An Enhanced Automated Test Item Creation Based on Learners Preferred Concept Space

Mohammad AL-Smadi
Computer Science Department
Jordan University of Science and Technology
P.O.Box: 3030 Irbid 22110, Jordan

Margit Höfler
Institute of Psychology
University of Graz
Graz, Austria

Christian Gütl
Graz University of Technology
Curtin University of Technology, Perth,
WA Graz, Austria

Abstract—**Recently, research has become increasingly interested in developing tools that are able to automatically create test items out of text-based learning contents. Such tools might not only support instructors in creating tests or exams but also learners in self-assessing their learning progress. This paper presents an enhanced automatic question-creation tool (EAQC) that has been recently developed. EAQC extracts the most important key phrases (concepts) out of a textual learning content and automatically creates test items based on these concepts. Moreover, this paper discusses two studies for the evaluation of EAQC application in real learning settings. The first study showed that concepts extracted by the EAQC often but not always reflect the concepts extracted by learners. Learners typically extracted fewer concepts than the EAQC and there was a great inter-individual variation between learners with regard to which concepts they experienced as relevant. Accordingly, the second study investigated whether the functionality of the EAQC can be improved in a way that valid test items are created if the tool was fed with concepts provided by learners. The results showed that the quality of semi-automated creation of test items were satisfactory. Moreover, this depicts the EAQC flexibility in adapting its workflow to the individual needs of the learners.**

Keywords*—*Automated Assessment; Automatic Test-Item Creation; Self-Regulated Learning; Evaluation of CAL systems; Pedagogical issues; Natural-Language Processing

I. Introduction

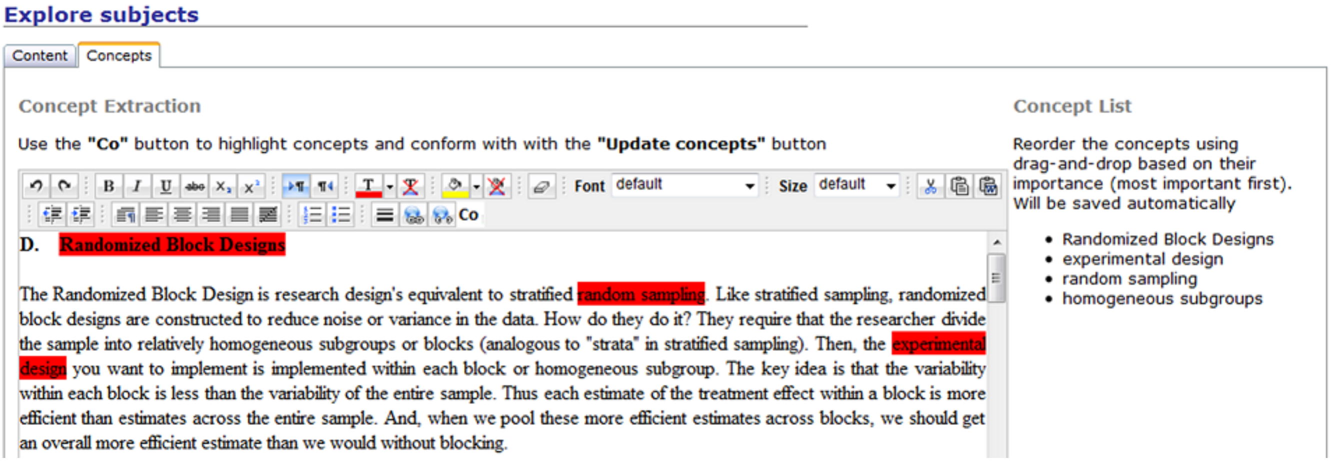
The ability of learners to self-regulate their learning process is a key competence in life-long learning. One efficient way of such self-regulation is to monitor and assess their learning progress by self-assessment (e.g., [1], [2], [3]). Self-assessment supports learners to focus on the most important aspects of the material. Moreover, it helps them to increase their involvement in the learning process [1]. However, learners often face problems if they have to generate questions on their own. For instance, in [4] medical students had to generate questions in order to enhance their metacognition strategies such as self-regulation and problem-solving. The results showed that some of the students had difficulties to create high-order questions (i.e., questions that ask for synthesis and evaluation of the information rather than for simple recall). Furthermore, even if the students had the opportunity to enhance their poorly formulated questions, they needed appropriate guidance to do so. Hence, it might be necessary to train learners in strategies and procedures of self-questioning to increase their competence in question generation [2].

But how can learners receive adequate support in creating good questions? Previous research has presented several guidelines and strategies to support the generation of effective questions (e.g., [5], [6], [7]). For instance, the Taxonomy of Bloom ([8], [9]) provides a powerful framework of how to build questions tailored to different learning goals and levels of cognitive processing (e.g., on a basic level, learners should be able to understand a formula whereas on a more sophisticated level they should be able to apply or explain it). Furthermore, [10] presented examples of guideline questions which might be useful for self-assessment:

1) Can I summarize the main idea of the text?
2) Can I list the five most important learning points of the chapter?
3) Can I write a short comment?
4) Can I discuss the topic raised in this chapter?
5) Are the important learning points I list consistent with those proposed by my classmates and teacher?

What the questions presented above have in common is that they rather address the general ideas and themes of a learning content than its details (e.g., there is no need to know every single detail about a topic if someone has to comment on it). In fact, Bugg and McDaniel [11] showed that conceptual questions addressed to the gist of the text (which required the integration of information across sentences) led to a better performance on a memory test compared to questions that required detailed knowledge about the text. Hence, if questions are generated from a learning content, it might be more valuable to focus on the most important keywords or key phrases than on the details.

