e-Examiner: Towards a Fully-Automatic Knowledge Assessment Tool applicable in Adaptive E-Learning Systems

Christian Gütl

Institute for Information Systems and New Media, Graz University of Technology, Graz, Austria

Key words: *Computer-based Assessment, Automated Answer Assessment, Pre-Knowledge Assessment, Knowledge State Assessment, User Modeling, Adaptive E-Learning, AdeLE*

Abstract:

Information about the learners' level or state of knowledge is a key aspect for effective provision of personalized learning activities. E-learning systems gain such information by users' self-assessment, predefined questionnaires and indirectly by inferences on observed user behaviors. However, most present solution approaches either cause a huge effort of manpower or a lack of required reliability. To overcome this, we have developed e-Examiner, a tool for automated knowledge assessment. It is designed to generate questions automatically and to assess student's natural language answers based on the learning content. In this paper, firstly, we introduce the overall concept and architecture. Secondly, we focus on our solution for automated assessment of answers, which is based on the ROUGE toolset for automatic summary evaluation.

1 Introduction

According to [1], community aspects as well as learner-centered, knowledge-centered and assessment-centered aspects are important for and have to be considered in modern learning environments. To narrow down to the assessment aspects, they can be distinguished further in summative assessment, performed at the end of a set of learning activities, and formative assessment, which is intended to give continuous feedback to students and teachers and to enable them to revise teaching and learning activities. Formative assessment results, especially information about the state of knowledge and knowledge acquisition, are needed by e-learning systems to adapt learning activities towards the user [2][3]. Some existing systems try to gain such information by inferences on observed user behaviors. An example for a simple but also wide-spread approach in that context is to analyze the duration and frequency of "consumed" learning content. A more sophisticated approach is to apply information about gaze patterns, which is implemented by the AdeLE system [4]. However, these approaches do not measure directly the actual knowledge acquisition and are more or less error prone. Another approach relies on users' self-assessment but also lacks required reliability. A commonly used approach is the application of questionnaires, composed by limited choice questions, completion tests and open-ended questions [5]. However, it causes huge efforts and costs to prepare tests appropriately to be applicable in computer-supported testing systems. Additionally, tests need to be updated for any changes in the learning content or pedagogic objectives, and personalizeable learning content in adaptive e-learning systems requires a variety of questions, which increases considerably the effort for formative assessment procedures. The situation stated so far led us to develop the e-Examiner, a computer-based assessment tool.

The remainder of the paper is organized as follows: Chapter 2 gives an overview over computer-assisted and computer-based assessment systems, Chapter 3 outlines the solution approach and an overall architecture of the e-Examiner system, Chapter 4 focuses on the prototype implementation of a ROUGE-based short free-text answer assessment, and finally Chapter 5 emphasizes lessons learned and future work.

2 Related Work

Formative assessment is an important part of modern teaching and learning processes in face-to-face courses as well as in e-learning environments. It provides valuable feedback to teachers and students, which allows revising and adapting teaching and learning activities. Furthermore, assessment activities and results could also be utilized for building and strengthen metacognitive skills. [1] However, continuous and frequent assessment in learning processes may cause huge efforts and costs. Therefore, computer-assisted assessment (CaAS) and computer-based assessment systems (CbAS) have become of increasing interest over the years. Such systems may provide parts or the entire chain of the assessment lifecycle, from assessment item authoring and management to specific test compilation to assessment performance and results compilation and management. In addition, emerging interest in sharing and reusing assessment items and assessment tests as well as an open format for exchangeability of assessment results have resulted in the IMS Question & Test Interoperability Specification (IMS QTI). [6] [7] [8]

In particular teachers', trainers' and tutors' increasing workload, caused by intensified formative assessment as well as the application of e-learning-based teaching and training and their need of immediate assessment feedback, has motivated research and development of automated assessment applications. Interesting approaches can be found for various application domains, such as language training [9], mathematics [10], computer science [11] and assessment for numeric disciplines [12]. The authors in [13] describe an interesting approach for automatic, or at least computer-aided, generation of multiple-choice tests from digital learning content. In general, limited choice questions and completion tests are very popular because of their simple methods for automatically assessing students' answers. However, they can not sufficiently assess more abstract educational objectives. [5] [14] To overcome that problem, at least partly, assessment items are applied, which result in students' short free-text answers and essays. Consequently, methods and procedures for automated assessment are much more challenging.

