Enhanced Automatic Question Creator – EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e-Education

Christian Gütl[1, 2], Klaus Lankmayr[1], Joachim Weinhofer[1] and Margit Höfler[1]
[1]Institute for Information Systems and New Media (IICM), TU-Graz, Austria
[2]Curtin University, Perth, Australia
Christian.Guetl@iicm.tu-graz.ac.at
lanki@sbox.tugraz.at
jowein@sbox.tugraz.at
mhoefler@iicm.tu-graz.ac.at

Abstract: Research in automated creation of test items for assessment purposes became increasingly important during the recent years. Due to automatic question creation it is possible to support personalized and self-directed learning activities by preparing appropriate and individualized test items quite easily with relatively little effort or even fully automatically. In this paper, which is an extended version of the conference paper of Gütl, Lankmayr and Weinhofer (2010), we present our most recent work on the automated creation of different types of test items. More precisely, we describe the design and the development of the *Enhanced Automatic Question Creator (EAQC)* which extracts most important concepts out of textual learning content and creates single choice, multiple-choice, completion exercises and open ended questions on the basis of these concepts. Our approach combines statistical, structural and semantic methods of natural language processing as well as a rule-based AI solution for concept extraction and test item creation. The prototype is designed in a flexible way to support easy changes or improvements of the above mentioned methods. EAQC is designed to deal with multilingual learning material and in its recent version English and German content is supported. Furthermore, we discuss the usage of the EAGC from the users' viewpoint and also present first results of an evaluation study in which students were asked to evaluate the relevance of the extracted concepts and the quality of the created test items. Results of this study showed that the concepts extracted and questions created by the EAQC were indeed relevant with respect to the learning content. Also the level of the questions and the provided answers were appropriate. Regarding the terminology of the questions and the selection of the distractors, which had been criticized most during the evaluation study, we discuss some aspects that could be considered in the future in order to enhance the automatic generation of questions. Nevertheless the results are promising and suggest that the quality of the automatically extracted concepts and created test items is comparable to human generated ones.

Keywords: e-assessment, automated test item creation, distance learning, self-directed learning, natural language processing, computer-based assessment

1. Introduction

Highest flexibility is required from the members of our modern world in terms of continuous adaptation of knowledge and skills. Formal education in primary and secondary settings but even academic settings is not sufficient any more for our ever-changing and knowledge-driven society. Thus life-long learning is the key in such an environment and new pedagogical approaches such as exemplary-based learning and self-directed learning are becoming increasingly popular. (Gütl, 2010) Commonly agreed and widely discussed in literature, such as in Bransford, Brown and Cocking (2000), assessment has not only be seen as an integrated part of the learning processes but also feedback to students and teachers is important to adapt the learning process and improve the learning outcome. Assessment activities are resource intensive and time-consuming which has motivated different computer-supported and computer-assisted approaches. The various approaches range from applications supporting human-based marking and feedback to applications, which support automated assessment. E-assessment tools can certainly reduce effort and improve feedback, however, the creation of appropriate test items is a time consuming task, in particular to assess content alternatives and different knowledge levels in adaptive e-learning environments. Moreover, in self-directed learning settings or more general in life-long learning settings there is no pre-defined learning content and students can select content from open or closed repositories or even Web content. Consequently it is almost impossible to provide prepared test items for such kind of learning. (Gütl, 2008)

This importance of assessment in the learning process has motivated the *Advanced Educational Media Technologies (AEMT) Group* at Graz University of Technology to initiate a research program on e-assessment to cover the entire life cycle of the assessment process by semi-automated and auto-

Reference this paper as: Gütl, C, Lankmayr, K, Weinhofer, J and Höfler, M. "Enhanced Automatic Question Creator – EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e-Education" *The Electronic Journal of e-Learning Volume 9 Issue 1 2011, (pp23-38), available online at* www.ejel.org

mated approaches. One important research strand in this context is semi-automated and fully-automated test item creation. A first simple solution has combined an approach for statistic text summaries and a named entity detection algorithm (Gütl, 2008). Findings of the first approach have led to an enhanced approach combining statistical, structural and semantic analysis for concept detection, and based on that different types of test items have been created (Gütl, Lankmayr, & Weinhofer, 2010). First pilot trials, a user study and findings from the development point of view have resulted in further improvements of the prototype.

In this paper, which is an extended version of the conference paper of Gütl, Lankmayr and Weinhofer (2010), we want to outline the enhanced version of the prototype and report about the most relevant finding of a user study focusing on the perception of the quality of the automatically created test items. To this end, the paper is structured as followed: first we will give background information and related work on both the concept extraction and automatic test item creation. This is followed by requirements, design and development of the enhanced prototype, the *Enhanced Automatic Question Creator (EAQC).* A discussion from the users' viewpoint as well as user study of the quality of the extracted concepts and created test items give first insights of the practical usage.

2. Background and related work

Following the basic idea of the proposed approach of the automated creation of assessment items, one of the most important tasks is the identification of the most relevant concepts form of natural language texts of the learning content, which is an active research topic in past and present, such as in (Moens & Angheluta, 2003; Villalon & Calvo, 2009). A short overview of the historic developments of concept extraction is based on (Gütl et al, 2010; Weinhofer, 2010). Early and initial ideas of concept extraction can be based on research of Luhn who found statistic relationships of words in textual content (Luhn, 1957). In the late 1970s Edmundson improved this method by combining cue phrases, word frequencies, title words and the position of words in a paragraph. Kupiec, Pederson and Chen (1995) extended this method by considering acronyms and proper nouns additionally. Frank et al (1999) created a domain-specific key phrase extraction (KEA) that uses a Naive Bayes classification depending on word frequency and the position of the first occurrence of the word. KEA was extended by Turney (2003) who enhanced the algorithm by co-occurrences which consider the customariness of two words together in the WWW. Song, Han and Rim (2004) generated lexical chains and a concept score depending on word association, the depth in WordNet hierarchy and a semantic relation weight. Hassan, Mihalcea and Banea (2007) use a text rank algorithm that takes account of the context of a word by transforming the document into a graph and calculating node weights. Ledeneva, Gelbukh and García-Hernández (2008) evaluate n-grams, consisting of n words, instead of single words to determine the importance of concepts. A more detailed discussion of methods and approaches can be found elsewhere, such as at (Liu & Yang, 2009; Hovy, Kozareva & Rillof, 2009).

By further focusing on research of *automated test item creation*, an extensive literature review has shown just few pre-existing approaches and tools where most of the available tools support multiple choice items (Gütl, 2008; Lankmayr, 2010; Gütl et al, 2010). In an early and simple approach, Coniam (1997) identified the concept/expression by two distinctive ways: a) user defined n-th word deletion depending on a predefined entry point, and (b) a part of speech tag. Distractors are extracted from a list derived from the Bank of England Corpus whereby these words have similar word frequencies in that corpus as the selected word. In the approach from Mitkov and Ha (2003), distractors based on given key terms are calculated by the use of WordNet. The questions are built by a rule based transforming of sentences into interrogative clauses. Machine learning was applied by Hoshino and Nakagawa (2005). Thus, k-Nearest neighborhood, naïve Bayes classification and a suitable training set are utilized to identify the positions of the blanks in news articles for creating multiple choice items. Goto et al (2010) introduce a solution, which combines the following process steps: (a) extract appropriate sentences based on preference learning, (b) identify blank part based on conditional random field, and (c) create distracters based on statistical patterns of existing questions. Brown, Frishkoff and Eskenazi (2005) developed the REAP system, that is able to provide texts suitable for users' reading levels and to generate appropriate multiple choice but also assignment items. Some work can also be identified focusing on other test item types. By the help of WordNet using definitions, synonyms, antonyms, hypernyms and hyponyms question items are formed to improve word knowledge by evaluating user statistics. Chen, Liou and Chang (2006) have built grammar tests by transforming sentences extracted from the WWW. The transformation is done by applying manually generated patterns and is used for creating multiple choice items and error detection tests. Rus, Cai and Graesser (2007) introduced methods for generating questions with the help of patterns, templates and a special markup language named QG-ML. The patterns are characterized by semantic, lexical and syntactical structures whereas the templates describe methods to implement these structures to generate questions. Heilman and Smith (2010) generate questions from reading materials by applying manually created rules and a ranking algorithm for items selection. Gütl (2008) described a system that uses automatic summary of a document to identify key concepts (named entities) and that generates completion tests as well as limited choice items.