According to [12], keywords are defined as a sequence of one or more words and provide a compact description of a document's content whereas key phrases consist of two or more key words and named identities (p.140). In general, key words and key phrases (in the following subsumed to the term concepts) help readers quickly to identify whether a text might be relevant to their needs or not (e.g. [13]). For example, some concepts of the current text (as reflected in the key words section of this paper) are Automated Assessment, and Self-Regulated Learning. Using these concepts, the readers of this journal can quickly identify whether this article might be relevant to their research or not. Likewise, in the context of learning, concepts provide an overview about the learning content and questions that are based on such concepts may



Explore subjects

Concept Extraction

Use the **"Co"** button to highlight concepts and conform with with the **"Update concepts"** button

D. Randomized Block Designs

The Randomized Block Design is research design's equivalent to stratified random sampling. Like stratified sampling, randomized block designs are constructed to reduce noise or variance in the data. How do they do it? They require that the researcher divide the sample into relatively homogeneous subgroups or blocks (analogous to "strata" in stratified sampling). Then, the experimental design you want to implement is implemented within each block or homogeneous subgroup. The key idea is that the variability within each block is less than the variability of the entire sample. Thus each estimate of the treatment effect within a block is more efficient than estimates across the entire sample. And, when we pool these more efficient estimates across blocks, we should get an overall more efficient estimate than we would without blocking.

Concept List

Reorder the concepts using drag-and-drop based on their importance (most important first). Will be saved automatically

- Randomized Block Designs
- experimental design
- random sampling
- homogeneous subgroups

Fig. 1: Screenshot with the learner view form the enhanced EAQC for manual selection of concepts as part of the automated creation of test items.

guarantee that the major themes are covered. However, it is typically left up to the learner to define which concepts reflect a topic best. This suggests that learners might fail to create appropriate questions for self-assessment activities because of inappropriately chosen concepts and the lack of required knowledge and/or skills.

Another approach to support learners in creating questions subsumes software tools that are capable to create questions automatically from a given (text-based) learning content. In recent years, research has become more and more interested in developing such tools (e.g., [14], [10], [15], see also [16], for an overview). However, generating questions automatically from a given text is still a challenge. For instance, such questions often lack meaningfulness, an adequate level of difficulty, and appropriate answers or even relevance (see e.g. [17] , [18], [19]).

A. Problem Statement

Automated question-creation tools typically generate only one type of questions (e.g. open-ended questions that ask for a free answer or multiple-choice questions that ask to find the correct answer among a set of provided distractors). For self-directed learning, having a flexible automated assessment tool that provides different types of questions and preferably also the correct answers has become a need. In addition, these questions should be based on the most relevant concepts of the learning material in order to be effective. However, this research is still active with room for a lot of improvements regarding how the most relevant concepts from natural language texts can be identified (e.g., [12], [13], [20], [21]; a discussion of concept extraction methods can be found in [22] or [23]). Moreover, an effective automated question-creation tool might not only provide questions which reflect the most important concepts of a text best but even allow learners based on their personal needs to determine on which concepts of the text the questions should encounter.

Recently, [24], [25] have presented an enhanced automated question creator (EAQC) tool. EAQC is able to create four types of test items out of English or German learning content. The provided test item types are open-ended items (which require a free answer), single-choice items (in which a given statement is true or false), multiple-choice items (in which one correct answer has to be found within a given number of distractors), and completion exercises (in which one key concept is missing in a statement). Moreover, the authors have conducted an evaluation study in order to investigate the quality of the test items created by EAQC. To this end, they had students rate the test items with regard to different quality criteria adapted from [26]. In addition to pertinence and terminology of test items, these quality criteria evaluate the quality of the answers (open-ended questions and completion exercises) and distractors (multiple-choice items). Preliminary results from the evaluation study [24] showed that EAQC-based test items did not differ in pertinence and level from manually (by an instructor) created test items and also the provided answers were qualitatively on a par with their hand-made counterparts. However, the terminology of some of the test items and the quality of the distractors for the multiple-choice items were rated as rather poor by the participants. Nevertheless, the overall quality of the test items created by the EAQC was satisfactory in a way that the test items were assumed to be of use for self-regulated learning or even in real test settings (see [27]).

Test items generated by EAQC are based on concepts that were extracted out of the learning content. Due to this process, the resulting questions are expected to cover the main aspects of the learning content. However, it is not clear whether all automatically extracted concepts are, at least from the viewpoint of an individual learner, important. In such case, the resulting automatically created test items might cover aspects that are not supportive for the individual learner at all. Previous research [24], [27] showed that learners experienced that not all concepts automatically extracted by EAQC were relevant [24]. This is a first indication that automatically extracted concepts might not always reflect the learner's view.

To the best of our knowledge, there is no study that directly addresses the question of whether the concepts extracted by an automated approach match the same concepts a learner would extract. Hence, it would be of avail if the EAQC

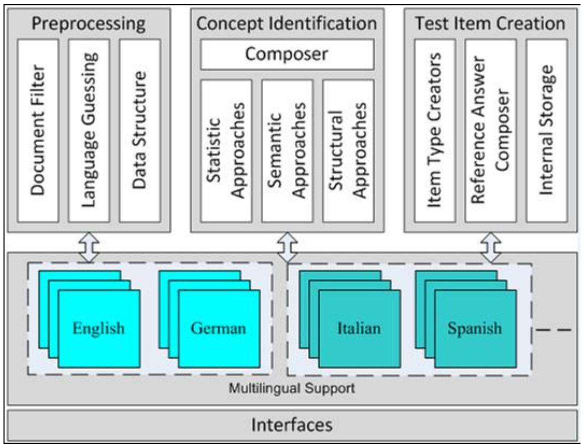


Fig. 2: Conceptual Architecture of EAQC.

provides the functionality to create test items on the basis of concepts which are important for the learner. Therefore, EAQC has been enhanced to enable manual selection of concepts based on learner interest (see Fig. 1). This means that learners are able with the enhanced version of EAQC to manually annotate learning content with the important concepts they aim to learn. Manually selected concepts (based on the annotated learning material) are then used by the EAQC algorithm to automatically create test items. For instance, during the learning process, the learner might highlight the most important concepts from a text and EAQC will use these concepts to automatically create relevant test items. With the resulting test items the learner may then self-assess her/his knowledge or comprehension about the text. Using the EAQC in such individual way has motivated our research interest in whether the EAQC provides such functionality in a sufficient way.