To narrow down to essay assessment, the complex evaluation process includes content aspects and style characteristics as well. In order to gain these aspects automatically, measurements are performed either by indirect characteristics (also termed as "proxes") or by actual dimensions. [8] Automated essay grading (AEG) has been an active research topic since the 1960s and some interesting prototypes and commercial tools have emerged. [15] The Project Essay Grad (PEG) has already started in the early days of AEG research, addresses style aspects and utilizes the concept of proxes and statistical methods (linear regression). The Intelligent Essay Assessor (IEA) focuses on content aspects and applies the latent semantic analysis (LSA) technique. The Bayesian Essay Test Scoring sYstem (BETSY) includes content as well as style aspects to separate assays automatically in 4 point nominal scale using Bayesian Text classification and statistical approaches. [8][16] By further focusing on content aspects of the essay grading process, the author in [17] investigates the applicability for grading on a four point scale of two different text categorization techniques combined with text-complexity features. Firstly, a set of marked reference essays is used to train binary Bayesian Independence Classifiers for distinguishing essay candidates on the four point scale. Secondly, for each level of the four point scale a K-nearest-neighbor classifier built on a probabilistic retrieval system performs the grading by the mean value of K-nearest reference essays. Despite the good results reported in this paper, the major drawback of such statistical approaches is that they rely on bag-of-words but do not include functional relationships between them. In contrast to that, the authors in [18] describe the CarmelTC approach, which focuses on *"correct answer aspects"* by hybrid text classification techniques. It is based on a rule learning text classification and combines results from syntactic functional analyses of text with a bag-of-words classification. A further interesting concept which goes behind statistical analysis of assay answers is discussed in [19]. The authors borrow the Lexical Conceptual Structure (LCS) approach from the machine translation domain for describing the content in a language-independent internal representation (interlingua), and discuss its application for content-based essay assessment.

Unlike the holistic assessment of content and style aspects for essays, the interest in Short Free-text Answer (SFTA) assessment is solely focused on content aspects. Typically, SFTA are written student responses form specific learning and testing activities, such as end-of-the-chapter review questions, classroom tests, and written homework assignment. [20] As opposed to open ended questions, SFTA are results from factual science questions, which can be assessed by objective criterations. [21] In order to perform the challenging assessment task, proposed and implemented solutions span approaches from statistical approaches to applying methods on artificial language to natural language technologies. The authors in [22] discuss the applicability of two simple statistic characteristics: recall for estimating the coverage of test answers compared to the reference answer, and precession for measuring the conciseness of test answers. For their first experiments they apply only the recall measure to assess answers, which results in a surprisingly good performance. However, they do not assess deeper text structure and meaning. A step towards language understanding and more reliable assessment of short answers could be the usage of a controlled natural language, but as one major drawback students are restricted to a limited vocabulary and a specific grammar. [23][24] By focusing on the assessment of natural language answers, the author in [25] describes an interesting approach, which uses a semantic network to represent candidate answers, assesses them against a model answer, identifies wrong and incomplete answers, and provides feedback in natural language. The C-rater tool applies apply a variety of natural language processing methods, such as context-sensitive spelling correction, predicate argument structure and pronominal reference processing, morphological analysis and synonym expansion. This results in a canonical representation of candidate answers and the reference answer, and a rule-based algorithm processes the assessment. [20] WebLAS, see [9], and the assessment engine from Intelligent Assessment Technologies, see [26],) also apply text pre-processing and tagging, and then assesses candidate answers against a number of mark scheme templates. The authors in [21] report about experiments applying further different interesting approaches. Firstly, they adopt an information extraction approach using text pre-processing and tagging, and then apply a set of patterns written and adjusted by hand to the tagged and chunked text for the assessment process. Secondly, they utilize three different machine learning methods, inductive logic programming, decision tree learning and Bayesian learning. A possible drawback of these approaches is that a representative set of training data is needed.

3 Solution Approach and Overall Architecture

Unlike most other existing computer-supported and computer-based assessment tools, we focus on an open and flexible approach, which is designed (1) either to be used as a stand-alone service or to be integrated easily in existing e-learning or other systems, (2) to focus on automated question generation and assessment of short free-text answers in natural language (from several word to several sentences answers) to gain information about user's pre-knowledge and knowledge acquisition in the learning process, (3) to deliver information for updating user models for the purpose of adapting course activities and learning content, and (4) to provide immediate feedback about the state of knowledge and metacognitive skills to learners and teachers.