The evaluation of the current state of research suggests that approaches using machine learning are strongly depending on the training set and the knowledge domain. Most of the illustrated systems are applying either statistical or semantic methods and are not able to fulfill the requests given by the variety of assessment item types. Moreover, pre-existing approaches and tools are not sufficiently flexible and extendable to support the above mentioned variety of application scenarios and learning settings. For this reason we developed a system, the *Automatic Question Creator (AQC)* as outline in Gütl et al (2010), which builds on a combination of statistical, semantic and structural analysis to accomplish a step-by-step extraction of relevant concepts from natural language texts. Insights of the first version of the prototype have led us to improve the system which is outlined in the subsequent sections,

3. Requirements, design and development

This section is an extended and updated version of the technical description outlined in Gütl et al (2010) and covers the technical aspects of the improved version of the automatic question creator tool, the *Enhanced Automatic Question Creator (EAQC).*

3.1 Objectives and high level requirements

Based on the findings and experiences of the first prototype development, the goal of the EAQC is to apply improved natural language processing methods which supports the creation of test items or even generates them automatically from the learning content of different languages. A flexible design should enable various groups to use the tool stand-alone or to integrate it in a learning platform as well as adjust the tool according to the specific learning setting. This has led us to specify to following requirements on an abstract level:

- Support of various input file formats from local file systems and from Internet resources
- Multilanguage support
- Domain knowledge and document structure independency
- Identification of most important concepts
- Creation of test items and reference answers based on identified concepts
- Support of open ended, single choice, multiple-choice and completion exercises
- Variability, configurability, modularity, extensibility and performance
- Interoperability with existing eLearning systems

3.2 Conceptual architecture and tools

The high-level conceptual design of the EAQC is outlined in Figure 1. It illustrates the core conceptual units and pre-existing tools as well. The system can be unfold into three main modules: (1) The *Preprocessing module* deals with format conversion of several file formats and online resources, text cleaning methods, language detection and transformation into an internal XML schema which contains all necessary data for further processing. In the current system English and German languages are supported, however, the flexible design easily enables to integrate other modules or tools to support other languages. (2) The *Concept Extraction module* performs structural, statistical and semantic analysis, runs term weighting and finally extracts the most suitable phrases; a detailed description is given in Section 3.3. (3) The *Assessment Creation module* determines the most appropriate sentence for each phrase and adds the previous and the following sentences to provide sufficient context information. Moreover the module identifies distractors and antonyms, creates question items and reference answers, and finally transforms those items in QTI standard.

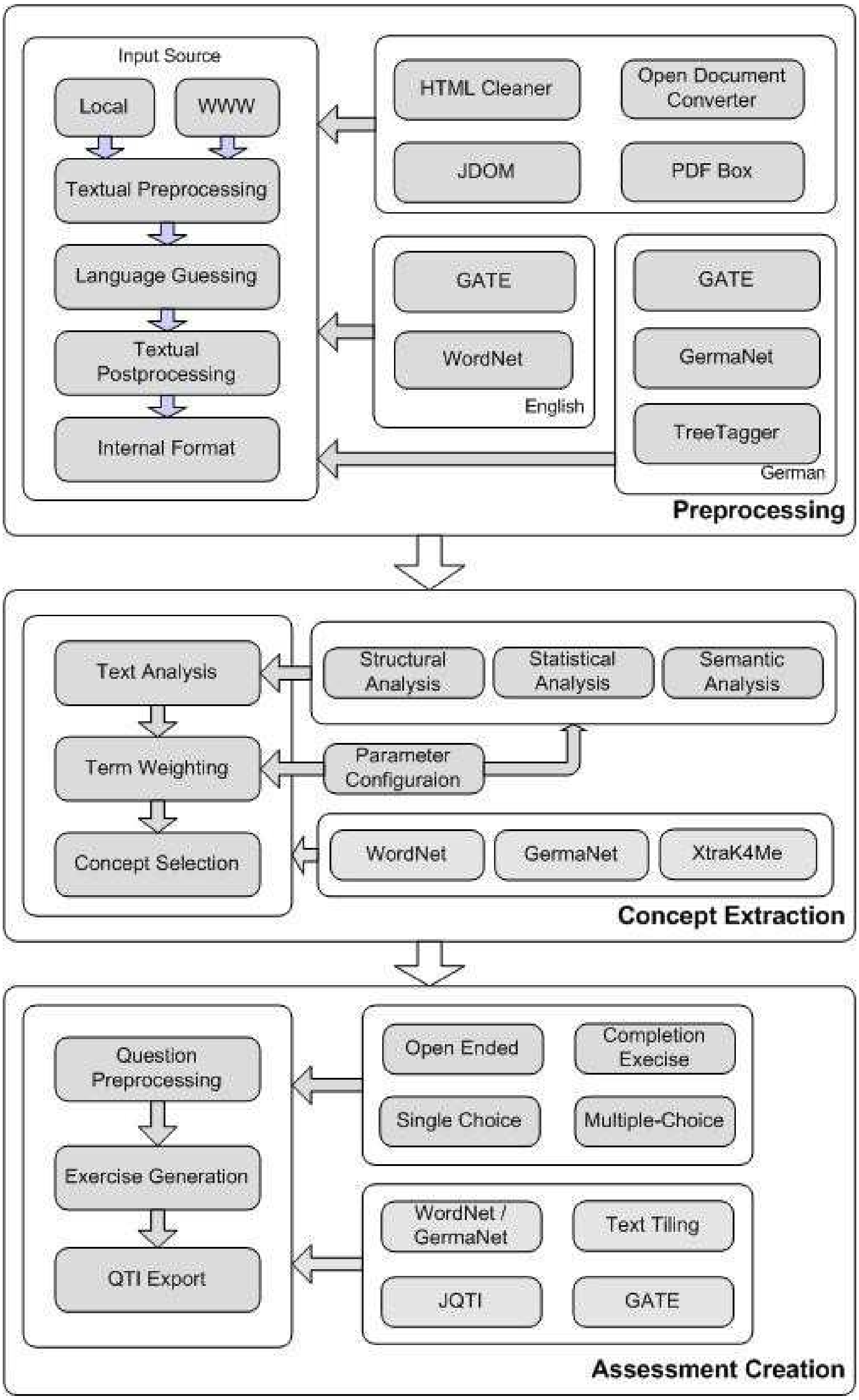


Figure 1: Conceptual design of Enhanced Automatic Question Creator (EAQC)

The main components integrated in the implemented system are GATE and two lexical databases. The GATE framework, especially the ANNIE plug-in, is used for basic text processing and annotation. Thereby the text is split up into tokens and sentences, the part of speech classification as well as name entity recognition, noun chunking and co-reference resolution of each token are performed. (GATE, 2010) The semantic analysis is processed with WordNet in case of English language or GermaNet when performing analysis on a German text. (WordNet, 2010; GermaNet, 2009) Thereby semantic and lexical relations between words are calculated as well as distractors and antonyms are selected. Format conversion for Word, Open Document Text and HTML is utilized to transform the input files into a HTML format by using JODConverter (2010). PDF files are transformed with the help of PDFBox (2010) that is able to extract the textual information from such files, structural information is added manually by applying predefined patterns. Content of the WWW, such as Wikipedia, is also supported as input source by the Automatic Question Creator. To ensure a high quality conversion especially to support Wikipedia content, a Wikipedia parser was implemented, to deal with the inconsistency of the provided HTML source code. Afterwards the generated HTML file is cleaned up using HTML Cleaner (2006) to ensure a conversion to XML with JDOM (2010). The concept extraction done by the EAQC is assisted by XtraK4Me of Schutz (2008) which was adapted to fit the requirements of the German language too. QTI exportation and rendering is done with JQTI (2008).

3.3 Data structure and applied methods

The main idea of our enhanced approach is to combine statistical, semantic and structural analysis to find most relevant words in learning content or more concrete concepts suitable for creating tests and exercises. Based on general word frequencies of the stemmed text the EAQC transforms those frequencies into weights for each word. In the second step of the process chain, these weights are adapted by a configurable set of algorithms that evaluate dependencies of the words according to the appearance in the text, such as in title, abstract, keywords, headlines. Also structure and formatting style as well as word types are considered in the process. Depending on the set of the highest weights and further configurable parameters the EAQC generates single choice, multiple-choice, completion exercises and open ended questions. Moreover the system is capable of exporting the test items including reference answers into the QTI format to allow integration into other learning and assessment systems.

In order to support the process chain, an internal data structure is applied which is organized into three main elements as illustrated in Figure 2. A *Word Element* contains all necessary textual and structural information of each token retrieved from GATE, WordNet and format conversion as well as from statistical and semantic analysis. According to the German language additional information is retrieved from GermaNet, the TreeTagger and the Durm German Lemmatizer. (TreeTagger, 1996; Durm, 2010) Each token is also associated with a *Weight Element* that stores a weight of each algorithm performed for concept extraction. The *Sentence Element* is calculated for each sentence in the text and contains the sentence boundaries, the related concepts and the sentence weight.

