klte.hu
http://tmcs.math.klte.hu

Teaching Mathematics and Computer Science

Development of classification module for automated question generation framework

Erika Gyöngyösi, László Kovács and László Bednarik

Abstract. Automatic question generation is in the focus of recent researches which includes bordering disciplines like education, text mining, knowledge-engineering. The elaborated system generates multi-choice questions from textbooks without using an external semantic database. One of the base modules of the system is the classification module defining the extracted word. This paper describes modules of the framework including a detailed analysis of the classification part. We show the operability of the elaborated system through a practical test. Key words and phrases:

Key words and phrases: question generation, clustering, classification, multi-choice questions, assessment.

ZDM Subject Classification: R40, R70, U70.

1. Task of the automatic question generation

One of the future directions of the development of educational tools supported by informatics is signified by the adaptive and semantic oriented computer tutorial systems. Nowadays the main feature of architecture is the appearance of semantic databases storing the knowledge of the target topic next to traditional frameworks. Building of the learning material of these curriculum databases should provide high level of flexibility. The adaptability primarily means the capability to adapt to students' requirements. The flexibility appears in several functions of the system even in the knowledge testing module being the object for the study. A good example to illustrate the logical structure of systems based on knowledge database is the system OntoEdu [8]. The general architecture of a next-generation semantic e-learning framework is presented in [10].

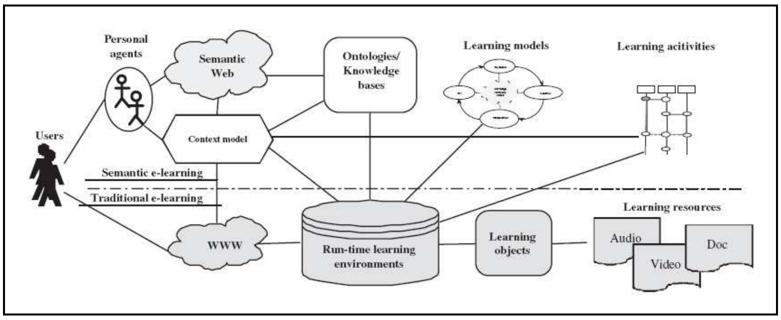


Figure 1. Architecture of a semantic e-learning framework in context (source: [10])

One of the main functions of the semantic oriented educational systems is checking students' knowledge namely testing. Generating questions is in itself a complex task as it has to suit to the topic, the goal of testing and the current state of the student. The general form of the question generation task can be characterized by six parameters. The notation and definition of each parameter can be summarized as follows [14]:

- S: the document source, according to which the question is generated,
- A: the answers required to give, information on which the question is about,
- Q: the questions generated,
- SA: the group of features of the source document and that of annotations (for example the format of the document, the description of the structure of the document),
- AA: the group of features describing the potential answers given to the question and that of annotations,
- QA: the group of features describing the question generated and that of annotations.

The general model of question generation on text documents can be seen in Figure 2.

One of the difficulties of the task is in connection with the definition of the question type. In the literature several question classification methods are known from which [2] can be emphasized for its clear, sound system. According to it the following main levels can be distinguished: recognition level, recall level, comprehension, application level and analysis level. Other researchers [3] worked out a more detailed system consisting of several hundreds of question types.

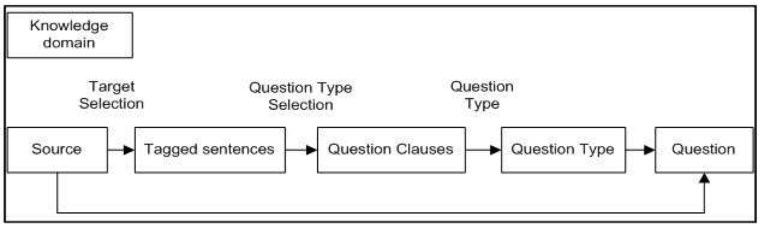


Figure 2. The general model of question generation on text documents (source: [13])

The automatic question generation cannot be considered without the semantic knowledge of concepts of the topic. George Miller et al at the Cognitive Science Laboratory of Princeton University created a lexical semantic net that is built around words and concepts of the English language [11]. This special semantic net representing synch language knowledge is called WordNet. WordNet is an electronic lexical semantic database in which the language concepts are built in a net. Concepts are represented by sets of synonyms (synsets) and connections between them are described by semantic relations (hyponym, meronym, antonym etc.).

