Learning Methods for Rating the Difficulty of Reading Comprehension Questions

Dorit Hutzler
Department of Mathematics and Computer Science
The Open University of Israel
Raanana, Israel
hutzler@g.jct.ac.il

Esther David
Department of Computer Science
Ashkelon College
Ashkelon, Israel
astrdod@acad.ash-college.ac.il

Mireille Avigal
Department of Mathematics and Computer Science
The Open University of Israel
Raanana, Israel
miray@openu.ac.il

Rina Azoulay
Department of Computer Science
Jerusalem College of Technology
Jerusalem, Israel
rrinaa@gmail.com

***Abstract*— This work deals with an Intelligent Tutoring System (ITS) for reading comprehension. Such a system could promote reading comprehension skills. An important step towards building a full ITS for reading comprehension is to build an automated ranking system that will assign a hardness level to questions used by the ITS. This is the main concern of this work. For this purpose we, first, had to define the set of criteria that determines the rate of difficulty of a question. Second, we prepared a bank of questions that were rated by a panel of experts using the set of criteria defined above. Third, we developed an automated rating software based on the criteria defined above. In particular, we considered and compared different machine learning techniques for the ranking system of the third part of the process: Artificial Neural Network (ANN), Support Vector Machine (SVM), decision tree and naïve Bayesian network. The definition of the criteria set for rating a question's difficulty, and the development of an automated software for rating a questions' difficulty, contribute to a tremendous advancement in the ITS domain for reading comprehension by providing a uniform, objective and automated system for determining a question's difficulty.**

Keywords- machine learning and analytics; inteligent tutoring systems; evaluation methodologies

I. INTRODUCTION

The technological advances over the past century have led to a number of negative outcomes - one of the most prominent of which is the growing dependence of human beings on machinery, automation and advanced technology. By relying on automation, human beings have, in sense, become automatons themselves, processing information only superficially without attempting to gain a deeper understanding of things. In consequence, we have been witnessing a gradual but significant decline in reading comprehension.

In light of the heightened awareness of the pressing need to promote the subject of reading comprehension, and given the lack of technological tools needed for its instruction, we elected to develop a software that can be used for this purpose - an ITS for reading comprehension.

The software in question consists of two sub-programs. One sub-program receives the texts and related questions and ranks their relative degrees of difficulty. This program basically facilitates the conversion of learning methods and progress monitoring into practical and measurable variables.

The second sub-program supports and conforms to a modern system of instruction, which relies more and more on e-learning, and is discussed in [1]. The goal of the program on the whole is to facilitate the development of more suitable and sophisticated learning systems for the instruction of reading comprehension, ranked by the degree of difficulty.

In this paper we concentrate on the development of the first subprogram that is in charge of the automated ranking of the questions' difficulty. In a computerized, technological world, this software will undoubtedly serve as a fundamental tool in developing learning systems for the instruction of reading comprehension.

In doing so, we present a comparison of various machine learning methods for the task of automated ranking of the question's difficulty level. In particular, we surveyed the following machine learning methods: (1) artificial neural networks (ANN), (2) Support Vector Machine (SVM), (3) decision tree, and (4) naïve Bayesian network. This comparison aimed to identify the outperforming method that best suits the domain of automated ranking of reading comprehension questions' difficulty level.

The paper is organized as follows. In section II we review the current state of the art in the domains related to our research questions. In section III we describe the machine learning methods we compared. In Section IV we present the dataset preparation. In section V we provide the experiment results of the various methods we tested for the automated question's difficulty ranking task. We conclude in section VI.

II. Related Work

In the following section we describe related work that considers automated tools for teaching reading comprehension and automated tools for rating the hardness of reading comprehension tasks.

A. Reading Comprehension Improvement

The best way to acquire reading comprehension skills is through reading. Self-study through frequent reading fosters and enhances reading comprehension skills. Books should complement classroom learning.

Snow et al. [2] discuss a study performed by the US' Department of Education's Office of Educational Research and Improvement. Their findings show that today's high school students exhibit a serious deficiency in reading comprehension.

Another research by Cornoldi and Cornoldi [3] shows that teachers consistently encounter students who do not fully understand what they have read. While there is a campaign underway to improve reading comprehension in high school grades 10-12 in various subjects, the results of reading comprehension tests in the United States as well as in other countries do not indicate any actual improvement.

These results are explained by the fact that, as shown by Snow [2], students in this generation find it difficult to sit and read a book; today's students are accustomed to having information provided to them in a manner that is highly visual, tangible, immediate and interactive. Apparently, this expectation developed due to the extensive use of computers, impeding the improvement of reading comprehension gained from reading books.