The requirements stated so far result in the architectural overview as depicted in Figure 1, which is described briefly as follows. The Assessment Test Management unit is the core module of the e-Examiner. It handles the entire lifecycle of the assessment procedures, from question generation to the storage of assessment items (question and reference answer(s)) to the assessment performance. Furthermore, it controls the communication flow with the other modules. The Question Generation module automatically identifies important concepts from a specified learning content, extracts a reference answer from this learning content, and creates simple question on this such as 'explain concept x' or 'describe concept y'. This module can either assist a teacher to create assessment items or it can be utilized to create them unsupervised on-the-fly. The Answer Assessment module assesses short free-text answers automatically and provides information for the Assessment Feedback tool, which prepares helpful feedback to students and teachers. The Interface module handles the information flow between the e-Assessor system and external systems (learning management systems (LMS), standalone assessment services or other systems) by applying the IMS Question & Test Interoperability Specification [7].

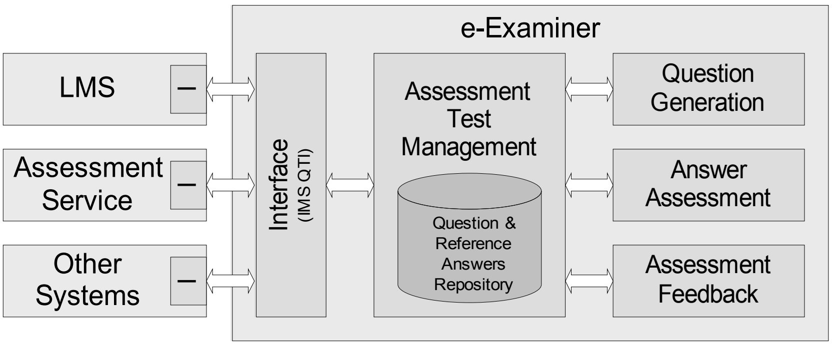

Figure 1: Overall Architecture

4 ROUGE-based Short Free-Text Answer Assessment

Along the line towards our flexible and open assessment system described in the previous chapter, we have developed a first prototype which is applicable to run as a test application, enables experiments and allows gaining first experiences for further research and development cycles. This prototype is designed as a Web-based stand-alone assessment service and focusses on (1) the management of assessment items, (2) the performance of student tests, (3) the automated assessment of short free-text answers, and (4) the immediate result presentation to students and teachers. For automated assessment performance we have decided to apply a hybrid approach. It is built on a natural language pre-processing chain and on ROUGE characteristics, originated for automated evaluation of text summaries [27]. Parts of this chapter are based on [24].

4.1 Prototype Architecture and Implementation

Figure 2 depicts the architecture of our Web-based prototype implementation. Teachers and students can access the system and perform their role-specific assessment tasks by Web Clients. Apache Tomcat [28], an open source JSP and Servlet Container, together with Apache Struts [29], an open source framework for building Servlet/JSP based web applications based on the Model-View-Controller (MVC) design paradigm, host the Web application and provides the Web gateway to the clients. The Control and View component handles the HTTP requests from the client side and the representation of the application. The model in the MVC design paradigm manages user data and roles by the User Management component. It also handles assessment items, assessment tests and performs the automated assessment by the Test Management component. For the purpose of persistent storage of data and states, a Data Storage and Retrieval component based on the open source framework Hibernate [30] was build. The underlying data are managed by the free available MySQL database system [31].