This paper presents an enhanced automated question creator (EAQC) that is able to create test items fully automatically from a textual learning content or semi-automatically based on manually provided concepts. Results show that the overall quality of the test items semi-automatically created by EAQC was comparable to manually created items by humans. The rest of this paper is organized as follows: Section II explains EAQC architecture and EAQC main modules, Section III presents two studies that have been conducted to evaluate EAQC improved functionality, Section IV discusses the studies' main results and findings, whereas, Section V concludes this research.

II. Enhanced Automated Question Creator (EAQC)

EAQC utilizes an automated process to create different types of test items out of textual learning content, more precisely EAQC automatically creates single choice, multiple-choice, completion exercises and open ended test items (questions). EAQC is capable of processing textual learning content stored in various file formats, extracting most important related concepts, creating different types of test items and reference answers that ad-her to the IMS Question & Test Interoperability (QTI) Specification[1]. As depicted in Fig. 2 EAQC architecture supports multilingual test item creation, currently English and German are supported, whereas a flexible extension to other languages such as Italian and Spanish is possible.

From a conceptual point of view, EAQC consists of three main modules [24] (a) the Pre-processing module, (b) the Concept Identification module and (c) the Test Item Creation module. The modules are explained as follows:

1) During the pre-processing module, EAQC detects the input material language (i.e. English or German), performs text cleaning and processing such as special characters and stop words removal, tokenization, and then converts the given learning content into an internal XML format for further processing. Several file formats and online resources are supported by EAQC (e.g. Microsoft Word, Open Document, PDF files, and HTML web content).
2) During the concept identification phase, a syntactic analysis based on part-of-speech tagging (POS) is applied using state-of-the-art natural language processing procedures for the identified language (i.e. English or German). This is followed by statistical analysis of term weighting based on terms co-occurrences. Furthermore, semantic word analysis using WordNet [28] is also performed. Results of the processed information further annotate the XML representation of the learning content. The GATE [29] text processing framework was used as part of this phase. The input text is analyzed into tokens, and then POS tagging, named-entity-recognition (NER), text chunking (noun-based), and co-reference resolution analysis are performed for each token.
   In the statistical analysis, the importance of nouns in the content is estimated and accordingly relevant candidates of word phrases (concepts) are extracted from the learning content. These candidates of word phrases are semantically analyzed using WordNet and prioritized by a relevance number based on the learning content. A configurable threshold value finally defines the concepts in descending order to be used in the third task - i.e. the assessment item creation. Special version of Wordnet is called GermaNet [30] was used for the semantic word analysis on the German text.
3) During test item creation phase, EAQC determines the most appropriate sentence out of the learning content for each of the previously extracted concepts and adds two neighboring sentences to the respective sentence in order to provide sufficient context information. In addition, EAQC computes the distractors for multiple-choice items and the antonyms for (incorrect) single-choice items by also making use of the previously outlined statistic, syntactic and semantic analyses of the concepts. Finally, EAQC creates question items applying a template approach and reference answers for the open-ended items and transforms all resulting items into IMS QTI Specification compliant format.

[1] https://www.imsglobal.org/question/qtiv2p1/imsqti_ oviewv2p1.html

III. Evaluation and Experimentation

Before testing whether EAQC is able to create test items based on manually extracted concepts, the concepts extracted by learners difference and relevance to the concepts extracted by EAQC were investigated. Therefore, in Study 1, data from [24] was reanalyzed. The concepts that participants had extracted from a text and the concepts extracted by EAQC using the same text were compared. Findings from Study 1 reanalysis were then used in Study 2 in order to investigate whether EAQC is able to create valid test items – not only on the basis of automatically extracted concepts but also on the basis of concepts extracted by learners. To this end, EAQC was fed with the manually created concepts from the participants of Study 1 and the resulting (semi-automated created) test items were evaluated by comparing their quality with fully automatically created ones.

A. Study 1

The aim of Study 1 was to investigate whether the concepts extracted by the EAQC from a learning textual material match the concepts extracted by learners from the very same text (Did learners extract similar concepts as the EAQC in terms of content and number?). To this end, data from a study presented in [24] was reanalyzed. The aim of that study was to evaluate the quality of the concepts and test items that were extracted by the EAQC from a text about 'natural-language processing' (NLP). Before participants evaluated the items, their familiarity with the learning topic – their ability to adequately assess the concepts and test items – was assured. Therefore, participants were required to read the text in order to extract - from their viewpoint - the most relevant concepts, then they had to create different types of test items from the text (see below). That is, participants followed similar phases as the EAQC does during test item creation. Comparing the concepts extracted by the participants with the concepts extracted by the EAQC from the same material should allow us to evaluate the level of agreement between the two approaches.

For a better understanding, although the analysis for the concept extraction will be provide afterwards only as outlined above, the full methodology of the study conducted in [24] is presented as follows. Such detailed description of the methodology of Study 1 is also important since the method of the subsequent study (Study 2) is fairly similar to the method of Study 1.

1) Participants: 29 participants (4 female) took part in this study. They were 25.4 years on average (SD = 3.3), ranging from 22 to 39 years. All of them were technical students. 93.1% of them already had a bachelor degree. The experiment took place within a regular course Information Search and Retrieval at Graz University of Technology. Because of the restricted number of computer-work places, the participants were divided in two groups (14 and 15 participants each, respectively). The two groups were tested separately on two consecutive days. All participants gave informed consent.