Gersten et al. [4] outline several techniques that have been shown by previous research to be effective in improving reading comprehension. They conclude that reading comprehension may best be improved by combining several different methods of instruction.

Dreyer and Nel [5] show that both successful students and those considered ‘at risk’ have become accustomed to a combination of different teaching methods.

B. E-Learning Systems and Question Ranking Criteria

In recent years, the use of e-learning technology has increased significantly. This appears to be one of the most successful methods for promoting student advancement, from both the student's and the teacher's perspective. The introduction of technology and its easy accessibility increases the student's motivation to learn, and it facilitates student assessment in terms of both response time and level of knowledge.

E-learning systems typically include question banks, in which questions are ranked by level. However, before these conclusions can be employed in the development of computer programs for teaching reading comprehension skills, questions on the written text must first be rated in order to enable the tutoring system to adjust the level of questions to the student’s level according to the learning algorithm used.

The most well-known system of classification is the first taxonomy, also known as Bloom's taxonomy [6], which consists of three domains: cognitive, affective and psychomotor. Bloom's cognitive taxonomy is divided into six levels, which are described below in order of difficulty (from easiest to hardest):

1. Knowledge - retrieval of data, information or specific items, remembering concept definitions.
2. Comprehension - retrieval as well as grasping the meaning, translation, summary or explanation of the retrieved item.
3. Application - using comprehension of learned material in order to apply it in a different way.
4. Analysis - classification of the learned material or idea into its component parts, and understanding its structure.
5. Synthesis - using components of the learned material to create a new structure.
6. Evaluation - appraisal of the material that was learned or the ideas that were presented.

Bloom's taxonomy has had a significant impact on the field of educational psychology, and in particular, on theories of evaluating academic achievement. Nevertheless, it has been subject to many critical reviews, resulting in the derivation of various updated versions of the taxonomy, such as that of Anderson and Krathwohl [7]. The main criticism of the first taxonomy lies in the difficulty in applying these levels consistently to a given set of questions.

Moreover, classification according to Bloom's taxonomy is not absolute; rather, it depends on context and on what is learned in class. In addition, the taxonomy purports to be applicable to all content areas, thereby ignoring the convention in cognitive psychology that states that each content area is associated with its own particular sphere of knowledge and skills.

These six levels can be further collapsed into three primary domains: knowledge, comprehension and higher mental processes, as in Anderson and Krathwohl [7]. However, it is difficult to determine definitively that all questions of knowledge are necessarily easier than questions of comprehension.

Davis [8] did a survey to determine the main nine groups of skills involved in reading comprehension. These include the following:

1. Knowledge of word meanings.
2. Ability to select the appropriate meaning of a word or phrase in the light of its particular contextual setting.
3. Ability to follow the organization of a passage and to identify antecedents and references in it.
4. Ability to select the main thought of a passage.
5. Ability to answer questions that are specifically answered in a passage.
6. Ability to answer questions that are answered in a passage but not in the words in which the question is asked.
7. Ability to draw inferences from a passage about its contents.
8. Ability to recognize the literary devices used in a passage and to determine its tone and mood.
9. Ability to determine a writer's purpose, intent, and point of view, i.e., to draw inferences abut a writer.

Interestingly, a high similarity exists between the set of nine criteria we established for defining a question's level as listed in Table 1, and the nine groups of skills derived by Davis [8] (especially between the fifth point in each of the lists). Although several skills are required to succeed in any reading comprehension task, nonetheless, for a given question not all skills are required to answer it correctly. Our contribution is exactly at this point where we define which of the skills are required for each question and in which level.

Levi and Dalal [9] provide explanations of how to correctly answer various reading comprehension questions. The explanations we provide for each criteria our panel of experts defined is actually taken from Levi and Daval's work [9].

C. Automated Learning Techniques for ITS

Till now we have reviewed several ranking methods; next we will review works relating to various learning approaches in ITS, such as Fuzzy Logic, Bayesian Networks and Neural Networks.

Dreiseitla and Ohno-Machadob [10] reviewed several classification algorithms and summarized the differences and the similarities of logistic regression and artificial neural networks (ANN). Logistic regression and ANN were compared to newer statistical machine learning algorithms, and were found to perform worse (ratio 2:5). Nevertheless, logistic regression is popular because of the interpretability of model parameters and the ease of use, and artificial neural networks are popular due to the fact that they can be seen as nonlinear generalizations of logistic regression. Dreiseitla and Ohno-Machadob [10] concluded that there is no single algorithm that performs better than all other algorithms in all the examined areas, and when using artificial neural networks, one hidden layer is generally sufficient to classify most datasets. In our research, an ANN with one hidden layer was found to perform well in the task of classifying the difficulty levels of the questions in the ITS.