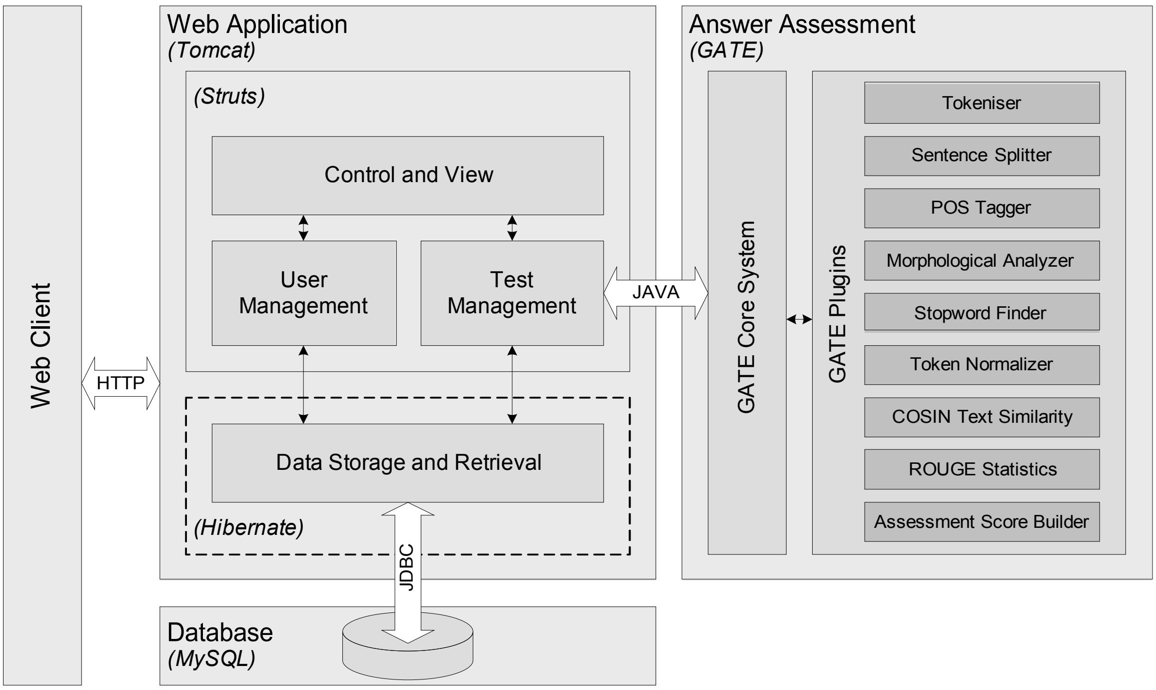

Figure 2: E-Examiner's Prototype Architecture

The automated assessment of short free-text answers is performed by the Answer Assessment component. It is built on GATE, an open source framework for "*developing and deploying software components that process human language"* [GATE User Guide]. One of the strengths of the GATE system is to enable easily the import of text-processing functions by plugins and to define processing chains based on these functions. It is important to note that a large number of pre-existing plugins is available. We have applied some of the ANNIE (A Nearly-New IE system) plugins in order to build the natural language pre-processing chain. This pre-processing chain includes the Tokeniser, the Sentences Splitter, the POS Tagger, the Morphological Analyzer, the Stopword Finder, and the Token Normalizer. For the automated assessment task we have implemented three plugins: the COSIN Text Similarity component estimates the similarity between candidate answers and one or more reference answers based on the vector space model [33]. The ROUGE Statistics component computes a variety of statistical characteristics on the level of word as described in the following subsection. Finally, the Assessment Score Builder component works out the final score by a linear combination of characteristics delivered by the two former mentioned plugins.

4.2 ROUGE Characteristics for Short Free-Text Answer Assessment

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) defines a set of statistical measures to determine the quality of a summary automatically by comparing it with reference summaries. The original intention was to reduce human efforts in the process of evaluating computer-generated summaries. [27]