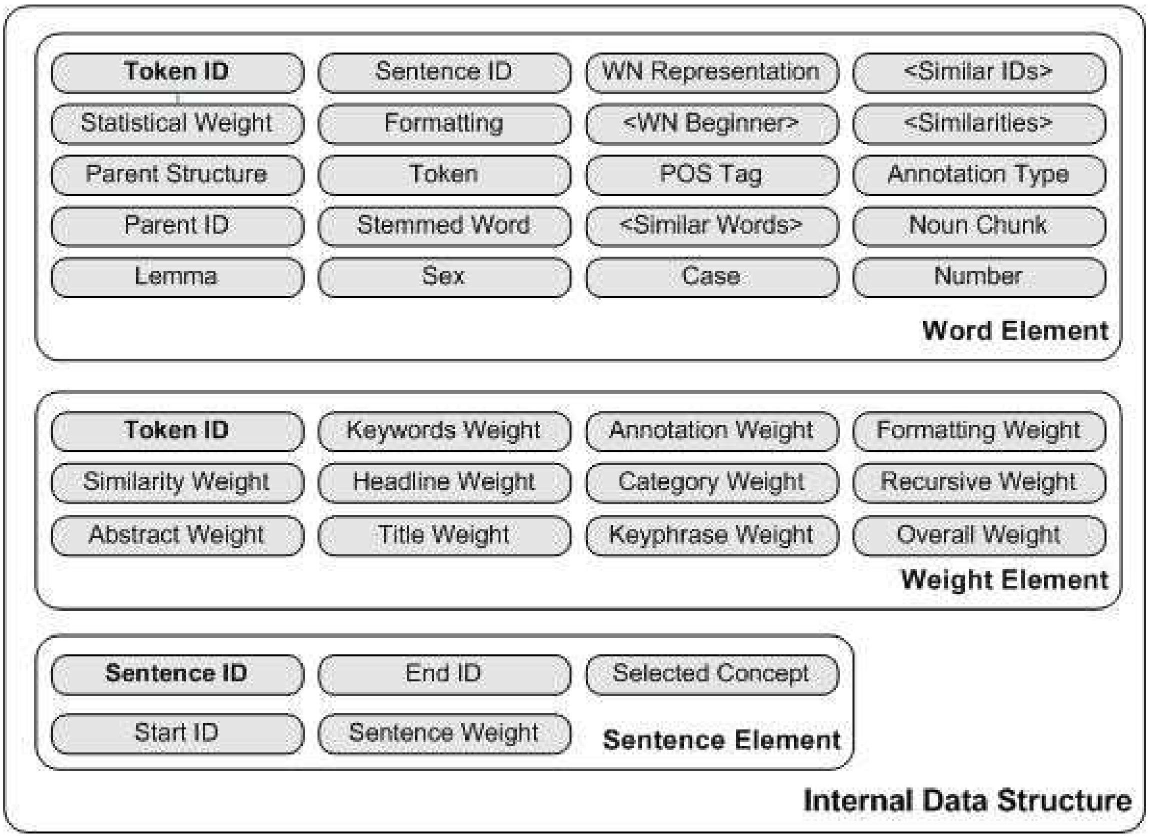


Figure 2: Internal data representation

The overall weight of words is composed of its statistical weight based on word occurrence w_1 (see Table 1, line 1) and several other weights w_l (see Table 1, line 2 - 11) that are retrieved by applying statistic, semantic and structural analysis. Most of these methods are subject to the distance of words in the used lexical databases hierarchies. The influence of each those weights on the overall weight can be adjusted by a set of independent parameters $k_{m,l}$. Our first approach to the calculation of the overall weight $\mathbf{w(i)}$ for a word $\mathbf{i}$ is shown in equation (1), further experiments and improvements are subject to future work. To ensure stop word elimination only nouns and verbs are considered. A more detailed description of the weighting process and the applied methods can be found in Weinhofer (2010).

Table 1: Algorithms, weights and configurable parameters

Module l_n	Weight	# Adjustable Parameters k_m	Description
1	$w_{stat}(i)$	1	statistical weight, normalized number of occurrences of a stemmed word in a section
2	$w_{sim}(i)$	2	weight derived from statistical weights of similar words, depends on similarity measures retrieved from Word-Net and GermaNet
3	$w_{title}(i)$	1	semantic relation to the words in the title
4	$w_{headline}(i)$	6	semantic relation to the words in the corresponding headline depending on the headline layer (up to 6)
5	$w_{abstract}(i)$	1	semantic relation to the words in the abstract
6	$w_{keywords}(i)$	1	semantic relation to keywords
7	$w_{annotation}(i)$	17	weight for the special annotations retrieved by GATE, the 17 annotation types can be handled individually
8	$w_{category}(i)$	25	weight according to the 25 unique beginners retrieved from WordNet and GermaNet
9	$w_{formatting}(i)$	1	Weight depending on the text formatting
10	$w_{keyphrase}(i)$	1	weight for phrases supplied form XtraK4Me algorithm
11	$w_{recursive}(i)$	2	recursive similarity weight calculation, consideration of lexical chains

In a further step, for each noun which is above a predefined threshold, a set of phrases that contain this word is built for each of the sections. Then all phrases of each set are weighted by summing up the overall weights of all words contained in a phrase. The highest weighted phrase of each set is chosen as potential concept. Finally the concept extraction is accomplished by building a collection of the best of these concepts for each section of the text.

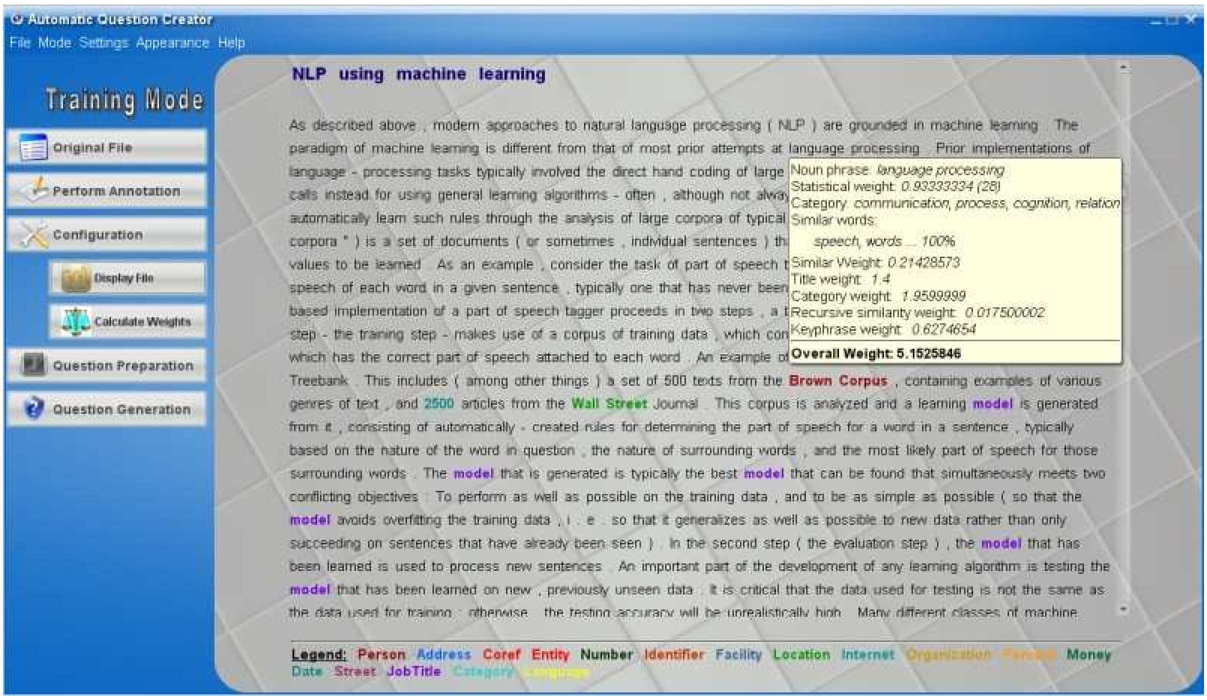
$$w(i) = w_{stat}(i) * \Bigg(k_1 + \sum_{l=2}^{11} [\![(w_l *]\!] \sum_{m} k_m \Bigg)\Bigg) \qquad (1)$$

For *Completion Exercises* the previous and following sentences are added to the selected sentence to offer additional context information to the user. In all of those sentences the selected concepts get replaced with fill-in blank areas to avoid unnecessary hints. *Multiple-choice item* also requires distractor calculation. Basically the distractors are determined by searching coordinate terms for the whole question phrase in WordNet respectively in GermaNet. If this calculation fails, the phrase gets split in all possible coherent n-grams and the coordinate terms for the longest sequence are randomly selected. In the worst case only a single word of a concept delivers suitable results. Due to the circumstance that there are very few proper nouns and no dates included in WordNet and GermaNet, a special case appears if the concept is assigned to a special annotation type. In this case three random phrases sharing the same annotation type are chosen as distractors from the underlying document. *Single choice items* can be generated by searching antonyms for single words in a concept and replacing the original word. Since the result of this procedure is seldom satisfying, the same method is repeated with all adjectives, verbs and nouns of the whole sentence. *Open ended exercises* are generated using several patterns depending on the special annotation type in the selected concept. Due to the fact of implementing a fully automatic assessment system the difficulty according to open ended questions is to compute a reference answer automatically. To meet this challenge the EAQC uses the text tiling algorithm to find the most proper text block containing the extracted concept.