2) Stimuli and Procedure: In advance of the study, EAQC was used to create test items from a learning content about 'Natural-Language Processing' (NLP). The text was taken (with slight changes) from Wikipedia[2] and consisted of approximately 2,600 words by that date. EAQC extracted 49 main concepts from this text. Example concepts were, for instance, "natural language processing"; "modern NLP algorithms", and "the Georgetown experiment". The concepts were automatically ranked with regard to their relevance by means of statistical and semantic analyses of the content (i.e., the first concept extracted by EAQC was statistically the most relevant one etc.; see Table 2, left column for the 10 most relevant concepts extracted by EAQC). For each of these concepts, EAQC then created four types of test items (one open-ended item, one completion exercise, one single-choice item and one multiple-choice item, respectively) and the respective answers. In the following, the resulting questions and answers for the concept "modern NLP algorithms" are presented (please note that for the sake of brevity the answer for open-ended items is not provided):

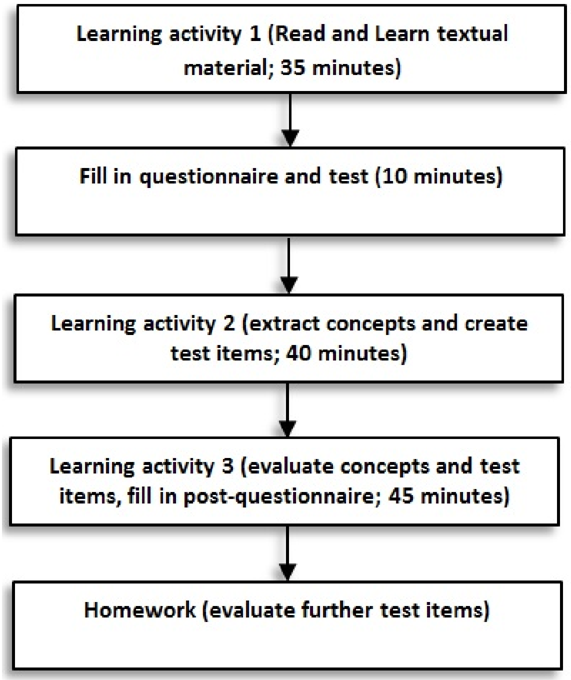

Learning activity 1 (Read and Learn textual material; 35 minutes)
Fill in questionnaire and test (10 minutes)
Learning activity 2 (extract concepts and create test items; 40 minutes)
Learning activity 3 (evaluate concepts and test items, fill in post-questionnaire; 45 minutes)
Homework (evaluate further test items)

Fig. 3: Flowchart explains the procedure of the experiment.

1) **Open-ended item**:
Example: What do you know about modern NLP algorithms in the context of Natural language processing?
Region of answer: (...)
2) **Completion exercise**:
Example: (...) are grounded in machine learning, especially statistical machine learning.
Answer: "modern NLP algorithms".
3) **Single-choice item**:
Example: Old style NLP algorithms are grounded in machine learning, especially statistical machine learning.
Answer: False.
4) **Multiple-choice item**:
Example: (...)are grounded in machine learning, es-

[2] NLP:http://en.wikipedia.org/wiki/Natural_ language_ processing

pecially statistical machine learning.

a) Answer 1: meta-rule NLP algorithm
b) Answer 2: algorithmic program NLP algorithms
c) Answer 3: modern NLP algorithms
d) Answer 4: heuristic NLP algorithms

Together, the EAQC created 196 test items in total (49 concepts × 4 test-item types). However, not all of the test items were evaluated during the study. In order to reduce the time and effort for the participants, they evaluated only 80 test items which were based on the 20 most relevant extracted concepts (i.e., 20 concepts × 4 test-item type = 80 test items). In addition, 24 test-items (six per test-item type) were provided. These items had been extracted by one of the authors in advance for checking purpose. Hence, each participant evaluated 104 test items in total (24 during the study session and 80 as a homework assignment). Furthermore, participants evaluated the relevance of the 49 automatically extracted concepts (and seven manually, by one of the authors, extracted concepts). The learning content (i.e., the text), the questionnaires for the evaluation tasks, and also the instructions were presented in English as web-based content. To collect the data, participants were asked to fill in five online questionnaires that were delivered successively at specific times during the study (see below). For creating these questionnaires LimeSurvey[3] was used. Participants were not aware about the immediate purpose of the study (i.e., evaluation of automatically extracted concepts and test items) but were told that they have to take part in several learning activities during the session. Crucially, they were not informed that most of the concepts and test items for evaluation were based on EAQC. Hence, participants did not know the source (i.e., if they were created by EAQC or human). Results from the tests delivered during the experiment were part of the final grading of the course. However, the participation in the experiment was not a prerequisite to complete the course successfully.

The study session consisted of several phases (see Fig. 3). At the beginning of the session, an outline about the session was presented by one of the authors. Then the learning content was presented online (web-based) to each participant. Participants were then asked to learn the prepared text about 'Natural-Language Processing' (Learning activity 1). After that, they had to fill in a questionnaire that was created with LimeSurvey in advance and sent to them via email. In this questionnaire participants were asked not only for specific demographic data such as age, previous education etc. but also to briefly summarize the text they had learned before. Then, after a short break, and crucially for the present paper, participants were required to extract relevant concepts from the text and to create eight test items (two of each test-item type as described before) using their extracted concepts (Learning activity 2). This learning activity aimed at ensuring the participants' familiarity with the text. The activity lasted about 40 minutes. The text was accessible to the students during this task and they were allowed to take notes if necessary. Participants inserted their answers into a new questionnaire that was delivered via email. After they had completed Learning activity 2, participants had to attend a second test in which they had to answer eight multiple choice items regarding the learning content.

After a further break, Learning activity 3 started. In this learning activity, participants were asked to evaluate the relevance of the 56 concepts (49 automatically extracted concepts and seven manually for control) using a 5-point Likert scale (5 = very relevant; 1 = not relevant at all). Participants were also asked to evaluate the 24 test items ((5 = very good; 1 = very bad) regarding their pertinence (i.e., relevance of the test item with respect to the major themes of the text), level (i.e., Is the test items trivial or does it expresses a significant meaning?), and terminology (appropriateness of the chosen words; see [26]). In addition, if an answer was provided, they had to evaluate the quality of the answer (i.e., Is the provided answer relevant?) and, in case of multiple-choice items, the quality of the distractors was also evaluated. The order of the concepts and test items to be evaluated was randomized. Finally, after completion of Learning activity 3, participants had to fill in a post-questionnaire in which they were asked to answer more general questions about the task (e.g. general attitudes regarding the different test-item types). In total, the experiment lasted approximately three hours. In addition, the evaluation of further test items (16 EAQC-based and four manually created test items per test-item type, resulting in 80 test items in total) was set for homework.