According to the information stored in WordNet, six types of questions can be created: definition, synonym, antonym, hypernym, hyponym and multi-choice questions. In order to extract special data from WordNet the proper meaning of the given word has to be given. The problem is that several words correspond to a given word in WordNet. Most often the input contains only the given word as part of a speech [4]. The defining question offers the definition concerning the word. It is available from the WordNet note part. In questions of synonym type the word is connected to its synonymous word and the system extracts the synonym of the given word from WordNet. In the case of antonym type, the students are required to pair a given word with its opposite. Questions of hypernym and hyponym types have a similar structure. The hypernym is a general definition that means a categorization, a special reference. In the case of multi-choice questions types, the main question is given followed by several answers from which only one is correct. Depending on the type of questions the given word can occur both in the given question and in the given answers.

2. The main modules of the question generation system

In the current first phase of the research, only the simplest type of question was investigated. In determining architecture of the question generation framework, the prototype system of Coniam [6], as one of the earliest model was used as a base model. The first module of the proposed architecture carries out the structural parsing and pre-processing of the text of the source documents. Afterwards the assignment of sentences containing the subject of the question takes place where a preference learning method is applied. In the next step the extracted word is determined inside the sentence under the procedure Conditional Random Fields. In the third phase the list of words getting into the option list is built according to an external data source, the semantic data base [7].

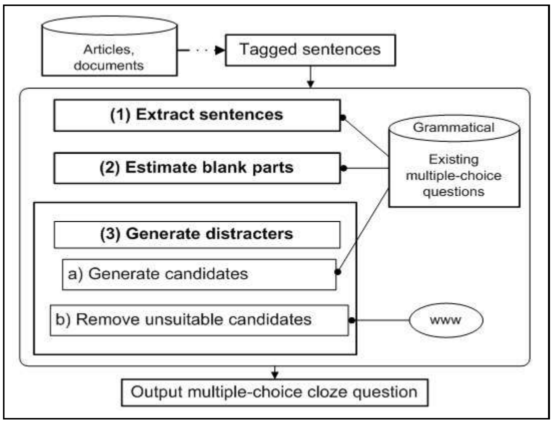


Figure 3. The base model of the question generation framework (source: [7])

In the developed prototype system a new architecture is needed which can work even without an external semantic database since the WordNet database is not available in the target language. Therefore an own dictionary of concepts was created within an internal module. Another difference compared to the base model is that it uses an individual classification method based on the neural net in order to determine words and sentences. The developed question generation framework is shown in Figure 4 [1].

The first important step of the question generation prototype system is the pre-processing whose main aim is to bring the documents into a form in which