The created test items are finally transformed into the QTI standard as single XML file for each question item. The reason for that exportation is to afford an opportunity of integrating the generated test items in learning management systems or other assessment tools. Currently a web service is developed to improve the flexibility in terms of submitting the learning content and to access the extracted phrases and the created test items.

4. Usage viewpoint

This section outlines EAQC from the user's point of view which is focused on the semi-automatic test item creation in a kind of interactive mode. The fully automated test item creation or batch mode processes the same steps but applies pre-configured settings. As the improvements of the enhanced tool (the EAQC) mainly have focused on methods of concept identification and test item creation, the graphical user interface has kept the same. Thus, the content in this section is a slightly adapted version of Gütl et al (2010) showing the process steps applied on the learning content of the case study (see also Section 5). The process steps are as follows: First, an input file in one of the supported formats has to be selected either from the local file system or from an Internet resource. The text is converted and filtered as well as a control output is generated. In the next step the user can induce the annotation process and the internal data structure is built. The result of the annotation is shown and the user can initiate the weighting process for concept identification. Figure 3 illustrates an example of a weighted text and the calculated weighting factors of a token which results from selected methods. In this step the user can initially set or change the weighting factors of the methods or even select and unselect methods to be applied (see Figure 4).



NLP using machine learning

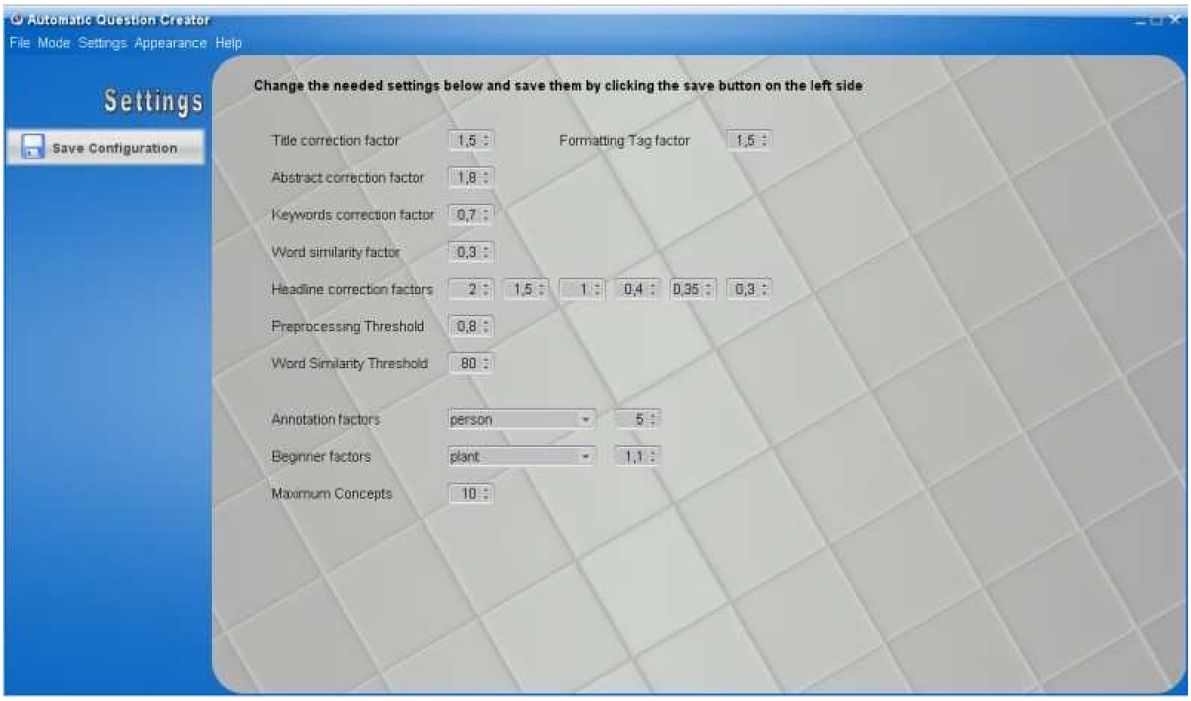
As described above , modern approaches to natural language processing (NLP) are grounded in machine learning . The
paradigm of machine learning is different from that of most prior attempts at language processing . Prior implementations of
language - processing tasks typically involved the direct hand coding of large
calls instead for using general learning algorithms - often , although not alway
automatically learn such rules through the analysis of large corpora of typical
corpora ") is a set of documents (or sometimes , individual sentences) th
values to be learned . As an example , consider the task of part of speech t
speech of each word in a given sentence , typically one that has never been
based implementation of a part of speech tagger proceeds in two steps , a t
step - the training step - makes use of a corpus of training data , which con
which has the correct part of speech attached to each word . An example of
Treebank . This includes (among other things) a set of 500 texts from the Brown Corpus , containing examples of various
genres of text , and 2500 articles from the Wall Street Journal . This corpus is analyzed and a learning model is generated
from it , consisting of automatically - created rules for determining the part of speech for a word in a sentence , typically
based on the nature of the word in question , the nature of surrounding words , and the most likely part of speech for those
surrounding words . The model that is generated is typically the best model that can be found that simultaneously meets two
conflicting objectives : To perform as well as possible on the training data , and to be as simple as possible (so that the
model avoids overfitting the training data , i . e . so that it generalizes as well as possible to new data rather than only
succeeding on sentences that have already been seen) . In the second step (the evaluation step) , the model that has
been learned is used to process new sentences . An important part of the development of any learning algorithm is testing the
model that has been learned on new , previously unseen data . It is critical that the data used for testing is not the same as
the data used for training ; otherwise , the testing accuracy will be unrealistically high . Many different classes of machine

Noun phrase: *language processing*
Statistical weight: *0.93333334 (28)*
Category: *communication, process, cognition, relation*
Similar words:
speech, words ... 100%
Similar Weight: *0.21428573*
Title weight: *1.4*
Category weight: *1.9599999*
Recursive similarity weight: *0.017500002*
Keyphrase weight: *0.6274654*
Overall Weight: 5.1525846

Legend: Person Address Coref Entity Number Identifier Facility Location Internet Organization Percent Money
Date Street JobTitle Category Language

Figure 3: Annotated and weighted text

The next step in the process chain is the selection of the most important concepts to finally create the test items. Figure 5 illustrates a sample of concept extraction whereby the highest weighted phrase for each section of the content is listed on the screen. The user is enabled to deselect unwanted phrases as well as add unconsidered phrases or single words in a chapter of the text. Based on the final settings the test item creation is initiated. Different types of test items can be selected and instances of created items can be viewed. An example of a generated multiple-choice test item is outlined in Figure 6 that shows the representation of the QTI item in HTML.