3) Results and Findings: Results of the evaluation task and test performance are presented in [24], [27]. For the purpose of this paper, only the concepts which the participants extracted during the experimental session (Learning activity 2) were needed.

All in all, the participants extracted 153 different concepts (491 in total) and 17.1 on average (SD = 10.3; ranging from 5 to 41 concepts per participant). Hence, participants typically extracted fewer concepts than the EAQC (49 concepts). Table 1 (middle column) depicts the 10 most frequently extracted concepts by the participants and the 10 statistically most relevant concepts extracted by the EAQC (left column). More than 90% of the students extracted ”machine learning” or ”natural language processing” whereas about 38% chose ”named entity recognition”. When two independent raters were asked to rate which of the concepts extracted by EAQC perfectly match a concept extracted by the students they agreed only on 9 (out of 147) perfect matches (e.g., ”natural language processing”, ”machine learning”); in further 57 cases there was some disagreement between the raters because the concepts matched at least partially (e.g., ”word/text segmentation” and ”evaluation”). However, most of the time (in 87 cases) participants extracted concepts that were not considered by EAQC (e.g., ”parsing”, ”word sense disambiguation”).

In sum, the findings of Study 1 suggest that there is some overlap between the concepts created by the EAQC and manually extracted concepts. However, students often also extracted concepts that were not considered by the tool. In addition, they considered fewer concepts as relevant as the tool. Hence, not all concepts extracted by the EAQC might be experienced as important for the individual learner and some perhaps relevant concepts are even missed. It may be therefore useful to allow the individual learner to enter his or her own concepts to the EAQC in order to receive appropriate test items. Therefore the EAQC was improved in order to support such functionality. In

[3] http://www.limesurvey.org/

TABLE I: Most important concepts extracted by the EAQC (left column) and by the participants in study 1(middle column; percentage of naming in parentheses). For a better comparison, also the concepts extracted by the participants of Study 2 are presented (right column).

Concepts by EAQC	Study 1: Concepts by students	Study 2: Concepts by students
natural language processing	machine learning (96.5%)	natural language processing (87.5%)
modern NLP algorithms	natural language processing (93.1%)	machine learning (75.0%)
languages text segmentation	part-of-speech tagging (72.4%)	artificial intelligence (50.0%)
the first statistical machine translation systems	NLP evaluation (65.5%)	linguistics (37.5%)
linear algebra and optimization theory	parsing (48.3%)	NLP evaluation (37.5%)
computer science and linguistics	word sense disambiguation (48.3%)	NLP tasks (37.5%)
machine learning	statistical NLP (41.4%)	Turing test (25%)
the Georgetown experiment	word segmentation (41.4%)	hand-written rules (25%)
evaluation metrics	topic segmentation and recognition (41.4%)	fully automatic translation (25%)
an evaluation step	Named entity recognition (37.9%)	statistical NLP (25%)

such scenario, user-defined concepts are used instead of the automatically extracted concepts. The remainder of the test-item creation procedure remains the same.

B. Study 2

To investigate the quality of the test items that are based on manually provided concepts, a second study was conducted. In this second study semi-automated test items (i.e., automatically generated test items that were based on manually extracted concepts) with fully automated test items (i.e., test items that were based on concepts extracted by the EAQC) were compared. Furthermore, a sample of test items for control that were created completely manually was included.

1) Participants: Eight participants took part in this study (2 out of them were females). Participants were 33.1 years on average (SD = 6.6), ranging from 25 to 41 years. 87.5% of them were PhD students in computer science and 12.5% were master students at Graz University of Technology. All participants gave informed consent.

2) Apparatus and Stimuli: Apparatus and Stimuli were the same as in the first study except for the tool enhancement of processing user-defined concepts for the test item creation process. To create the test items using EAQC, the concepts provided by EAQC and the concepts that had been extracted by the participants during the first study were used. In particular, the 10 – out of 15 – most frequently extracted concepts from the participants in the first study were used to create test items using EAQC. Some of the participants extracted concepts had to be slightly rephrased in order to create test items automatically out of them as there was no fuzzy matching mechanism implemented in that version of EAQC.

In total, 120 test items were presented based on three categories with 40 test items each (10 per each of the four test-item types). The first category included test items fully provided by the EAQC (EAQC-a). These items were the same as in the first study. The second category consisted of 40 items created by the EAQC based on the concepts that had been extracted by the participants in the first study (EAQC-m). Finally, the third category (manual) included a randomly selected sample of the test items that had been created by the participants during the second learning activity in Study 1 (see above). For this latter category only fully elaborated test items were considered (i.e. in which the participants of Study 1 provided not only the question but also the respective - correct - answer).

3) Procedure: The procedure of Study 2 was similar to the first study with the following exceptions. The study was not conducted during a course but participants attended the study as an online session. In addition, participants had to evaluate 120 test items in total. There was no time restriction for the tasks although participants were given the same guidelines as described in the first study (see Fig. 1). Participants were also asked to complete a further questionnaire in which they had to rate the quality of multiple-choice questions distractors. As this task was of no relevance for the aims of the current study, results of this questionnaire are not reported here.

4) Results and Findings: First, the concepts that were extracted by the participants were analyzed and compared with the concepts extracted by the participants in Study 1 and the concepts extracted by EAQC. Participants in Study 2 extracted 53 different concepts (100 in total) and 12.5 on average (SD = 8.7; ranging from 3 to 24 concepts per participant). Table 1 shows the 10 most frequently extracted concepts, by EAQC (which were the same for both studies in column 1), by the participants during the first study (column 2), and by the participants during the second study (column 3). Despite the fact that fewer participants took part in Study 2, the concepts were quite similar to the concepts extracted by the participants Study 1. That is, in both studies, the most important concepts were ”natural language processing” and ”machine learning”. Also ”NLP evaluation” and ”statistical NLP” were mentioned by both groups. However, there were also concepts that were only considered by the participants of one study (e.g., ”parsing” and ”history of NLP” only in the first study vs. ”artificial intelligence” and ”linguistics” only in the second study). This could be - at least to some extent - attributed to differences in prior knowledge between the two groups (note that most of the participants in the second study were PhD students) and is further evidence that learners differ slightly in their individual views of which concepts are relevant.