Abu Naser [11] used an artificial neural network to predict the academic performance level of a student who uses the Linear Programming Intelligent Tutoring System. The Expert system (part of the ITS) was used to determine the proper difficulty level that suits the predicted academic performance of the learner. To train the ANN, Abu Naser [11] used the feed forward back-propagation algorithm. Though we used the ANN for a different purpose (to learn the questions' difficulty level), we also used the same algorithm to train our ANN.

Venkatesh et al [12] aimed to reduce the time and the cost required to develop an ITS. They built an ITS relay on an independent domain. The questions from a particular domain-course are queried from the students and the answer is obtained from a bank of answers using the Question Answering System which uses an ANN with one hidden layer. However, in their system the difficulty level is given as input, while we wanted to automatically predict the level.

D. Automated Techniques for Question Ranking

We proceed by comparing previous work on automated techniques for ranking the questions' hardness. The need for such an automatic process stems from the desire to have an adaptive system that will enable the insertion of more texts and questions to its bank of questions in a classified manner.

Yahya and Osman [13] investigated the effectiveness of support vector machines (SVMs) for the classification of item bank questions into Bloom's taxonomy of cognitive levels. The study used a dataset of questions that had been pre-rated. Before loading the questions into the dataset, each question underwent preliminary processing to remove punctuation and stop words, and individual words were stripped down to their basic root form. The database was divided into a training set ($\approx$70% of the dataset) and a testing set ($\approx$30% of the dataset). Results indicated that when questions are classified using this system, the probability that the assigned classification is accurate (i.e., the precision) is 85%, but with a recall of only 29% probability that the question will be assigned to the proper classification. The lower percentages were attributed to the small size of the dataset that was used (272 questions).

Namba [14] describes a new system to classify a learner's understanding level by using an associative Binary Cellular Neural Network (BCNN). All the questions (Java programming course) were divided into three levels (easy, standard and difficult) according to the accuracy rate of the learners. Namely, a question which many students answered incorrectly was classified as a difficult question. Similarly, a question which most students answered correctly was classified as an easy question. The author compared the performance of the BCNN with that of the Multi-Layered Perceptron (MLP) and found that the BCNN outperformed the MLP .

Namba's [14] work is similar to our work as we both aimed to associate the level of difficulty to the questions in the system. However, our work differs in the way it is done. Namba subjectively associates the difficulty level of a question based on the relative success of students in answering it. In contrast, we associate an objective difficulty level of a question based on some well defined criteria (established by experts) in an automated manner using a neural network.

III. Developing an Automated Software for Rating the Difficulty of Reading Comprehension Questions

In this section we describe the process we used to establish an automate software for rating the difficulty of reading comprehension questions. The section is organized as follows. In section A we discuss the definition of the set of criteria that will determine the rate of difficulty of a question. In section B we provide details about the various learning algorithms we compared.

A. Criteria Definition

In order to define a proper and meaningful set of criteria that will be the basis for the decision of the difficulty of a particular question, we approached a panel of experts in the domain of teaching reading comprehension using a set of

comprehensive criteria as we describe below and summarize in Table I:

1) Question style: As defined by Levi and Dalal [9], there are several different question styles: true/false questions, multiple-choice questions – questions that contain several possible answers, of which only one is correct, or open-ended questions – questions that address more in-depth information.

2) Answer location: Can the answer be retrieved from a single location in the text or is the required information presented across relatively large sections of the text? Does the question direct students to a specific paragraph or line within the text or is no guidance provided concerning where the answer is located in the text? If no guidance is provided, can the question be answered based on a relatively small section of the text (one or two paragraphs) or does it require a global inference (e.g., "What is the objective of the author?")

3) Question complexity: How many actions are included in the task presented to the student? How many words and instructions appear in the question? Some questions contain only one instruction, while others contain several instructions, as in the following example: "What is the logical relationship between the position of environmental groups, as presented in the article, and the position of the general public? Provide evidence from the article to support your answer." This question contains two instructions: "What is the logical relationship...?" and "Provide evidence…" The next three criteria (4, 5, 6) are relevant only for questions that require students to locate written information in the text.

4) Information extraction: When the answer is extracted from the text, in some cases the information appears in the text in its entirety and can be copied verbatim, while in other cases the information must be restructured after extraction (e.g., as a full sentence). That is, for some questions, the student only needs to copy or directly quote from the text, while in other cases the student must state the answer in his or her own words. For example, the necessary information may appear as part of a dialogue, while the answer must be worded as a statement.

5) Answer explicitness: To what degree is the answer presented explicitly in the text? Whereas the answer is sometimes provided explicitly, at other times it is provided in a more obscure dimension in the text. Or alternatively, the answer may be provided explicitly, but the formulation of the question does not explicitly reveal the answer, perhaps due to the use of unusual words, as in the following examples: The article uses the word 'hardships' and the question is, "What are the problems …?"; or the text uses the word 'factors' and the question is, "What brought about the results?"