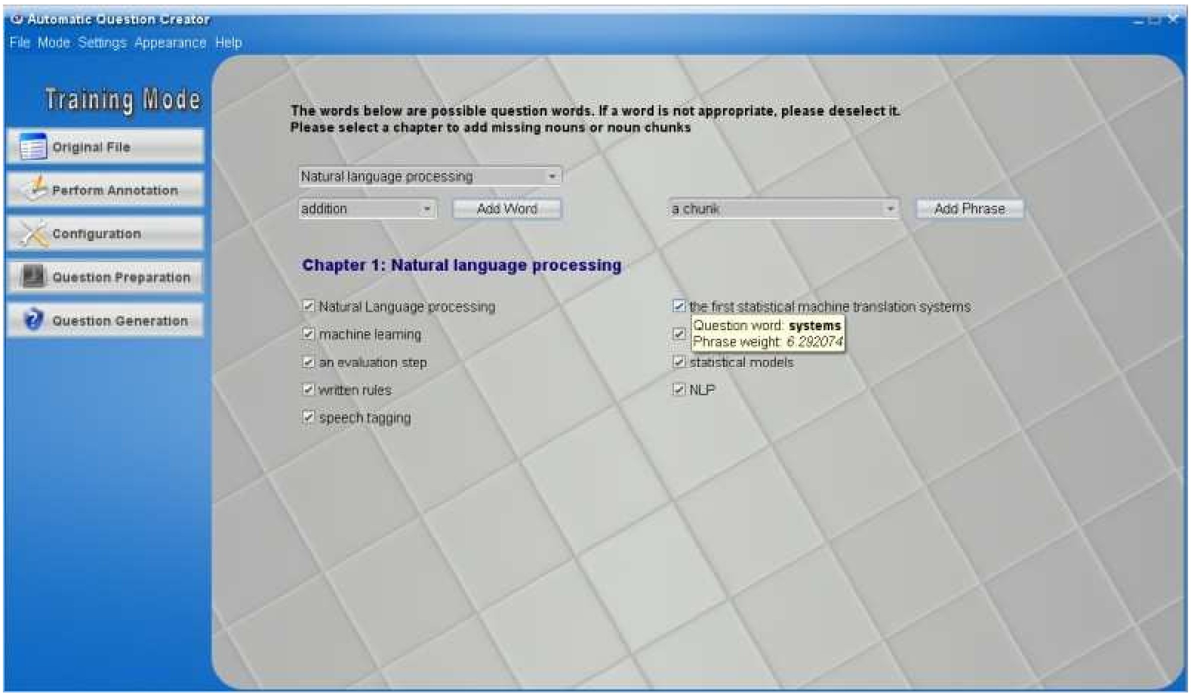

Automatic Question Creator
File Mode Settings Appearance Help
Settings
Save Configuration

Change the needed settings below and save them by clicking the save button on the left side

Title correction factor 1,5
Formatting Tag factor 1,5
Abstract correction factor 1,8
Keywords correction factor 0,7
Word similarity factor 0,3
Headline correction factors 2 1,5 1 0,4 0,35 0,3
Preprocessing Threshold 0,8
Word Similarity Threshold 80

Annotation factors person 5
Beginner factors plant 1,1
Maximum Concepts 10

Figure 4: EAQC configuration panel



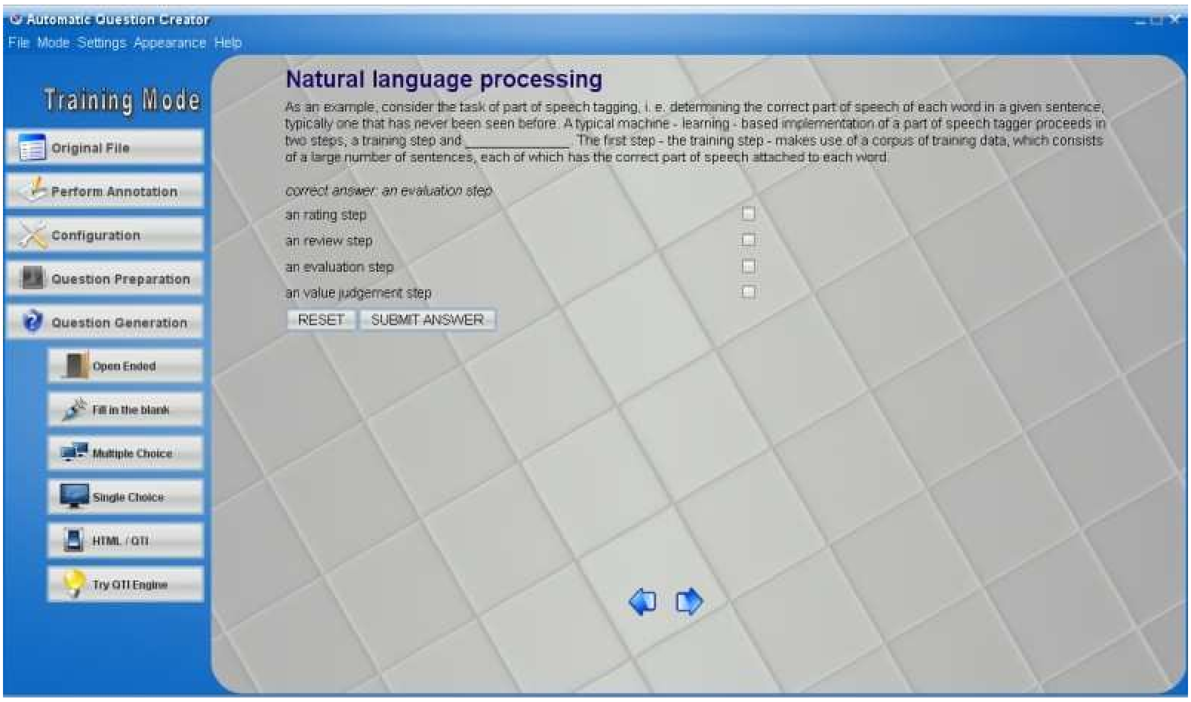
The words below are possible question words. If a word is not appropriate, please deselect it.
Please select a chapter to add missing nouns or noun chunks

Natural language processing
addition
a chunk

Chapter 1: Natural language processing

☑ Natural Language processing
☑ machine learning
☑ an evaluation step
☑ written rules
☑ speech tagging
☑ the first statistical machine translation systems
☑ Question word: **systems** Phrase weight: *6.292074*
☑ statistical models
☑ NLP

Figure 5: Example of extracted concepts



Natural language processing

As an example, consider the task of part of speech tagging, i. e. determining the correct part of speech of each word in a given sentence, typically one that has never been seen before. A typical machine - learning - based implementation of a part of speech tagger proceeds in two steps, a training step and _____________. The first step - the training step - makes use of a corpus of training data, which consists of a large number of sentences, each of which has the correct part of speech attached to each word.

correct answer: an evaluation step

an rating step □

an review step □

an evaluation step □

an value judgement step □

Figure 6: Example of a multiple-choice item

5. Case study in academic education

To verify the implemented system, especially the quality of the extracted concepts and created test items, we conducted a study within the regular course “Information Research and Retrieval (ISR)” at Graz University of Technology at the end of the winter semester 2010/11. In particular, we were interested in how students evaluate automatically extracted concepts and test items (namely open-ended, single choice, multiple choice, and completion exercises, respectively) compared to concepts and test items generated by human.

5.1 Study setup

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. Most of them (93.1%) were bachelor students; the rest were master students. Results from the tests delivered during the study (see below) were part of the final grading of the course, but note that the participation in the study was not a prerequisite for the completion of the course. All participants gave informed consent before attending the study. In order to generate questions with the EAQC, we modified a learning content (approximately 2,600 words) about “Natural Language Processing” (NLP) from Wikipedia (http://en.wikipedia.org/wiki/Natural_language_processing).

The procedure of the study was as follows: At the beginning, the scope and the time schedule of the study was briefly outlined by the experimentators. Participants were informed that they had to attend several learning activities during the session (see also Lankmayr, 2010, for a similar approach). The whole material (including the text, the instructions and all questionnaires) was presented as Web-based content. Furthermore, although almost all of the students were German-speaking, the learning content and the questionnaires were presented in English in order to enable comparing studies on international level. Participants were also asked to provide - if necessary - answers in English. After the introduction, students were asked to learn the text about “Natural Language Processing” (NLP) for 35 minutes and to briefly summarize it afterwards (Test 1; 10 minutes). Participants were not allowed to consult the given text during the test.

After a short break, the first of two main learning activities started. The goal of the first learning activity was that the students became familiar with the learning content. Similar to the operation method of the EAQC, students were asked to extract relevant concepts from the text first and to create eight test items (labeled as questions in the following) concerning the text afterwards. According to the test items generated by the EAQC, students had to generate two open ended questions, two completion exercises, two single choice questions, and two multiple-choice questions, respectively. Example concepts and example questions for each of the four question types concerning a different topic were provided. Participants were allowed to use the text while working on this task. This learning activity lasted about 40 minutes. Subsequently, participants again had to attend a test without any help. Contrariwise to the first test, this test included eight prepared questions and lasted 15 minutes. Four questions in this test based on the EAQC; four had been generated by human.

After a further break in the second learning activity, participants were asked to evaluate concepts and questions that had been generated beforehand by the EAQC or by human. In total, 56 concepts and 24 questions (six per each of the four question types) had to be evaluated. From the 56 concepts, 49 had been extracted by the EAQC (highest ranked by the tool) and seven by human. The 49 automatically extracted concepts corresponded to the suitable phrases calculated by the EAQC in descending order from the text (see Section 3 for details). From the 24 questions to be evaluated during the study, 16 questions had been generated by the EAQC and eight questions had been generated by human. The 16 automatically generated test items (four per each question type) based on the four highest ranked concepts that had been extracted by the EAQC.