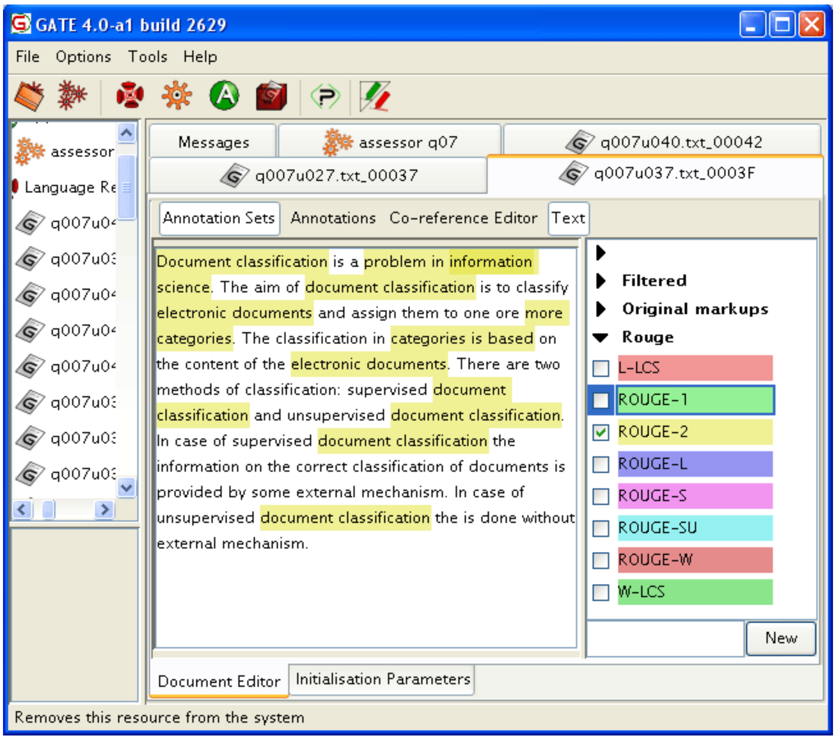

Figure 3: Example of GATE's graphical output for ROUGE-2 characteristic (matching word bigrams) applied on a candidate answer

We were inspired by the ROUGE idea stated so far to study its applicability for short free-text answer assessment. From our point of view, this application scenario is in a sense similar to the original one: students' free-text answers related to learning content can be seen as summaries of this learning content and they are compared with reference answers or reference summaries. The ROUGE metrics are briefly described in the following. A detailed description and discussion can be found in [27] and [34].

- ROUGE-N defines word n-gram co-occurrence statistics between candidate and reference answers where n stands for the length or number of words to be applied for the co-occurrence statistic. To give an example, for n=2 the word bigram statistic is addressed, as illustrated in Figure 3.
- ROUGE-L compares the longest common subsequence (LCS) of words between candidate and reference answers. It is important to note that this similarity measure does not require consecutive matches but focuses on in-sequence matches.

- ROUGE-W stands for weighted longest common subsequence. Unlike ROUGE-L it differentiates several spatial relations in sequences and favors LCS with consecutive matches.
- ROUGE-S describes the skip-bigram co-occurrence statistic. Skip-bigrams are any pairs of words in sentence order which allow arbitrary gaps between them. This metric measures the overlap ratio of skip-bigrams between candidate and reference answers.
- ROUGE-SU is an extension of the ROUGE-S measure. The latter does not give any credit candidate answers which have not any pair of words in common. To overcome this, ROUGE-SU extends ROUGE-S by adding the count of unigram to this statistical measure.

4.3 First Experiment Results

The aim of this section is to report our first results. For our experiments we have collected a dataset for short free-text answer assessment from two different sources, from Institute for Information Systems and New Media (IICM), Graz University of Technology, Austria, and from Faculty of Engineering, Al-Quds University, Jerusalem. The IICM dataset consists of 5 questions related to computer science topic, one reference answer for each of them and 23 sets of student answers. Out of the Al-Quds dataset we have selected three questions also related to a computer science topic and taken 23 answers for each of these questions. Both together, the IICM data and the selected Al-Quds data, are compiled to a dataset, which consist of 8 questions, a reference answer for each of them, and 23 sets of answers. All free-text answers in the dataset are manually assessed by a domain expert according to the reference answers and marked by a number between zero (inappropriate) and ten (very good).

The experiment setup and first results are briefly described in the remainder of this section. Eleven of the 23 sets of answers (88 answers) constitute the set of training data. The remaining 12 sets of answers (96 answers) comprise the set of test data. So far we have focused on three various linear combinations of selected similarity metrics. For each of them the parameters for the linear combination is computed by applying linear regression on the results of the training data set. For the set of test data, the average number of the absolute error and the correlation is computed. A comprehensive view of the experiments and results is given in Table 1. Detailed result plots for the IICM test dataset and the Al-Quds test dataset for one of the experiment setups is given in Figure 4 and Figure 5.