6) Data filtering: Is the information summarized in the text or does the reader have to scan the text and filter the required information while skipping over irrelevant details. Such a question requires the ability to focus on relevant information and to distinguish between general and detailed information in the text.

7) Interpretation and inference: Questions of reasoning that require readers to draw conclusions, i.e., questions that deal with matters that are not mentioned explicitly in the text. Levi and Dalal [9] describe reasoning questions that require readers to draw conclusions or deal with matters that are not mentioned explicitly in the text, such as: 'What is the subject of the text?' and 'What is the author's objective?' Such questions may also require an understanding of phrases and words that have multiple meanings, as well as logical relationships and connections between different segments appearing in one or more paragraphs.

8) Evaluation and critique: Questions that require readers to appraise the logic of the writer. These questions require an understanding of the content, structure and style of language, or alternatively to infer the purpose and intent of the author. At times readers must identify the means of persuasion employed by the author or the tone and mood of the text, and express an opinion about it. Sometimes the student is required to evaluate the credibility and reliability of the text.

In any case, the student must express a reasoned opinion. As Levi and Dalal [9] explain, such questions require the students to engage in dialogue with the text, evaluate it and express their personal opinion regarding the contents of the text, the purpose of the text, and its structure and style. Since such questions can have more than one correct answer, as different students will have different opinions, the students must present a reasoned argument. This argument can be based on information taken from the text or on logic, personal taste and general knowledge. One example of such a question could be, "Is the picturesque language used in the text appropriate for its subject?".

9) Presentation of information: Is the information presented in the text or in a graph/table/diagram.

All of the criteria's values can easily be observed given the text and the question, and some of the question (question 2, 3, 8 and 9) can be acquired by an automated system. The main advantage of the definition of these criteria is that all of them can be obtained without the need of an expert in the field. Moreover, in the next step, by giving these criteria's values, the automated difficulty rating system will be able to rank the question's difficulty level without the need of an expert.

Note also that it is known that the variable to be learned (the question level) depends on the criteria value, and for some of the criteria for example, parameters 3-4, the origin of the influence is known in advance. However, the network will learn the accurate effect rate of each of the criteria on the question difficulty level, in order to predict the difficulty level of new questions given their criteria's values.

Note that criteria values were assigned such that a higher numerical value indicates greater difficulty in terms of the corresponding characteristic. For example, if no data retrieval is required for a given question, the Information Extraction criterion receives a value of 0; extraction of information that appears explicitly in the text corresponds to a value of 1, while the more difficult requirement of rewriting information in the student's own words corresponds to a value of 2.

TABLE I. CRITERIA THAT HAVE AN IMPACT ON QUESTION DIFFICULTY

Criterion Name	Input Value
1.Question style	0 - True/false 1 - Multiple-choice 2 - Open-ended
2.Answer location	0 - Location provided in the question 1 - Local; location not provided 2 - Global/ inclusive
3.Question complexity	0 - Simple question with a single action 1 - Complex question; more than one instruction
4.Information extraction	0 - No information is extracted 1 - Information can be copied verbatim 2 - Information must be restructured
5.Answer explicitness	0 - No information is extracted 1 - Information is provided explicitly 2 - Question wording is not explicit 3 - Information is alluded in an obscure dimension
6. Data filtering	0 - No information is extracted 1 - Information is summarized within the text 2 - Relevant information must be filtered
7. Interpretation & inference	0 - No interpretation or inference is required 1 - Text subject 2 - Understanding expressions 3 - Relationships between parts of the text 4 - Author's purpose
8. Evaluation & critique	0 - No evaluation or critique is required 1 - Personal opinion 2 - General knowledge
9. Presentation of information	0 – Textual 1 - Graphic, table or diagram

B. *Learning Methods for Rating Questions' Difficulty*

In this section, we present a comparison of various machine learning methods for the task of automated ranking of the questions' difficulty levels. In particular, we surveyed the following machine learning methods: (1)artificial neural networks (ANN), (2) Support Vector Machine (SVM), (3) decision tree, (4) naïve Bayesian network. This comparison aims to identify the outperforming method that best suits this domain of automated ranking reading comprehension questions' difficulties.

Each Machine learning method receives information about a question on reading comprehension (i.e., question criteria's values) as input and returns the question's level of difficulty as output. The use of machine learning methods provides a solution for both teachers and students, for teaching and learning purposes, respectively.