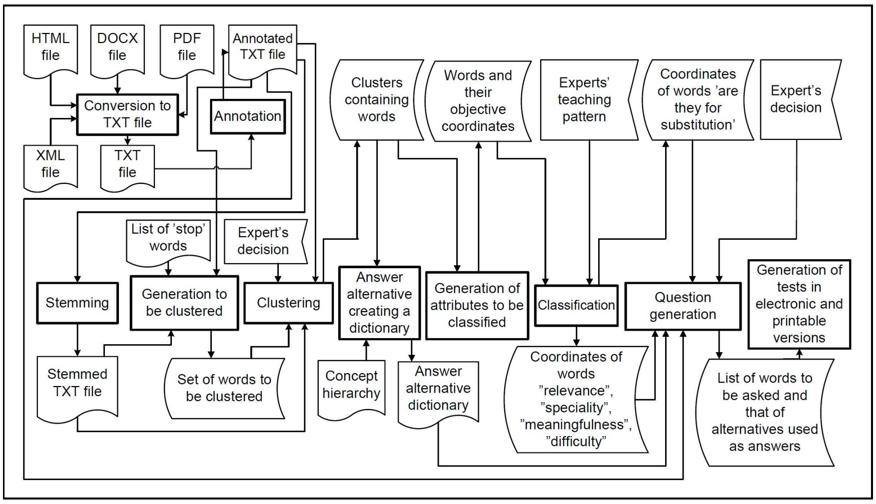


Figure 4. Modules of the developed question generation framework system (source: [1])

classification, clustering tasks can be carried out efficiently. The aim of the annotation module is to assign a role to each sentence. The annotation was carried out manually in the first phase of the prototype system. Currently, the goal of next phase is to automatize the sentence annotation process. The annotation language was created as an enhancement of the descriptive scheme language DocBook XML. Stemming is a procedure to reduce words with a suffix or prefix modifying the meaning into their stem. Through stemming the set of words to be recognized and handled can be considerably reduced as in the Hungarian language a basic word may occur in 20–50 inflected forms in the text. A free module the Szószablya framework [12] was built into the system whose algorithm works with the adaptation of the well-known Porter algorithm [15]. Porter's model applies elementary reduction processes to produce the stem. The main advantage of this method is the high speed. The task of the clustering module is to substitute an external semantic database with an own hierarchical word clustering. In the clustering the distance between words corresponds to distances between the topics of the target words. The clustering was performed with the adaptation of the algorithm BIRCH which provides a more efficient result than the other methods for large datasets. In the classification module the usefulness of words as extracted words is measured. The usefulness can be managed as a binary classification problem where the class code depending on the objectively immeasurable dimensions of words needs to be determined. The classification task was solved by using the most convenient neural network. Determination of sentences to be questioned and words extracted from sentences to be questioned takes place in the question generation module. Furthermore in this module the possible answers be given to the posed questions are generated. The module provides this information to modules displaying and evaluating results.

3. The operational principle of the classification module

One of the main elements concerning the problems of the computer intelligence is the classification task during which documents are assigned to pre-defined class codes. While classification we assume that the objects can be characterized by a given set of features and the class code of the object depends on the features of the object. The goal of classification is to define the function describing the unknown relation $f : O \to 2^C$ as well as to determine a classifying function

$$g : O \to 2^C \tag{3.1}$$

for which

$$E(f, g, S) \to \min \tag{3.2}$$

holds where the error function is denoted by E, the function describing the unknown relation is f, the classifying function is g, the teaching set is S, and the set of codes is C. The function g obtained in this way can be used to converge to the function f on the whole set O.

During our research we developed a classification algorithm that is capable to create connection between the exactly measurable objective coordinates of the words in the document and the subjective coordinates representing human's judgements. The main task of this classification module is to determine the candidate words for substitutions (blank positions in the questions). As human's opinions cannot be accessed directly, the visible parameters are used to infer the human's decisions. Linguistic features and features objectively measurable by statistical methods of words are called *objective coordinates*. The objective coordinates used are: part of speech, measure of inflection, number of occurrence in the document, number of occurrence in the sentence, place in the sentence, serial number of cluster, its distance from the rest of the words in the sentence containing the given word. The network was set to the special features of the particular task according to both structural and functional aspects. Revealing of rules of the transformation between objective and subjective coordinates was achieved as an optimisation task through which the target function of the optimisation of the classifying algorithm was defined as the maximum of the level of knowledge described by the following expression: $\frac{\textit{Number of words classified correctly}}{\textit{number of classified words}}$. The limit condition of the optimisation is the time spent on learning or not exceeding the maximum number of steps.