Experiment Setup		Average Absolute Error			Correlation		
No.	**Description of applied Similarity Measures**	**Training Dataset**	**Test Dataset**	**Total Dataset**	**Training Dataset**	**Test Dataset**	**Total Dataset**
1	Number of Tokens and Precision Measure of ROUGE-1	1.47	1.51	1.49	0.80	0.80	0.80
2	Number of Tokens, COSIN Similarity and ROUGE Recall and Precision Measures (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-L, ROUGE-S, ROUGE-SU, ROUGE-W)	1.03	2.00	1.54	0.84	0.79	0.81
3	Number of Tokens, Recall of ROUGE-L, Recall and Precision of ROUGE-1	1.32	1.46	1.39	0.82	0,81	0,81

Table 1: Overview about experiment setup and results

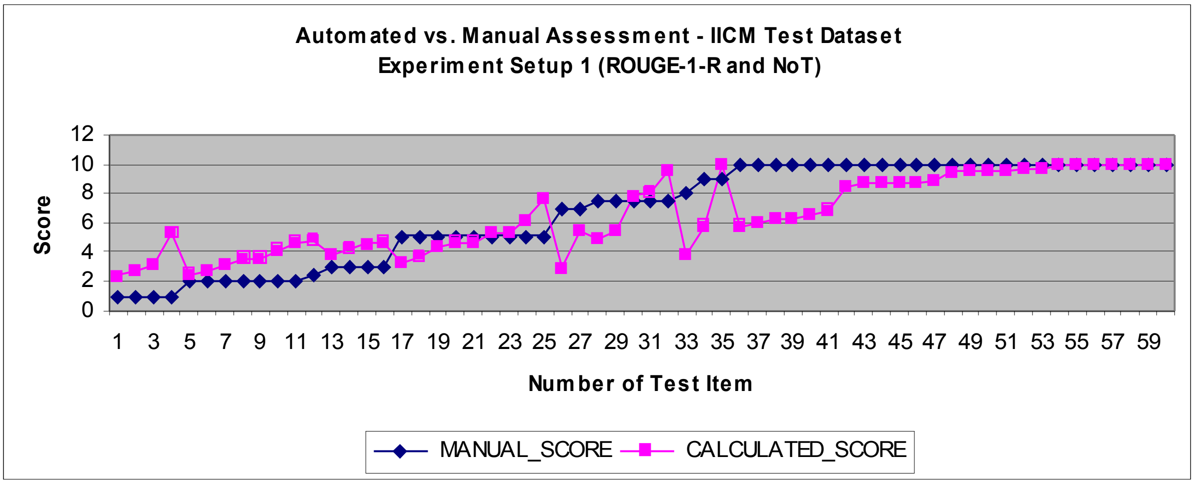

Figure 4: Result plot of experiment setup 1 (ROUGE-1-R and NoT) shows human vs. computer-based scores in ascending order of human scores for the IICM test dataset.

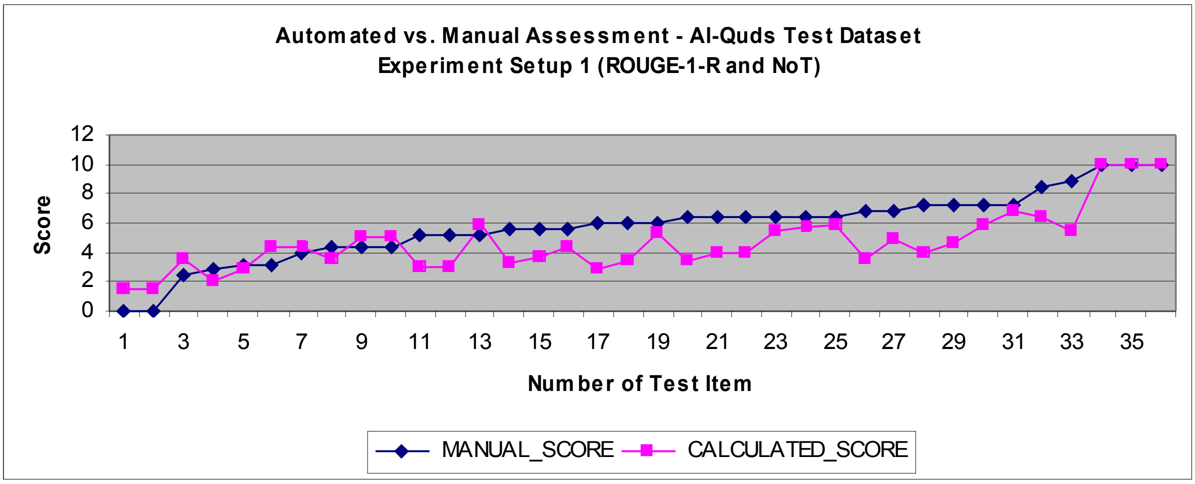

Figure 5: Result plot of experiment setup 1 (ROUGE-1-R and NoT) shows human vs. computer-based scores in ascending order of human scores for the Al-Quds test dataset.