These methods can be used to adjust the level of reading instruction to the individual level of each student, without the intervention of a human teacher. This enables the creation of tests at a predefined level of difficulty using questions taken from an existing bank of questions, as well as graded tests. The significant advantage of these machine learning techniques is that after training, there is no need for the intervention of a skilled expert for the rest of the questions. Namely, the only human intervention required is the entry of input parameter values, which can be performed even by people who are not highly experienced.

Moreover, determining the difficulty of questions is not an exact science. Ratings may vary from person to person and from time to time, depending on the rater's subjective opinion, resulting in inconsistent ratings. In contrast, the machine learning method uses predefined criteria, thus providing consistent ratings.

Hence, the learning algorithm actually provides better results than human raters, due to the resultant uniformity and consistency in ratings. However, it is critical that the input to this network, namely, the criteria's values for each question, be extremely accurate. Entry of accurate values for each question is essential and directly impacts the veracity of the results. Therefore, those who actually enter the criteria's values for each question must be instructed carefully and precisely regarding each possible value of the various criteria.

Next we shortly describe each of the methods that were compared

1) Artificial Neural Network - ANN

ANNs are described in detail in Duda and Hart's paper [15] and Bishop's work [16]. The multi-layer feed-forward error propagation algorithm that we chose to use in this work is described in [17]. This network belongs to the class of supervised networks.

Our ANN version uses the error back propagation algorithm, as described in the study of Kugblenu et al. [18]. This algorithm initializes all weights $w_1 \ldots w_n$ (where n is the number of weights) for random low values. The network uses a quasi-Newton algorithm to determine the true value of the weights for each training sample. Then the network calculates the result for the training example.

Weights are updated until convergence, or until a maximum of 1000 iterations.

As illustrated in Fig. 1, such a multi-layer network comprises three layers: an input layer, a hidden layer and an output layer. The network comprises nine input parameters that represent the questions' characteristics. The hidden inner layer contains nine neurons, each of which is linked to each of the nine input parameters. The output layer has one output parameter, which indicates the question's level of difficulty, represented by an integer between 1 and 5.

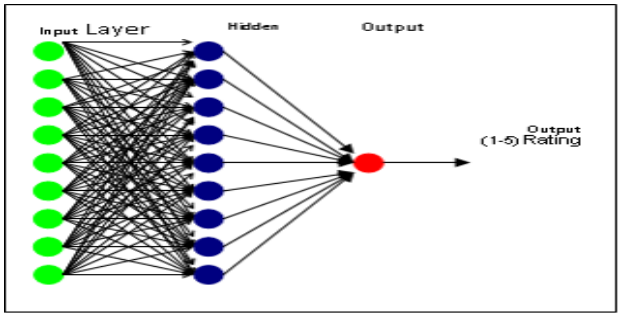

Figure 1. Network model

When using the error back propagation algorithm, weights are initialized to a random value, meaning that a different neural network may be generated each time. Since the characteristics chosen by the experts adequately cover all factors that can influence question difficulty, similar neural networks and similar performances were produced even when the same samples were run several times. This result proves the accuracy of the parameters selected.

2) Support Vector Machines – SVM

Support vector machines (SVMs), introduced by Vapnik [19], are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis [20]. Their advantage is that they can make use of certain kernels in order to transform the problem, thus enabling the application of linear classification techniques with non-linear data. Using kernels, the SVM arranges the data points in a multi-dimensional space in such a way that there is a hyper-plane that separates them. The SVMs find the best hyper-plane to separate the two classes. SVMs are based on the Structural Risk Minimization principle taken from the computational learning theory.

The idea of structural risk minimization is to find a hypothesis h for which we can guarantee the lowest true error. The true error of a hypothesis is the probability that the hypothesis will make an error on an unseen and randomly selected test example. An upper bound can be used to connect the true error of a hypothesis h with the error of h on the training set and the complexity of H (measured by VC-Dimension), the hypothesis space containing. SVMs find the hypothesis h which (approximately) minimizes this bound on the true error by effectively and efficiently controlling the VC-Dimension of H.

Given a set of training examples, the SVM training algorithm builds a model that assigns new examples to one of the categories, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

3) Decision Tree

Another machine learning approach described in [21] and in [22] is based on building a decision tree. According to the decision tree approach, a training set of objects is given, whose class is known. The induction task is to develop a classification rule that can determine the class of any object from its values of the attributes, and the classification rule will be expressed as a decision tree.

The internal nodes of the decision tree represent attribute-based tests with a branch for each possible outcome and the leaves of the decision tree indicate the resulting class for each particular path of the tree. In order to classify an object, we start at the root of the tree, evaluate the test, and take the branch appropriate to the outcome. The process continues until a leaf is encountered, at which time the object is asserted to belong to the class of the leaf.