Participants were asked to evaluate the relevance of a concept using a 5-point Likert scale (1 = not relevant at all; 5 = very relevant). The quality measure for assessing the questions was derived from the observation matrix of Canella, Ciancimino and Campos (2010). This observation matrix originally consisted of the pertinence, level, terminology, and the interdisciplinarity regarding test items created by students. In our context the interdisciplinarity is not appropriate due to the usage of patterns and the focus on specific topics. Therefore we adapted the procedure to evaluate the quality of the automatically or manually generated questions. Participants were asked to evaluate the questions with respect to the following criteria, again using a 5-point Likert scale (1 = very bad; 5 = very good):

- Pertinence: relevancy of a question in the given context
- Level: level of difficulty of a question
- Terminology: appropriateness of the words chosen
- Answer: quality of the reference answer
- Distractors: quality of the listed distractors (for multiple-choice items only)

The order of the concepts and questions to be evaluated was randomized. This second learning activity lasted approximately 45 minutes. At the end of the evaluation task, students had to fill in a questionnaire in which they were asked to answer more general questions about the task (e.g., how difficult it was to generate and evaluate the questions, respectively, or whether the time schedule for each task was appropriate). In total, the whole experiment lasted approximately three hours. Students were also asked to evaluate further questions for homework (results are not included to the analysis presented here).

5.2 Results

In the following, we concentrate on the students' evaluation of the concepts and the questions in the second learning activity. We first investigated the quality of the concepts extracted by the EAQC by comparing those concepts with manually generated concepts. The mean rating for the concepts extracted by the EAQC was 2.6 (*SD* = 0.4), for manually extracted concepts it was 4.0 (*SD* = 0.6; see Figure 7). A two-tailed *t*-test for dependent measures showed that this difference was reliable, $t(28) = 14.87$; $p < .001$. This means that students evaluated automatically extracted concepts as less relevant compared to concepts extracted by human. However, when we only investigate the relevance of the seven highest ranked concepts provided by the EAQC, mean ratings for the automatically extracted concepts increased to 3.9 (*SD* = 0.3; Figure 8). In this case, ratings for concepts extracted by the EAQC were equal to concepts extracted by human, $t(28) = 1.21$, $p = 0.23$, meaning that the most suitable automatically extracted concepts were as relevant as manually extracted concepts.

Based on the automatic concept extraction (see Section 3) it can be assumed that the perceived relevance of the concepts provided by the EAQC decreases with their ranking; i.e., we expected that lower ranked concepts after the extraction phase should be evaluated worse compared to the higher ranked concepts. We investigated this assumption by comparing the mean ratings for the first half of the automatically extracted concepts (higher ranked concepts) with the second half of the concepts (lower ranked concepts). A two-tailed *t*-test for dependent measure showed that students evaluated higher ranked concepts extracted by the EAQC indeed better compared to lower ranked concepts, *t*(28) = 10.27, *p* < .001. Taken together these results showed that the concepts extracted by the EAQC differ as expected in their relevance: Higher ranked concepts were perceived as more relevant compared to lower ranked concepts. Furthermore, these automatically extracted higher ranked concepts did not differ in their relevance from manually extracted concepts.

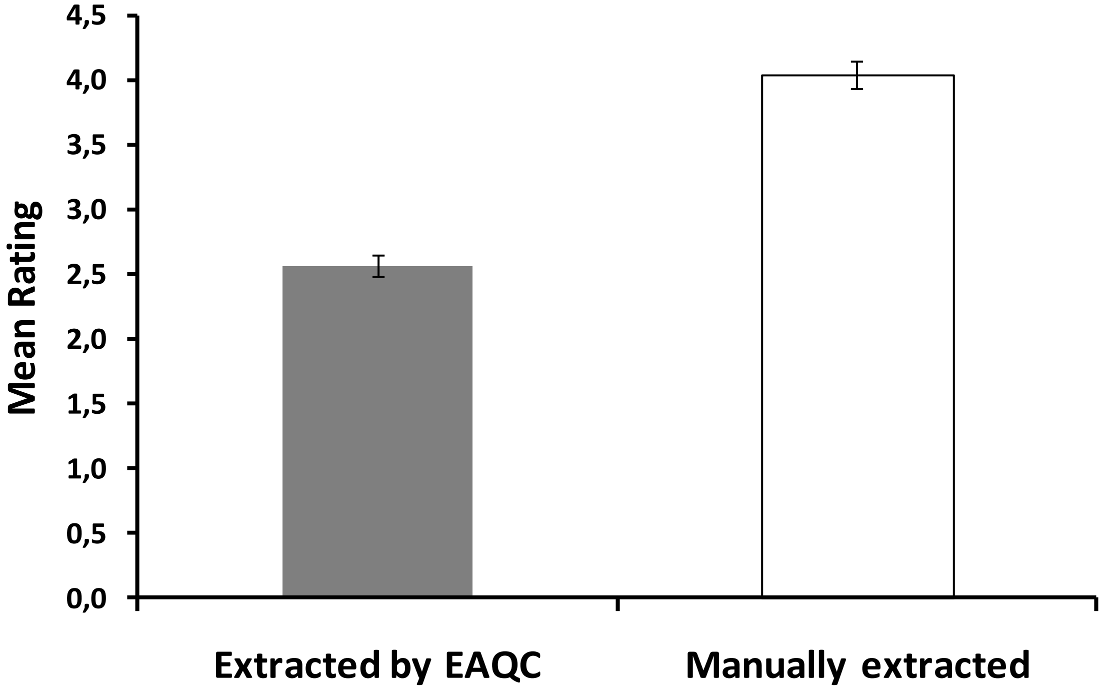


Figure 7: Mean ratings for concepts extracted by the EAQC compared to manually extracted concepts. Error bars represent the standard error

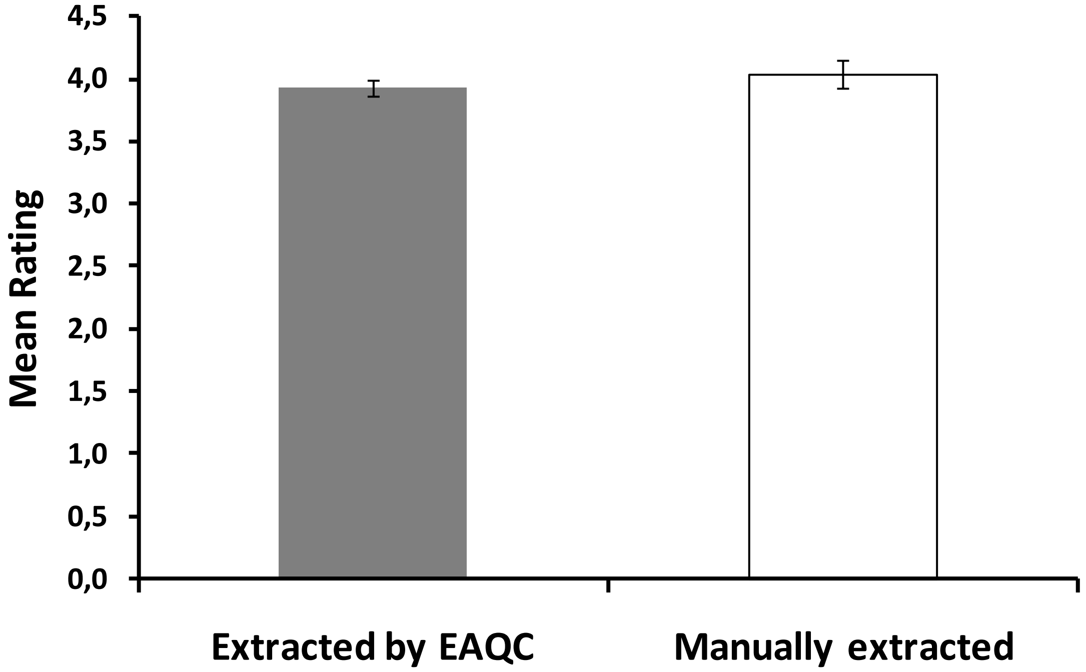


Figure 8: Mean ratings for the seven highest ranked concepts extracted by the EAQC and the seven concepts extracted manually. Error bars represent the standard error

Before discussing the results of the concept analysis more in detail, we present the analysis regarding the quality of the questions provided by the EAQC. For this analysis we only investigated the questions evaluated in the second learning activity because these questions based on the highest ranked concepts by the EAQC. Hence, we assumed that these questions should be evaluated as relevant as the manually created questions. Table 2 shows examples for “good” and “bad” questions with respect to each question type. We defined a question as “good” regarding an evaluation criterion, when the average rating for this criterion was above 3.5. The respective criterion is presented in parentheses in Table 2. Accordingly, a question was “bad” regarding a specific criterion when the mean rating was below 3.0. For instance, the “good” open-ended question presented in the example received higher ratings regarding its pertinence and terminology; the “bad” multiple-choice question received lower ratings regarding its terminology and its distractors.