5 Summary and Future Work

In this paper we have shown a variety of computer-assisted and computer-based assessment systems and methods which can support formative assessment activities. It may reduce human workload in the assessment process and it can provide prompt feedback to students, which is of particular interest in the e-learning application domain.

Assessment of short free-text answers enables to evaluate more abstract educational objectives or skills and competences. The automated assessment in that context has become an increasing and active research topic. Various solution approaches, from statistic methods to natural language understanding solutions, are present in contemporary research. For the evaluation process we have applied natural language pre-processing together with statistic ROUGE metrics, originally used for the automatic evaluation of computer-generated summaries. Despite the application of only statistic measures of text similarity, our first experiment results are promising good. We believe that such a system can be used in an e-learning environment for providing automatic feedback to students and for updating user-profiling information as well. However, further experiments by applying a larger dataset and the application of more than one reference answer for each of the test items will be investigated. Furthermore, we also want to adapt and use our approach to evaluate short free-text answers in German language.

Acknowledgement

The author grateful acknowledges Josef Moser for contributing valuable input for this paper and for the prototype implementation as part of his master theses. Furthermore, the author also gratefully acknowledges Labib Arafeh and Yousef Oriquat from Al-Quds University, Jerusalem, for providing a dataset of short free-text answers.

References:

[1] Bransford, J.D., Brown, A.L., & Cocking; R.R. (Eds.) (2000). How People Learn: Brain, Mind, Experience, and School. Expanded Edition. Washington DC: National Academies Press.

[2] De Bra, P., Houben, G.-J., & Wu, H. (1999). AHAM: a Dexter-based reference model for adaptive hypermedia. Proceedings of the tenth ACM Conference on Hypertext and Hypermedia, pp. 147-156.

[3] Brusilovsky, P. (2001). Adaptive hypermedia. User Modeling and User Adapted Interaction, Ten Year Anniversary Issue (Alfred Kobsa, ed.), 11 (1/2), pp. 87-110.

[4] García-Barrios, V.M (2006). Real-Time Learner Modeling: Using Gaze-Tracking in Distributed Adaptive E-Learning Environments, Proceedings of MIPRO, Opatija, Croatia, May 22-26, 2006.

[5] Mödritscher, F., & Sindler, A.: Quizzes are not enough to reach high-level learning objectives! In: KOMMERS, Piet (Edt.) ; RICHARDS, Griff (Edt.): Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005 (ED-MEDIA 2005). Montreal, Canada : AACE, June 2005, pp. 3275–3278.

[6] Valenti, S., Cucchiarelli, A., & Panti, M. (2002). Computer Based Assessment Systems Evaluation via the ISO9126 Quality Model. Journal of Information Technology Education Volume 1 No. 3, 157-176, and http://jite.org/documents/Vol1/v1n3p157-175.pdf

[7] IMS (2005). IMS Question and Test Interoperability Overview. Version 2.0 Final Specification. Last retrieved February 18th 2007 from http://www.imsglobal.org/question/qti_v2p0/imsqti_oviewv2p0.html

[8] Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. Journal of Information Technology Education Volume 2, 319-330, and http://jite.org/documents/Vol2/v2p319-330-30.pdf

[9] Bachman, L.F., Carr, N., Kamei, G., Kim, M., Pan, M.J., Salvador, C., & Sawaki, Y. (2002). A reliable approach to automatic assessment of short answer free responses. In Proceedings of the 19th international conference on Computational linguistics, Taipei, Taiwan, pp. 1-4.

[10] Beevers C.E., & Paterson J.S. (2001): Automatic Assessment of Problem Solving Skills in Mathematics, Maths CAA Series: July 2001, last retrieved February 18th 2007 from http://ltsn.mathstore.ac.uk/articles/maths-caa-series/july2001/index.shtml

[11] Saikkonen, R., Malmi, L., & Korhonen, A. (2001): Fully Automatic Assessment of Programming Exercises. Proceedings of the 6th annual conference on Innovation and technology in computer science education (ITiCSE '01). (pp. 133-136)

[12] Patel, A., Kinshuk, K., & D. Russell, D. (1998). A computer-based intelligent assessment system for numeric disciplines. Information Service, Vo 18, No 1-2, 53—63.