The algorithm that builds the tree can be recursive. At each time, it observes the set of examples given to it. If all of them have the same classification class, the algorithm creates a leaf with this classification class. Otherwise, it chooses the attributes that gains most information about the examples, and generates a node that creates a branch according to this attribute. In our case, where the criteria are not binary, each criterion may have more than one node, and in each node, one binary comparison is performed. The decision tree output for the training set of questions is provided in Fig. III, Section 4.

4) Naïve Bayesian Classifier

The Naïve Bayes algorithm described, for example, in [17a] is a simple probabilistic classifier based on applying Bayes' theorem of conditional probability with strong (naive) independence assumptions [22]. It uses all the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other.

All model parameters (i.e., class values and the attribute probability distributions) can be approximated with relative frequencies from the training set. To estimate the parameters for a feature's distribution, one must assume a distribution or generate nonparametric models for the features from the training set. Then, the training set is provided, and generates updates of the probabilities of the class and of the parameters using the Bayesian rule and assuming independent probability of the parameters.

The Naïve Bayes is a popular method for document classification due to its computational efficiency and relatively good predictive performance, and it is also applicable and has achieved good performance for the problem of question classification. Each question is viewed as a collection of attributes (criteria) and their order is considered irrelevant.

IV. Dataset Preparation

The preparation of the bank of questions entailed two parts. In part I the characteristics of each question in terms of the values of their parameter was determined, and in part II the level of difficulty of each question was determined based on expert opinion. (We used an expert to determine the difficulty of the questions of the training set as well as the questions of the test set, to check the learning accuracy. However, in real world situations, only the difficulty of the training set questions should be determined by an expert). The neural learning algorithm was trained using a limited subset of data (training group).

The algorithm generated a function that matches the difficulty level of questions with the parameter values entered as input (See Fig. 2). The output for each question was the level of question difficulty, rated on a scale of 1 (a very easy question) to 5 (a very difficult question).

At this stage, the accuracy of the learning method was tested by running both the training and test groups through the system. We examined the ability to rate the questions that were used for training, as well as new questions, for which we received both input (parameter values) and output (difficulty ratings) from the group of experts. After running all questions through the system, we compared the machine learning's output with the expected ratings.

V. A Comparison of the Various Machine Learning Methods' Performances for Automatic Question Difficulty Ranking

In our study, 136 reading comprehension questions were used, taken from 9 different articles. In this section we describe the results obtained in the different experiments we performed. We used the following evaluation criteria. The first criterion, called *Boolean success rate*, is the average fraction of accurate ratings of the total number of ratings. The second method, called *Average relative error*, is the average ratio of the distance between the actual and expected rating from the maximum possible error for each question, and the third method, called *distance average*, is the average differences between the actual and the expected ratings.

Next, we describe the different experiments we ran to compare the results of the different learning methods.

A. The Results When 70% of the Questions were in the Trainins Set

In our first experiment, we compared the results of the different Machine learning algorithms by using the same set of 70% of the questions as a training set, and the remaining 30% of the questions as the test set. The input for each machine learning method was the training set. At the end of the training process, each method built a model that was used to automatically rank the unobserved question from the testing set simply based on their criteria's values. Given these learning models we compared their successes as summarized in the Fig. 2.

Step 1:
Each machine learning algorithm receives as input a set of questions (from the training set) with their criteria's values and their ranking determined by experts.

Step 2: For each question in the test set:
a. The parameters of the question are sent as inputs to the learning model of each method.
b. The model calculates the question's difficulty based on its criteria's values.
c. The question's real difficulty level determined by the expert, is observed.
d. The difference between the calculated difficulty and the real difficulty is calculated.

Step 3: The average accuracy of the algorithm is calculated, using the average difference between the calculated and the real difficulty.

Figure 2. The template of the learning algorithm

The learning methods tested in this set of experiments were: ANN (artificial neural network, using back propagation), SVM, Naïve Bayesian Classifier, and Decision Tree.

The rows of the tables are defined as follows. The Boolean success rate is the ratio of values which were inferred accurately. The average relative error is the average ratio of the calculated distance between the real level and the predicted level divided by the maximum possible distance. The average distances and distance's standard deviation are based on the distance between the question's level predicted by the learning method and the correct question level determined by the expert.

The results found using the above technique are described in Table II. As shown, the results of the different methods were very close, with a slight advantage of the ANN (artificial neural network method).