Table 2: Examples of “good” and “bad” questions by the EAQC for each question type, the respective evaluation criteria which were evaluated best and worst are presented in parentheses. To simplify matters, we did not include answers for open ended questions

	“Good” question	“Bad question”
Open ended	What do you know about Modern NLP algorithms in the context of Natural language processing? *(Pertinence & Terminology)*	What do you know about Natural Language processing in the context of Natural language processing? *(Terminology)*
Single choice	Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. [true] *(Pertinence & Terminology)*	However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in those languages trade [correct: text] edition segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language. *(Terminology)*
Completion exercise	[...] Little further research in machine translation was conducted until the late 1980 s, when _______________ were developed. [...] Answer: the first statistical machine translation systems *(Answer)*	______________ (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. [...] Answer: Natural Language processing *(Level)*
Multiple choice	[...] Little further research in machine translation was conducted until the late 1980 s, when _______________ were developed. [...] A1: the first statistical machine translation systems A2: the first statistical robotics systems A3: the first statistical mt systems *(Pertinence)*	[...] However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in _______________ is a significant task requiring knowledge of the vocabulary and morphology of words in the language. A1: those hyponyms text segmentation A2: those indications text segmentation A3: those languages text segmentation A4: those expressive styles text segmentation
		(Terminology & Distractors)

Figure 9 shows the comparison between manual and automatically created test items (averaged across question types) regarding the five evaluation criteria (i.e., pertinence, terminology, level, answer, and distractors, respectively) described before. Ratings were generally high with an average of *M* = 3.4 (*SD* = 0.4) for automatically generated questions and *M* = 3.7 (*SD* = 0.3) for manually generated questions. We compared questions created by EAQC and manually created questions for each quality criteria by computing individual two-tailed *t*-tests for depended measures. Results showed that mean ratings for questions created by EAQC did not differ from the manually created questions regarding pertinence, level, and answer (all *p*’s > .05, Bonferroni corrected). However, regarding terminology and quality of the distractors, questions created by the EAQC were rated worse compared to manually created questions (all *p*’s < .001).

Although comparison between the two conditions (i.e., automatically vs. manually created questions) should be interpreted with caution, because there were less questions created by humans than by the EAQC, results nevertheless suggest, that the quality of the questions created by the EAQC is quite good. As expected from the analysis of the underlying concepts, results indicate that the questions provided by the AGQ were as relevant as questions provided by humans. This is further evidence that the key concepts extracted by the EAQC and hence, the questions that base on these concepts are indeed equally relevant for the students. However, further experimentation is necessary in order to evaluate the quality of questions that base on less suitable (i.e., lower ranked) concepts.

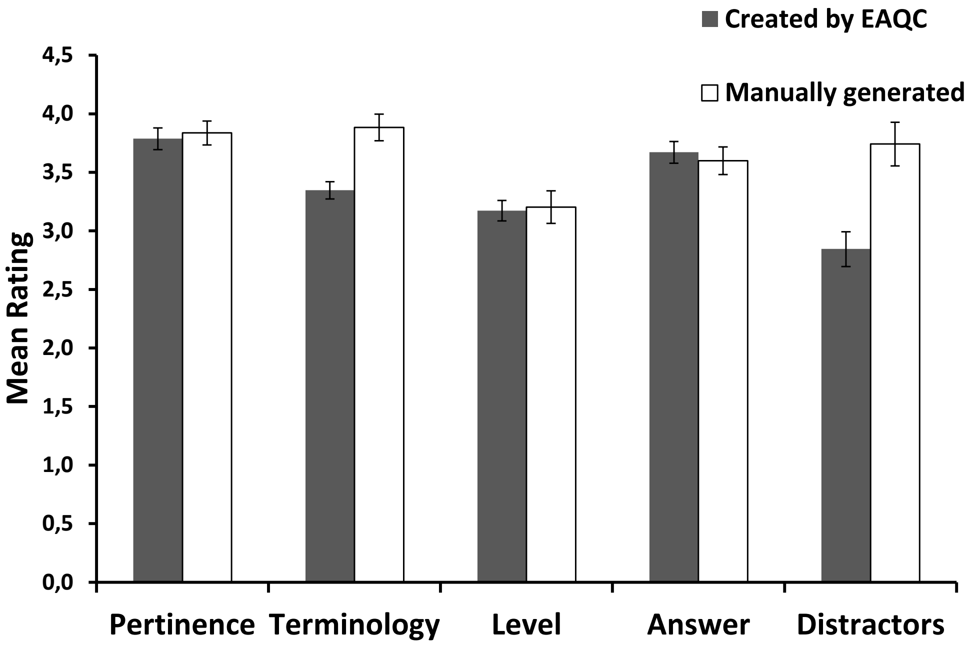


Figure 9: Comparison of manually and automatically created questions with respect to the defined evaluation criteria. Error bars represent the standard errors

Furthermore, results also showed that the level of the questions and the provided answers seem to fulfill the needs of the students. Regarding these criteria of the items' difficulty and the answers, there was no difference between automatically and manually created questions. However, students' perception of the terminology and the quality of the distractors created by the EAQC was worse compared to their perception of the same aspects regarding manually created questions. A closer look to the data suggests that the terminology was worse especially for completion exercises and multiple choice questions. This is insofar somewhat surprising as the terminology of those question types - when automatically created - did not differ that much from the terminology of the original sentences in the text. For instance, a completion exercise is created by using an existing sentence or paragraph of the text, leaving blank the main concept (= answer) (see also Table 2). Perhaps students are not that familiar with such a style. For instance, when students were asked to create themselves completion exercises and multiple-choice questions during the first learning activity, they typically constructed new sentences and did not simply use the existing ones. Hence, it is possible that not the terminology of the questions *per se* but their terminology in context of questioning is inappropriate. In any case, further experimentation is necessary to investigate this issue in more detail. For instance, students could be asked to define why the terminology of a question is inappropriate or how it could be improved.

Results also showed that the quality of the distractors provided by the EAQC was worse compared to human created questions. Automatic generation of distractors is still very challenging. Previous research suggests that the chosen distractors should be as semantically close to the correct answer as possible (Mitkov, Ha, & Karamanis, 2005). Our current approach builds on antonyms and related terms on concept or word level. Improvements could be gained by more carefully choosing distractors which we are currently working on. Another alternative for improvements could be the deep study of the process of distractor creation by subject domains in order to implement a similar process chain in the tool. Hence, also in this case further experimentation is necessary in order to create appropriate distractors for multiple choice questions. Furthermore, it might be worth investigating why a specific distractor is suitable or not in order to define enhanced criteria for the improvement of our tool.

Finally, the analysis of the automatically extracted concepts showed that not all 49 concepts extracted by the AGC were equal in their relevance. On the one hand, this is in accordance with the concept extraction strategy described in Section 3: The automatically extracted concepts are ranked regarding their suitability and are therefore a priori not expected to be equally relevant at all. On the other hand, however, this nevertheless raises - from a pedagogical viewpoint - three important questions. First, when exactly is a concept relevant or not? Second, what is the appropriate number of concepts that should be extracted in general so that only “relevant” concepts are used for question generation? Third, is it perhaps also worth providing questions that base on “less relevant” concepts? Regarding the first two objections, analysis of one task of the first learning activity of the study showed that students themselves extracted 17.1 concepts on average (*SD* = 10.3); ranging from 5 to 41 extracted concepts per student. Note at this point that the students were asked to extract the “main” concepts of the text; i.e., to extract such concepts they perceive as relevant. The variance in the number of self-extracted concepts indicates that there are big individual differences between students. Such individual differences should also be taken into account by the EAQC when questions are automatically created. As described before, the user has the possibility to add or deselect phrases during the phase of the automatic concept extraction. In doing so, the EAGC already supports the creation of questions on the basis of the individual students’ requests. A further improvement of the tool could be that a user simply enters relevant concepts (based on his or her individual viewpoint which concepts are relevant) into the system to receive questions from the EAQC. Such an approach would also support the benefit of the EAQC with respect to self-regulated learning activities. However, students sometimes might face the problem that they cannot estimate, which concepts are relevant and which are not. In this case they would miss important concepts for question creation, which, in turn, might impair their learning progress. Therefore, also “less relevant” concepts and the resulting questions might be valuable for a deeper understanding of the learning content. Once again, investigating these issues will be one challenge in future studies.

6. Conclusions and future work

Assessment has to be seen as an integrated and important activity in the learning process. In particular modern educational approaches - such as self-directed or exemplary learning - and personalized learning activities cause a tremendous effort or make it even impossible to prepare appropriate and individualized test items, assess them and provide feedback. To overcome this problem, we advocate an approach which automatically creates test items from learning content, administer knowledge assessment and provide feedback.