The prototype system applies a CPN (Counter-Propagation Network) classifier [9] based on a three-layer feed forward architecture network for the classification phase. The learning with neural network was published first by McCulloch and Pitts in 1943. Neural networks contain units (neurons) linked by directed connections. The link a_j from the neuron denoted by j leading to the neuron i is in charge of forwarding the activation state of neuron j towards neuron i. Every single link is connected to a numeric weight $W_{j,i}$ which determines the strength and sign of the link. Every single neuron first calculates the weighted sum of its inputs according to the following formula (3.3).

$$in_i = \sum_{j=0}^{n} W_{j,i} * a_j \tag{3.3}$$

The activation state of the ith neuron is determined by the activation function g applied on the weighted sum of its inputs according to the formula (3.4).

$$a_i = g(in_i) = g\left(\sum_{j=0}^{n} W_{j,i} * a_j\right) \tag{3.4}$$

According to the flow of information neural networks can be classified into two main classes. These are the feed-forward networks and the recurrent networks [16]. The feed-forward networks can be distinguished from each other according to the number of layers, even the number of neurons in each layer and the links between neurons. Typically, a neural network consists of three layers carrying out three different tasks. These are the input layer, the hidden layer and the output layer.

The input layer of the CPN network contains the description of the input objects to be classified. In the hidden layer there is a layer called Kohonen SOM whose aim is to cluster the documents. The output layer carries out a nearest-neighbour classification. During the learning process in the SOM layer the winning neuron always gets closer to the object linked to the input. After setting the initial weights, the algorithm takes the words of the training set one after the other linking them to the input of the network. This is carried out through a special adaptor module which determines the neuron of the neural network in the dynamically built layer to each value of the objective coordinate of the word. As a result of linking the input neurons representing the position of the given word in the objective space get into an active level. After setting the activation level of neurons in the hidden layer determining the activation level of neurons in the output layer takes place. To sum up the activation level of the neuron i – in the output layer – can be defined by the formula in (3.5) according to the activation level of neurons in the input layer.

$$a_i = \left(\sum_{j=0}^{n} W_{j,i} * g \left(\sum_{k=0}^{m} W_{k,j} * a_k \right) \right) \tag{3.5}$$

where

a_k: The activation level of the neuron in the input layer k,

a_i: The activation level of the neuron in the output layer i,

m: The number of neurons in the input layer,

n: The number of neurons in the hidden layer,

$W_{k,j}$: The strength of the link between the neuron in the input layer k and the neuron in the hidden layer j,

$W_{j,i}$: The strength of the link between the neuron in the hidden layer j and the neuron in the output layer i,

g: Activation function which takes the value 1 if the sum obtained as a parameter is positive and 0 otherwise.

The basic structure of the applied neural net can be seen in Figure 5 [1].

As the set of range of objective coordinates can considerably vary in different documents or in the function of clustering parameters therefore in order to determine the exact number of neurons in the input layer another algorithm was developed. This algorithm reveals the set of values taken by the objective coordinates before starting the classification and it takes a separate neuron to each possible value of each objective coordinate in the input layer.

One of the key questions of building neural nets is how to determine the number of neurons in the inside layer. In the case of applying fewer neurons than the optimal value the net will not be able to store then necessary amount of information for learning the task. In the case of applying more internal neurons than the optimal value the capacity of generalisation of the neural set decreases.