TABLE II. THE ALGORITHMS' RESULTS FOR 70% TRAINING SET

	Boolean Success Rate (accurate or not)	Ralative Error Rate	Average Distances	Distance - Standard Deviation
ANN	0.35	0.231	0.703	0.609
SVM	0.35	0.221	0.73	0.599
Naïve Bayesian Classifier	0.35	0.290	0.81	0.729
Decision Tree	0.35	0.243	0.73	0.599

B. The Cross Validation Results

In order to be able to check if the differences between the learning methods are significant or not, we tested them using a cross validation method. Using this method, for each test, a different set of the inputs (criteria of questions and the questions' difficulty level) was used as a training set, and the rest of the input was used as the test set. Our results are presented in Table III. As can be observed, the results are very close to the results presented in table II, and the ANN shows a slight advantage w.r.t. the other methods. However, for both types of experiments (70% training and cross-validation), none of the differences between the algorithms' results was found to be significantly better (for a significance level of 0.05, using the two-tailed test).

TABLE III. THE ALGORITHMS' CROSS-VALIDATION RESULTS

	Boolean Success Rate (accurate or not)	Relative Error Rate	Average Distances	Distances - Standard Deviation
ANN	0.456	0.232	0.6397	0.649
SVM	0.463	0.221	0.647	0.670
Naïve Bayesian	0.397	0.291	0.750	0.725
Decision Tree	0.440	0.243	0.679	0.718

In other words, for both types of experiments the artificial neural network shows a slight advantage w.r.t. the other methods, but the advantage is insignificant.

On the other hand, the decision tree output is more clearly understood by humans, as shown in Fig. 3. Consequently, when the learning model needs to be understood by humans, the decision tree may be preferable even though its results were shown to be slightly lower.

Thus, one may prefer this method as a learning method, whenever it is important to understand the output level.

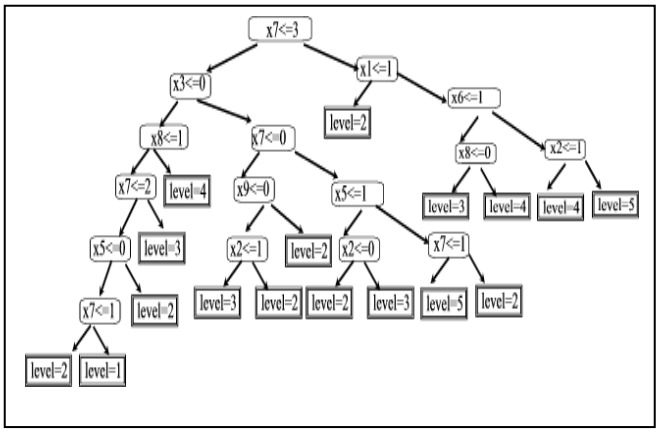

Figure 3. The decision tree created for the 70% training set

C. The Value of Information of each Parameter

In the final set of experiments, we tested the behavior of the learning algorithms (concentrating on ANN and on the decision tree algorithm) when one or more of the criteria is omitted. The motivation of this test is the fact that acquiring the different criteria is costly. Even though no expert is needed, human effort is still required for most of the criteria. Thus, we would like to be able to omit them, if possible. For each criterion omitted, we computed the difference between the average distance when the criterion is omitted and when the criterion exists (using cross-validation without and with the criterion), and we denote this difference the value of information for this criterion. Our results are presented in Table IV.

TABLE IV. VALUE OF INFORMATION FOR EACH CRITERION

Criteria	Value of information using ANNs	Value of information using Decision trees
1. Question style	0.081	0.056
2. Answer location	0.048	0.097
3. Question complexity	-0.024	0.161
4. Information extraction	0.016	0
5. Answer explicitness	0.024	0.008
6. Data filtering	0.024	-0.008
7. Interpretation & inference	**0.242**	**0.202**
8. Evaluation & critique	0.081	0.089
9. Presentation of information	0.04	0.065

As demonstrated in the results, the contribution of criterion 7 (interpretation & inference) is the most important criterion ,and its absence causes the distance to increase by 39% (from 0.64 to 0.88) for the ANN and by

24% (from 0.68 to 0.85) for the Decision tree. The influence of the other criteria is much smaller, and the absence of each of them causes the distance to increase by 12.5% at most There are also cases where the addition of a criterion causes the average distance for the test set to increase, as shown for criterion 3 (for ANN) and for criterion 6 (for the decision tree). Finally, criterion 4 was not included in the decision tree (as depicted in Fig. 3), thus its absence causes no difference in the average distance.

To conclude, if a criterion which is hard to achieve exists, there are cases where it may be omitted completely, and the learning algorithm will still be able to accurately predict the level of future questions.

Consequently, the ability to construct an automated software that learns the level of the questions given criteria that can be easily, and sometimes even automatically, acquired from the text of the question, the answer and the article text, might be possible. Future work is required to find the minimal set of criteria which is required to be able to accurately learn the question level, in order to be able to save in human efforts when creating and categorizing new questions for the ITS.