[13] Mitkov, R., & Ha, L. (2003). Computer-aided generation of multiple-choice tests. In Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing, pp. 17-22

[14] Lister, R. (2001). Objectives and objective assessment in CS1. SIGCSE Bull., Vo 33, No 1, pp. 292-296.

[15] Kukich, K. (2000). Beyond Automated Essay Scoring. IEEE Intelligent Systems, September/October 2000, pp. 22 – 27.

[16] Williams, R. (2001). Automated essay grading: An evaluation of four conceptual models. In Proceedings of the 10th Annual Teaching Learning Forum, February 2001, Perth, Western Australia, Last retrieved February 18th 2007 from http://lsn.curtin.edu.au/tlf/tlf2001/williams.html

[17] Larkey, L.S. (1998). Automatic essay grading using text categorization technique. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, pp. 90-95.
[18] Rosé, C.P., Roque, A., Bhembe, D., & VanLehn, K. (2003). A hybrid approach to content analysis for automatic essay grading. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada, pp. 88-90.
[19] Whittington, D., & Hunt, H. (1999). Approaches to the computerized assessment of free text responses. In Proceedings of the 3rd Annual CAA Conference, Loughborough, pp. 207-219.
[20] Leacock, C., & Chodorow, M. (2003). C-rater: Automated Scoring of Short-Answer Questions. Computers and the Humanities, Volume 37, Number 4 / November, 2003, pp. 389-405.
[21] Pulman, S.G., & Sukkarieh, J.Z. (2005). Automatic Short Answer Marking. In Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, Ann Arbor, June 2005, pp. 9–16.
[22] Hirschman, L., Breck, E., Light, M., Burger, J.D., & Ferro, L. (2000). Automated Grading of Short-Answer Tests. IEEE Intelligent Systems, September/October 2000, pp. 22 – 27.
[23] Schwitter, R. (2002). English as a Formal Specification Language. In Proceedings of the 13th International Workshop on Database and Expert Systems Applications. Washington, DC, USA : IEEE Computer Society, 2002, pp. 228–232.
[24] Moser, J.R. (2007). The Electronic Assessor. Automatic Knowledge Assessment in the AdeLE E-Learning Environment. Masters Theses, Graz University of Technology, Graz, Austria, to be published.
[25] Lütticke, R. (2006). Using Semantic Networks for Assessment of Learners' Answers, In Proceedings of ICALT 2006, pp. 1070-1072.
[26] Mitchell, T., Aldridge, N., & Broomhead, P. (2003). Computerised Marking of Short-Answer Free-Text Responses. Manchester IAEA conference, October 2003.
[27] Lin, C.Y (2004) ROUGE: A Package for Automatic Evaluation of Summaries. In Proc. of Workshop on Text Summarization Branches Out, pp. 74-81.
[28] [TOMCAT] APACHE. Apache Tomcat. Official Tomcat Web site, Apache Foundation, last retrieved March 3rd 2007 from http://tomcat.apache.org/
[29] APACHE. Struts Framework. Official Struts Web site, Apache Foundation, last retrieved March 3rd 2007 from http://struts.apache.org/
[30] REDHAT. Hibernate. Official Hibernate Web site, Red Hat Middleware LLC., last retrieved March 3rd 2007 form http://www.hibernate.org/
[31] MYSQL. MySQL Database. Official MySQL Web site, MySQL AB, last retrieved March 3rd 2007 form http://www.mysql.com/
[32] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., & et al. (2007). Developing Language Processing Components with GATE - Version 4 (a User Guide), University of Sheffield, last edited February 8 2007, last retrieved March 3rd 2007 form http://gate.ac.uk/sale/tao/index.html
[33] WIKIPEDIA (2007). Vector space model. Wikipedia, last modified February 18th 2007, last retrieved March 3rd 2007 form http://en.wikipedia.org/wiki/Vector_space_model
[34] Lin, C.Y. (2004). Looking for a Few Good Metrics: Automatic Summarization Evaluation – How Many Samples Are Enough? In Proceedings of the NTCIR Workshop 4, Tokyo, Japan, June 2004, last received March 3rd 2007 form http://research.microsoft.com/~cyl/download/papers/NTCIR4.pdf

Author(s):

Christian, Gütl, Dr.techn. Dipl.Ing.
Graz University of Technology, Institute for Information Systems and New Media,
Inffeldgasse 16c, 8010 Graz, Austria
cguetl@iicm.edu and cguetl@acm.org