For each test-item type, the quality of 10 EAQC-created items based on automatically extracted concepts (EAQC-a), 10 EAQC-created items based on manually extracted concepts (EAQC-m), and 10 fully manually created test items were analyzed. The evaluation criteria were the same as in the first study (see [26]). That is, participants evaluated the

TABLE II: Mean ratings for test items: automatically created based on automatically extracted concepts (EAQC-a), automatically created based on manually extracted concepts (EAQC-m), and manually created test items (Manually) for each test-item type with regard to the evaluation criteria. Standard deviations are presented in parentheses.

Question Type	Approach	Pertinence	Terminology	Level	Answer	Distractors
Completion Exercises	EAQC-a	3.7 (0.7)	3.6 (0.7)	3.4 (0.7)	3.7 (0.7)	-
	EAQC-m	3.6 (0.7)	3.6 (0.7)	3.4 (0.6)	3.7 (0.8)	-
	Manually	3.7 (0.9)	3.7 (0.7)	3.6 (0.8)	3.7 (0.7)	-
Single-Choice Items	EAQC-a	3.7 (0.7)	3.8 (0.8)	3.7 (0.7)	-	-
	EAQC-m	3.5 (1.0)	3.5 (0.5)	3.4 (0.8)	-	-
	Manually	3.3 (0.8)	3.3 (0.8)	3.2 (0.8)	-	-
Open-Ended Items	EAQC-a	3.9 (0.7)	3.6 (0.7)	3.9 (0.5)	3.6 (0.6)	-
	EAQC-m	3.9 (0.7)	3.8 (0.6)	3.9 (0.6)	3.6 (0.6)	-
	Manually	4.2 (0.7)	4.2 (0.7)	4.0 (0.6)	3.8 (0.6)	-
Multiple-Choice Items	EAQC-a	3.6 (0.6)	3.6 (0.6)	3.3 (0.7)	3.5 (0.5)	3.1 (0.8)
	EAQC-m	3.6 (0.8)	3.5 (0.7)	3.1 (0.7)	3.4 (0.7)	2.9 (0.8)
	Manually	3.8 (0.7)	3.9 (0.6)	3.6 (0.8)	3.9 (0.8)	3.9 (0.7)

pertinence, the terminology and the level of the test items as well as, when appropriate, the relevance and quality of the answers and distractors. Table 2 shows mean ratings (1 = very bad; 5 = very good) for each test-item type with regard to the evaluation criteria. In order to investigate possible quality differences between the three item's sources (EAQC-a; EAQC-m, and manual, respectively) regarding the evaluation criteria repeated-measures ANOVAs separately for each test-item type were computed. In case the assumption of sphericity was violated, the Greenhouse-Geisser correction was used to correct for degrees of freedom.

For completion exercises, the ANOVA showed no effect of item source, F <1, but a main effect of evaluation criteria, F(3, 21) = 4.66, p <0.05, $\eta_p^2 = 0.40$, and no interaction, F <1. This finding suggests that the quality of both categories of automatically created completions exercises was comparable to the manually created items.

For single-choice items, a main effect of item source was found, F(2, 14) = 7.78, p $<.01$, $\eta_p^2 = 0.53$, but no effect of evaluation criteria, F(2, 14) = 2.94, p = .09, and no interaction, F <1 were found. Post-hoc analysis (Bonferroni corrected) showed that manually created items were evaluated even worse compared to automatically created items based on automatically extracted concept (p $<.05$). No difference in quality was found between the manually created items and EAQC-m items and between the two types of automatically created items.

For open-ended items, a main effect of item source was found, F(2, 14) = 5.88, p $<.05$, $\eta_p^2 = 0.46$, and a main effect of evaluation criteria as well, F(1.33, 9.33) = 4.74, p $<.05$, $\eta_p^2 = 0.40$. There was no interaction, F(1.69, 11.82) = 1.01, p = .38. Post-hoc analysis showed that manually created items were evaluated better compared to EAQC-m items. There was also a tendency (p = .09) that manually generated test items were evaluated better than EAQC-a items. There was again no difference between EAQC-m and EAQC-a items.

Finally, for multiple-choice items, a main effect of item source was found, F(2, 14) = 7.58, p $<.01$, $\eta_p^2 = 0.52$, and also a main effect of evaluation criteria, F(4, 28) = 6.13, p $<.01$, $\eta_p^2 = 0.47$. The interaction was also significant, F(8, 56) = 2.70, p $<.05$, $\eta_p^2 = 0.28$. Post-hoc analysis showed that manually created items were again evaluated better as compared to EAQC-m items (p $<.05$). There was no such difference between manually and EAQC-a items and between EAQC-a and EAQC-m items.

IV. Discussion

In order to evaluate the quality of EAQC automatically extracted concepts and test items two studies were conducted. In the first study, whether and to what extent learners might extract the same concepts as EAQC out of a learning content were tested. Results showed that there was an overlap between the manually and automatically extracted concepts. In addition, participants extracted on average fewer concepts than EAQC. Together, this suggested that not all concepts provided by EAQC might be relevant for learners. Therefore, the second study tested whether it is possible to personalize EAQC in such a way that the tool has the functionality to create valid test items out of concepts that were manually selected and inserted to the tool. To this end, EAQC was fed with manually extracted concepts and compared the quality of the resulting test items with test items that were either fully automatically or fully manually created.