VI. CONCLUSTIONS

In this paper we were faced with the challenge of developing an intelligent tutoring system to advance students' reading comprehension skills. The paper describes two main contributions to this new research area. Our first contribution is the establishment of a set of criteria that define the difficulty of a reading comprehension question, in an absolute and objective manner. This was accomplished by consulting a panel of experts in this domain. Our second contribution is the development of an automated ranking system in order to predict the question difficulty given the question criteria.

We checked different methods of learning the question level given its attributes: Artificial Neural Network, Support Vector Machine, Naïve Bayesian network and decision tree, and we found that their results were relatively close, with a slight advantage on the part of the Artificial Neural Network.

In future work, we intend to reveal which set of criteria is important for the level question learning task, and which solution can be obtained if we concentrate on criteria which can be automatically determined. Moreover, it may be interesting to check whether this set of criteria can help predict the questions' levels in other fields of study.

REFERENCES

[1] R. Azoulay, E. David, D. Hutzler, and M. Avigal. "Adaptation Schemes for Question's Level to be Proposed by Intelligent Tutoring Systems", International Conference on Agents and Artificial Intelligence, France, pp. 245-255. 2014.

[2] C. Snow, Chair, "Reading for Understanding", RAND reading study group. RAND Corporation, 2002.

[3] C. Cornoldi, and J. Oakhill, "Reading Comprehension Difficulties", Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1-13. 1996.

[4] R. Gersten, L. S. Fuchs, J. P. Williams and S. Baker, "Teaching reading comprehension strategies to students with learning disabilities", Review of Educational Research,1998.

[5] C. Dreyer and C. Nel, "Teaching reading strategies and reading comprehension within a technology – enhanced learning environment. Vol. 31, Issue 3. pp. 349-365. 2003.

[6] B. S. Bloom, M. D. Engelhart, E.J. Furst, W. H. Hill, and D. R. Krathwohl, "Taxonomy of educational objectives: The classification of educational goals". Handbook I: Cognitive domain. New York: David McKay Company. 1956.

[7] L. W. Anderson and D. R. Krathwohl, "Taxonomy for learning, teaching and assessing", Longman. ISBN13: 9780801319037, 2000.

[8] F. B. Davis. Fundamental factors of comprehension in reading. Psychometrika, 9, pp. 185-197. 1944.

[9] D. Levi and M. Dalal. "Lashon, havana, v'haba'ah; hachana likrat bagrut kayitz, she'elon 1". [Grammar, comprehension and writing; preparation for summer matriculation exam, questionnaire 1]. Reches Educational Projects, Ltd, 2012.

[10] S. Dreiseitla and L. Ohno-Machadob, "Logistic regression and artificial neural network classification models: a methodology review". Journal of iomedical Informatics 35. pp. 352–359. 2002.

[11] S. S. Abu Naser, "Predicting Learners Performance Using Artificial Neural Networks in Linear Programming Intelligent Tutoring System". International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2., pp. 65-73. 2012.

[12] R. Venkatesh, E. R. Naganathan and N. Umamaheswari, "Intelligent Tutoring System Using Hybrid Expert System With Speech Model in Neural Networks". International Journal of Computer Theory and Engineering, Vol. 2, No. 1, 2010.

[13] A. Yahya and A. Osman, "Automatic Classification of questions into Bloom's cognitive levels using support vector machines", ACIT, website: http://www.acit2k.org/ACIT/index.php?option=com_content&task=view&id=328&Itemid=524 2011.

[14] M. Namba, "Intelligent Tutoring System with Associative Cellular Neural Network". E-Learning Organizational Infrastructure and Tools for Specific Areas. Vol 8 pp. 123-134. 2012.

[15] R. O. Duda and P. E. Hart,"Pattern Classification and Scene Analysis", New-York, John Wliey & Sons, 2.

[16] C. M. Bishop, "Pattern recognition and machine learning, Springer, Information Science and Statistics series, 2006,

[17] J. Hertz, A. Krogh and R.G. Palmer, "Introduction to the ZTheory of Neural Computation", Lecture notes vol. 1, Santa Fe Institute Studies in the sciences of complexity, Addison-Wesley, Redwood City, CA 94065, 1991.

[18] S. Kugblenu, S. Taguchi, S. and T. Okuzawa, "Prediction of the geomagnetic storm associated Dst index using an artificial neural network algorithm". Earth Planets Space, 51., pp. 307–313. 1999.

[19] V. Vapnik, "The nature of statistical Learning Theory", Springer, New York, 1995.

[20] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", ECML. pp. 137-142. 1998.

[21] J.R, Quinlan, "Induction of Decision Trees", Machine Learning Vol. 1 pp. 81-106. 1986.

[22] N. Friedman, D. Geiger and M. Goldszmidt, "Bayesian Network Classifiers", Machine Learning Vol. 29, pp. 1-37. 1997.