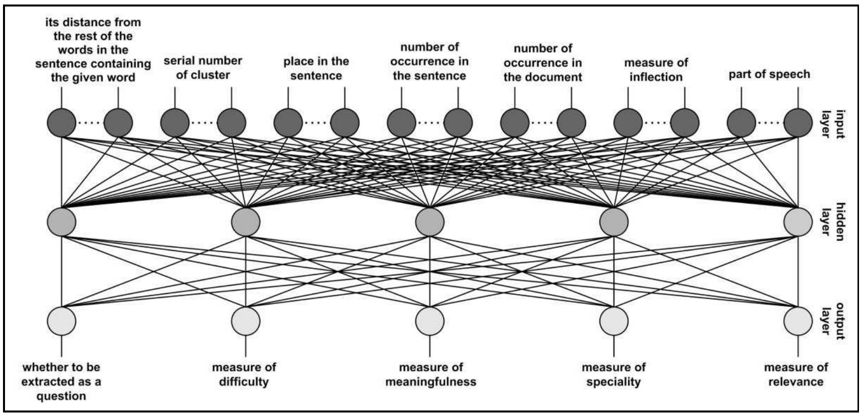
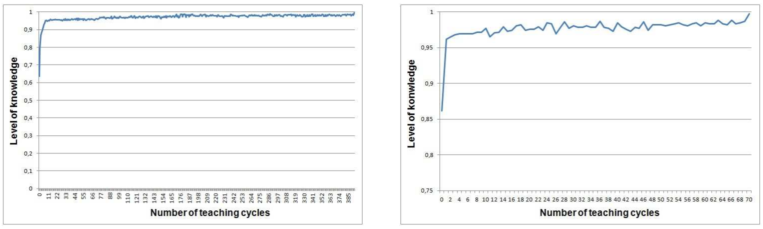


Figure 5. Basic structure of the applied neural net (source: [1])

The rate of decrease is in direct proportion with the number of internal neurons used in addition to the optimal value. Consequently, the decrease in the capacity of generalisation implies that the neural set can reveal rules in patterns used for learning at a decreased rated.

With neurons in the output layer the possible values of subjective coordinates of words are modelled. *Subjective coordinates* denote features of words that can not be exactly defined expressing human's judgement related to a given word. The set of range of subjective coordinates is fix hence the number of output neurons is set to 22. Dimensions of the subjective space are described by the following five coordinates: relevance, speciality, meaningfulness, difficulty and "whether to be extracted as a question". The set of range of the first four coordinates is described by five different values. With each value we determined the relation of words of documents to the feature expressed by the given coordinate. With the fifth subjective coordinate the expert's judgement was represented regarding the extraction of the given word.

The process of learning is completed if the average error of the classification of words in the training set does not exceed the allowable threshold or if the number of cycles gets at the number of repetition allowed. Figure 6 shows changes experienced in the level of knowledge in the network in the case of different values of the parameters of the teaching weight unit.



(a) unit of weight: 0.01 (b) unit of weigth: 0.1

Figure 6. The dependence of the learning efficiency from the learning weight

4. The assessment of the question generation pattern system

Testing and the assessment of results of the automatic question generation pattern system took place in the second term of the academic year 2011/12 at Comenius Faculty of the University of Miskolc. 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 to test practically the results of the automatic question generation 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.

40 out of 45 people who took part in testing 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 with the test created manually by the instructors of the subject 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 [5] where the extracted concepts were ranked manually. Ranking of concepts was carried out with the help Likert scale 5 point.

Each participant in the survey filled in both test papers. The one created by the automatic question generation pattern 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 7.

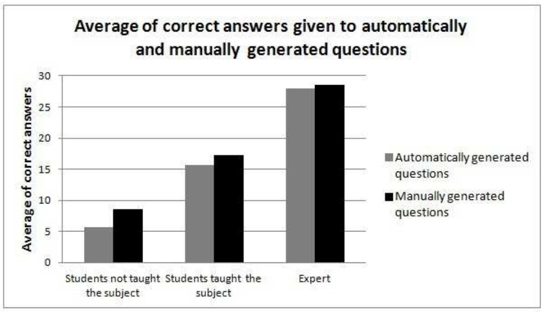


Figure 7. The ratio of correct answers in groups

According to Figure 7 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.

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.

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

Figure 8. The relationship between the correct answers and the subjective coordinates

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.

5. Summary

The developed framework generates multi-choice questions from textbooks without an external semantic database. The framework first carries out text mining pre-processing steps and then it builds up an own semantic database through clustering words on the base of concept similarity. One of the basic modules of the system is the classification module determining the extracted word in which a method based on the neural network was worked out. Experiments show that the newly developed conception can be applied instead of the more time consuming manual question generation. With the several tuning possibilities (distance of words obtained by clustering, concept hierarchy, subjective coordinates) used in the new method it is achievable that questions will represent the essentials of the curriculum under testing better than manually generated questions. The effect of this was shown by the results of tests.