Results showed that, in general, the quality of both the semi- and fully automatically created test items was comparable to manually created ones. In particular, for completion exercises, both types of automatically created items did not differ from the manually created test items with regard to various quality criteria such as pertinence, terminology, and level. For single choice items, the automatically created test items were evaluated even slightly better than the manually created items. However, there were two test-item types in which manually created items outperformed their (fully and semi-) automated created counterparts: open-ended items and multiple-choice items. For multiple-choice items, the results might be caused by the relatively low quality of the distractors created by EAQC. In the same time for multiple-choice items the participants'ratings for manually and (both fully and semi-) categories differed only slightly with regard to all other quality criteria (pertinence, terminology, level of the test items and the relevance of the answer). The average ratings for the quality of EAQC-based distractors were approximately one point lower than for the manually created distractors (see Table 2; note that, because of a possible lack of statistical power, the respective post-test did not yield a significant difference). Anecdotal reports of the students also suggest that the distractors were sometimes too easy. They stated that learners might be able to guess the correct answers by simply excluding the obviously inappropriate alternatives.

Previous research has shown that creating appropriate and valid distractor items is in fact demanding (e.g. [31], [32]). The creation of appropriate distractors can be difficult even if the distractors are created manually. For instance, DiBattista and Kurzawa [33] showed that many distractors created for various classroom tests were flawed and needed revision. Creating meaningful distractors using question-creation tools is even more challenging (e.g., [17]). In general, good distractors should be as semantically close to the correct answer as possible [34]. Currently, the EAQC uses antonyms and related terms of the respective concept in order to compute distractors. This approach might, however, lead to the creation of distractors that are clearly odd when presented in a specific context (e.g., correct answer: ”speech tagging”; example distractor computed by the EAQC: ”tongue-lashing tagging”). Thus, further experimentation is necessary to improve the quality of distractors automatic creation as part of multiple-choice questions.

With regard to the open-ended items, the cause of the worse quality of both types of automatically created test items as compared to manually created test items is not that clear as for the multiple-choice items. One reason for the poor evaluation of the quality of these items might be the ”uniform” terminology of the main phrase that scaffolds open-ended items. That is, for automatically created test items, the standard phrasing used was ”What do you know about [concept] in the context of (or subjected to) NLP?”. In contrast, the structure of the manually created items was more diversified (e.g., ”Explain......”, ”What is [the difference between] ...?”, ”Describe......”). Perhaps this made the test items more interesting and less ”artificial” for the participants. However, this does not mean that participants were not able to understand the meaning of the questions created by EAQC. The results suggest that the automatically created test items were nevertheless syntactically valid. Hence, even if the open-ended automatically created questions may be less quality than human created ones, they still can be used for self-assessment without difficulty.

In Study 2, test items that were either based on fully automatically extracted concepts or on manually provided ones have been evaluated. Note that, as the content of the automatically and manually extracted concepts sometimes overlapped –i.e. they use similar terminology (see Table 1) – this had created typically similar test items (e.g., EAQC created the same test items for the concept ’natural-language processing’ which had been extracted by both the EAQC and the participants in Study 1). It can be assumed that such overlap between test items of different item categories (semi automatically vs. fully automatically) led to similar ratings of the respective test items, and hence, the lack of finding differences between the quality of these two test item categories. From a technological and methodological point of view, such finding reflects the validity of EAQC architecture, when the same concepts had led to the creation of the same test items. However, from a pedagogical point of view, a possible enhancement of the tool might be that other alternative items are computed from the same concept. Such variation in the test items might prevent having the same test item being presented again and again when learners self-assess their learning progress more than once. Moreover, it might prevent learners of becoming bored or even being less challenged during learning.

The number of concepts extracted by the participants in both studies varied from about five concepts per person to more than 40 concepts per person. Hence, there were large differences between individual learners with regard to how many concepts they considered as relevant. Noticeably, the EAQC extracted far more concepts (49 concepts) from the same learning content than the average learner did. Thus, there might be a mismatch between the ”statistical” relevance of a concept and its relevance from the learner's individual point of view. In Study 1, only 9 (i.e., 6.1%) out of the 147 concepts that were extracted by the students in total were perfectly consistent with the concepts extracted by the EAQC. The same analysis for Study 2 revealed a perfect match for only 2 (i.e., 3.8%) out of 53 concepts. This is insofar critical as the test items that result on these concepts might then also be often ”worthless” for a learner.

The EAQC already provides the opportunity to reject less important concepts after the concept-extraction phase (i.e., before starting the actual test-item creation, see Fig. 1). Therefore, learners may read through all concepts the EAQC has provided and have the EAQC create test items only from the most appropriate concepts. This functionality (which was purposefully not used in the present studies) may already help to improve the quality of the automatically created test items. In addition, it is also possible to reduce the number of concepts that the EAQC is supposed to extract from the learning content before the concept-extraction phase (i.e., the learner can determine the number of concepts that should be extracted).

From a pedagogical point of view, it might not always make sense to reduce the number of concepts in advance, because such ”less relevant” concepts might sometimes cover aspects that the learner had missed so far during learning. Rather, it might be suitable to have EAQC first extract the concepts based on a statistical analysis from which the learner can then select the most appropriate ones. This process of reflecting on the concepts extracted by EAQC might help learners to deepen their understanding of the text. In a final step, learners should then be allowed to include concepts that are still missing from their viewpoint. Based on these three steps, the resulting test items created by EAQC should then cover the main topics of the learning content which in turn are assumed to support learners efficiently. This paper shows that EAQC is partially able to support this functionality outlined above. The tool allows selecting specific concepts from the extracted ones and it was able to create test items from 10 out of 15 manually inserted concepts. However, some of the concepts had to be slightly rephrased in order to receive valid test items. Such rephrasing is of course quite time-consuming. Thus, further improvements for fuzzy term matching and synonym and hypernym relations are considered in this regard.

V. CONCLUSION

In the context of self-regulated learning learners often lack adequate feedback of whether they have covered all aspects of a topic or whether the important points they have extracted are consistent with the points proposed by colleagues or instructors [10]. As a consequence, they often have problems to create appropriate test items for self-assessment.

In this paper an enhanced automated question creator (EAQC) is presented. EAQC is able to either create test items fully automatically from a learning content or on the basis of manually provided concepts. Two studies evaluating EAQC improved functionality were discussed. The results showed that the overall quality of the test items semi-automatically created by EAQC was comparable to manually created items. These findings are a further step in developing a tool that is able to effectively support learners during self-regulated learning and self-assessment activities.

REFERENCES

[1] T. Janssen, "Instruction in self-questioning as a literary reading strategy: An exploration of empirical research," *L1-Educational Studies in Language and Literature*, vol. 2, no. 2, pp. 95–120, 2002.

[2] A. King, "Effects of self-questioning training on college students' comprehension of lectures," *Contemporary Educational Psychology*, vol. 14, no. 4, pp. 366–381, 1989.

[3] M. Scardamalia and C. Bereiter, "Text-based and knowledge based questioning by children," *Cognition and instruction*, vol. 9, no. 3, pp. 177–199, 1992.

[4] T. Papinczak, R. Peterson, A. S. Babri, K. Ward, V. Kippers, and D. Wilkinson, "Using student-generated questions for student-centred assessment," *Assessment & Evaluation in Higher Education*, vol. 37, no. 4, pp. 439–452, 2012.

[5] D. Buehl, *Classroom strategies for interactive learning*. International Reading Assoc., 2013.

[6] N. Silveira, "Towards a framework for question generation," in *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, 2008.

[7] A. C. Graesser, S. E. Gordon, and L. E. Brainerd, "Quest: A model of question answering," *Computers & Mathematics with Applications*, vol. 23, no. 6, pp. 733–745, 1992.

[8] B. S. Bloom and M. D. Engelhart, *Taxonomy of Educational Objectives: The Classification of Educational Goals: By a Committee of College and University Examiners: Handbook 1*. David McKay, 1969.

[9] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon, 2001.

[10] C. K. E. Cheng, "The role of self-regulated learning in enhancing learning performance," 2011.

[11] J. M. Bugg and M. A. McDaniel, "Selective benefits of question self-generation and answering for remembering expository text." *Journal of educational psychology*, vol. 104, no. 4, p. 922, 2012.

[12] M. Dostál and K. Jezek, "Automatic keyphrase extraction based on nlp and statistical methods." in *DATESO*, 2011, pp. 140–145.

[13] Y. Lui, R. Brent, and A. Calinescu, "Extracting significant phrases from text," in *Advanced Information Networking and Applications Workshops, 2007, AINAW '07. 21st International Conference on*, vol. 1, May 2007, pp. 361–366.

[14] M. Agarwal, R. Shah, and P. Mannem, "Automatic question generation using discourse cues," in *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, ser. IUNLPBEA '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1–9. [Online]. Available: http://dl.acm.org/citation.cfm?id=2043132.2043133

[15] D. L. Lindberg, "Automatic question generation from text for self-directed learning," Ph.D. dissertation, Applied Sciences: School of Computing Science, 2013.

[16] P. Piwek and K. E. Boyer, "Varieties of question generation: introduction to this special issue," *Dialogue and Discourse*, vol. 3, pp. 1–9, 2012.

[17] 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 (KM&EL)*, vol. 2, no. 3, pp. 210–224, 2010.

[18] L. Vanderwende, "The importance of being important: Question generation," in *Proceedings of the 1st Workshop on the Question Generation Shared Task Evaluation Challenge, Arlington, VA*, 2008.

[19] O. Sitthisak, L. Gilbert, and H. Davis, "An evaluation of pedagogical informed parameterized questions for self-assessment," *Learning, Media and Technology*, vol. 33, no. 3, pp. 235–248, 2008.

[20] J. Villalon and R. A. Calvo, "Concept extraction from student essays, towards concept map mining," in *2009 Ninth IEEE International Conference on Advanced Learning Technologies*. IEEE, 2009, pp. 221–225.

[21] W. J, "Extraction of relevant semantic data from natural language texts in the view of automatic question generation," 2010.

[22] N. Liu and C. C. Yang, "Keyphrase extraction for labeling a website topic hierarchy," in *Proceedings of the 11th International Conference on Electronic Commerce*. ACM, 2009, pp. 81–88.

[23] E. Hovy, Z. Kozareva, and E. Riloff, "Toward completeness in concept extraction and classification," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, 2009, pp. 948–957.

[24] C. Gütl, K. Lankmayr, J. Weinhofer, and M. Hofler, "Enhanced automatic question creator–eaqc: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education." *Electronic Journal of e-Learning*, vol. 9, no. 1, pp. 23–38, 2011.

[25] M. Al-Smadi and C. Guetl, "Supporting self-regulated learners with formative assessments using automatically created qti-questions," in *Global Engineering Education Conference (EDUCON), 2011 IEEE*, April 2011, pp. 288–294.

[26] S. Cannella, E. Ciancimino, and M. López Campos, "Mixed e-assessment: An application of the studentgenerated question technique," in *Education Engineering (EDUCON), 2010 IEEE*, April 2010, pp. 769–773.

[27] M. Höfler, M. AL-Smadi, and C. Gütl, "Investigating content quality of automatically and manually generated questions to support self-directed learning," in *CAA 2011 International Conference, University of Southampton*, 2011.

[28] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[29] H. Cunningham, "Gate, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, no. 2, pp. 223–254, 2002.

[30] B. Hamp, H. Feldweg *et al.*, "Germanet-a lexical-semantic net for german," in *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Citeseer, 1997, pp. 9–15.

[31] T. M. Haladyna, *Developing and validating multiple-choice test items*. Routledge, 2012.

[32] J. R. Moser, C. Gütl, and W. Liu, "Refined distractor generation with lsa and stylometry for automated multiple choice question generation," in *AI 2012: Advances in Artificial Intelligence*. Springer, 2012, pp. 95–106.

[33] D. DiBattista and L. Kurzawa, "Examination of the quality of multiple-choice items on classroom tests." *Canadian Journal for the Scholarship of Teaching and Learning*, vol. 2, no. 2, p. 4, 2011.

[34] R. Mitkov, H. LE AN, and N. Karamanis, "A computer-aided environment for generating multiple-choice test items," *Natural Language Engineering*, vol. 12, no. 02, pp. 177–194, 2006.