References

[1] L. Bednarik, *Automatizált kérdésgenerálás annotált szövegből*, Ph.D. disszertáció, 2012, `http://www.hjphd.iit.uni-miskolc.hu/files/disszertaciok/Bednarik_disszertacio.pdf`.

[2] B. Bloom, *Taxonomy of Educational Objectives*, Handbook1: Cognitive Domain, Addison Wesley, 1956.

[3] K. E. Boyer, R. Phillips, M. D. Wallis, M. A. Vouk and J. C. Lester, *Proc. of 2nd Workshop on Question Generation*, 2009, 9–16.

[4] J. Brown and M. Eskenazi, *Retrieval of Authentic Documents for Reader-Specific Lexical Practice*, in: *Proceedings of InSTIL/ICALL Symposium*, Venice, Italy, 2004.

[5] S. Canella, E. Ciancimino and M. L. Campos, *Mixed e-Assessment: an application of the student-generated question technique*, Paper read at IEEE International Conference EDUCON, Madrid, Spain, April, 2010.

[6] D. Coniam, A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests, *CALICO Journal* **14**, no. 2–4 (1997), 15–33.

[7] T. Goto, T. Kojiri, T. Watanabe, T. Iwata and T. Yamada, Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation, *Knowledge Management & E-Learning: An International Journal* **2**, no. 3 (2005), 1–15.

[8] C. Guangzuo, C. Fei, C. Hu and L. Shufang, *OntoEdu: A Case Study of Ontology-based Education Grid System for E-Learning*, Proc. of GCCCE2004 International conference, Hong Kong, 2004.

[9] R. Hecht-Nielsen, Counterpropagation Networks, *Applied Optics* **26**, no. 23 (1987), 4979–4984.

[10] W. Huang, D. Webster, D. Wood and T. Ishaya, An intelligent semantic e-learning framework using context-aware Semantic Web technologies, *British Journal of Educational Technology* **37**, no. 3 (2006), 351–373.

[11] G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller, *Five Papers on WordNet*, CSL Report 43. Cognitive Science Laboratory, Princeton University, 1990.

[12] L. Németh, A Szószablya fejlesztés, 2003, 3–4.

[13] S. Pal, T. Mondal, P. Pakray, D. Das and S. Bandyopadhyay, *QGSTEC System Description – JUQGG: A Rule based approach*, Proceedings of the Third Workshop on Question Generation, 2010, 84–90.

[14] P. Piwek, J. Mostow, Y. Chali, C. Forascu and D. Gates, *Data Requirements, Sources and Annotation Schemes for Question Generation Shared Tasks*, The Question Generation Shared Task and Evaluation Challenge, Institute for Intelligent Systems, University of Memphis, 2009, 15–20.

[15] M. F. Porter, *Snowball: A Language for Stemming Algorithms*, 2001.

[16] S. P. Victor and S. Ramasundaram, Text Categorization by Backpropagation Network, *International Journal of Computer Applications* **8**, no. 6 (2010), 1–6.

ERIKA GYÖNGYÖSI and LÁSZLÓ BEDNARIK
COMENIUS FACULTY
UNIVERSITY OF MISKOLC
EÖTVÖS ÚT 7.
3950 SÁROSPATAK
HUNGARY

E-mail: gyerika@hotmail.com

E-mail: laszlobednarik@gmail.com

LÁSZLÓ KOVÁCS
DEPARTMENT OF INFORMATION TECHNOLOGY
UNIVERSITY OF MISKOLC
H-3515 MISKOLC-EGYETEMVÁROS
HUNGARY

E-mail: kovacs@iit.uni-miskolc.hu

(Received July, 2012)