We have introduced a concept and prototype implementation, that is capable of handling various text formats and WWW resources, that annotates the corpus using GATE, that applies statistical, semantic and structural methods for identifying key concepts. Based on these concepts the Enhanced Automatic Question Creator (EAQC) generates open ended, single choice, multiple-choice and completion exercises and exports those into QTI items. The evaluation confirmed first promising results and showed the applicability of the system. Encountered problems include (a) the high time complexity for text annotation and WordNet-based operations, (b) problems with specific structures of the content and versions of file formats, (c) partly inappropriate concepts selection due to lack of common sense knowledge and domain knowledge, and (d) the quite low quality of selected distractors.

On the technical level, future work include improvements of better dealing with different content structures, applying common sense and domain knowledge as well as to improve the process of the distractor selection. On the cognitive science and pedagogic level, further pilot studies and evaluations in concrete learning scenarios will be performed.

Acknowledgements

This research is partially supported by the European Commission under the Collaborative Project ALICE "Adaptive Learning via Intuitive/Interactive, Collaborative and Emotional System", VII Framework Program, Theme ICT-2009.4.2 (Technology-Enhanced Learning), Grant Agreement n. 257639. Some of the research work of Klaus Lankmayr and Joachim Weinhofer has been financially supported by an internal research grant of Graz University of Technology. Some contributions of this paper also result of visiting academic activities of Christian Gütl in Australia. The visits were supported and sponsored by School of Information Systems and Curtin Business School, Curtin University, and Institute for Information Systems and Computer Media (IICM) at Graz University of Technology. We are grateful to Isabella Pichlmair and Claudia Pieslinger for their support in the analysis of the study.

References

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.

Brown, J.C., Frishkoff, G.A. and Eskenazi, M. (2005) "Automatic Question Generation for Vocabulary Assessment", *Proceedings of the Human Language Technology Conference on Empirical Methods in Natural Language Processing: 6 – 8 October 2005, Vancouver, British Columbia, Canada*, pp 819- 826.

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

Chen, C.Y., Liou, H.C. and Chang, J.S. (2006) "FAST: An Automatic Generation System for Grammar Tests", *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp 1-4.

Coniam, D. (1997) "A Preliminary Inquiry Into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests", Calico Journal, Vol 14, No. 2–4, pp 15-34.

Durm (2010) "*The Durm German Lemmatizer*", [Online], Concordia University Montreal, http://www.semanticsoftware.info/durm-german-lemmatizer

Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C. and Nevill-Manning, C.G. (1999) „Domain-specific Keyphrase Extraction", *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence IJCAI*, pp 668-673.

GATE (2010). "*GATE: general architecture for text engineering*", [Online], The University of Sheffield, http://gate.ac.uk

GermaNet (2009) "*GermaNet – an Introduction*", [Online], Eberhard Karls Universität Tübingen, http://www.sfs.uni-tuebingen.de/GermaNet/

Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). "Automatic Generation System of Multiple-choice Cloze Questions and its Evaluation", Knowledge Management & E-Learning: An International Journal (KM&EL), Vol 2, No 3, 2010

Gütl, C. (2008) "Automatic Limited-Choice and Completion Test Creation, Assessment and Feedback in modern Learning Processes", Paper read at LRN Conference 2008, Guatemala, February 12th – 16th.

Gütl, C. (2010) "The Support of Virtual 3D Worlds for enhancing Collaboration in Learning Settings", in Francesca Pozzi and Donatella Persico (Eds.) *Techniques for Fostering Collaboration in Online Learning Communities: Theoretical and Practical Perspectives*, IGI Global, 2011, 278-299.

Gütl, C., Lankmayr, K., & Weinhofer, J. (2010) "Enhanced Approach of Automatic Creation of Test Items to Foster Modern Learning Setting", in Proc. of the 9th European Conference on e-Learning, Porto, Portugal, 4-5 November 2010, 225-234.

Hassan, S., Mihalcea, R. and Banea, C. (2007) "Random-Walk Term Weighting for Improved Text Classification", *Semantic Computing, ICSC 2007*, pp 242–249.

Heilman, M., & Smith, N. A. (2010) "Good Question! Statistical Ranking for Question Generation", Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 609–617, Los Angeles, California, June 2010.

Hoshino, A. and Nakagawa, H. (2005) "A real-time multiple-choice question generation for language testing: a preliminary study", *Proceedings of the second workshop on Building Educational Applications Using NLP*, pp 17-20.

Hovy, E., Kozareva, Z. and Riloff, E. (2009) "Towards Completeness in Concept Extraction and Classification", *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Vol. 2*, pp 948-857.

HTML Cleaner (2006) "*HTML Cleaner Project*", Sourceforge.net, http://htmlcleaner.sourceforge.net

JDOM (2010) "*JDOM*", http://www.jdom.org

JODConverter (2010) "*JODConverter*", Art of Solving, www.artofsolving.com/opensource/jodconverter

JQTI (2008) "JQTI", University of Southampton, http://jqti.qtitools.org

Kupiec, J., Pederson, J. and Chen, F. (1995) "A trainable document summarizer", *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp 68-73.

Lankmayr, K. (2010) "Design and Development of Concepts for Automatic Exercise Generation", [online], Graz University of Technology, http://www.iicm.tu-graz.ac.at/thesis/MA%20Lankmayr.pdf

Ledeneva, Y., Gelbukh, A. and Garciá-Hernández, R. A. (2008) "Terms Derived from Frequent Sequences for Extractive Text Summarization", *Computational Linguistics and Intelligent Text Processing, 9th International Conference, CICLing 2008, Haifa, Israel, February 2008, Proceedings,* pp 593-604, Springer Verlag, Heidelberg.

Liu, N. and Yang, C.C. (2009) "Keyphrase Extraction for Labeling a Website Topic Hierarchy", *Proceedings of the 11th International Conference on Electronic Commerce, Taipei, Taiwan, 2009*, pp 81-88.

Luhn, H. P. (1957) "A statistical approach to mechanized encoding and searching of literary information", IBM J. Res. Dev. 1, 4 (October 1957), 309-317. DOI=10.1147/rd.14.0309 http://dx.doi.org/10.1147/rd.14.0309

Mitkov, R. and Ha, A. L. (2003) "Computer-Aided Generation of Multiple-Choice Tests", *Proceedings of the HLT-NAACL 2003 workshop on Building educational applications using natural language processing*, pp 17-22.

Mitkov, R., Ha, A. L., and Karamanis, N. (2005). "A computer-aided environment for generating multiple-choice test items", *Natural Language Engineering*, Vol. 12, pp 177-194.

Moens, F.M. and Angheluta, R. (2003) “Concept extraction from legal cases: the use of a statistic of a coincidence”, *Proceedings of the 9th international conference of Artificial intelligence and law, Scotland, United Kingdom, 2003*, pp 142–146.

PDFBox (2010) “*Apache PDF Box, Java- PDF Library*“, The Apache Software Foundation, http://pdfbox.apache.org/

Rus, V., Cai, Z. and Graesser, A. C. (2007) “Experiments on Generating Questions About Facts“, *Computational Linguistics and Intelligent Text Processing, 8th International Conference, CICLing 2007, Mexiko City, Mexiko, February 18-24, 2007, Proceedings*, pp 444–455.

Schutz, A. (2008) “SMiLE: XtraK4Me - Extraction of Keyphrases for Metadata Creation“, SmILE: Semantic Information Systems and Language Engineering Group, http://smile.deri.ie/projects/keyphrase-extraction

Song, Y.-I., Han, K.-S. and Rim, H.-C. (2004) “A Term Weighting Method based on Lexical Chain for Automatic Summarization“, *Computational Linguistics and Intelligent Text Processing, 5th International Conference, CICLing 2004, Seoul, Korea, February 2004, Proceedings*, pp 636–639, Springer Verlag, Heidelberg.

TreeTagger (1996) “*TreeTagger – a language independent part-of-speech tagger*”, [Online], University of Stuttgart, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

Turney, P.D. (2003) “Coherent Keyphrase Extraction via Web Mining*”, Proceedings of IJCAI’03*, pp 434–439.

Villalon, J. and Calvo, R.A. (2009) “Concept Extraction from student essays, towards Concept Map Mining”, *Proceedings of the 9th IEEE International Conference on Advanced Learning Technologies, ICALT 2009*, pp 221-225.

Weinhofer, J. (2010) “Extraction of relevant semantic data from natural language texts in the view of automatic question generation”, [online], Graz University of Technology, http://www.iicm.tu-graz.ac.at/thesis/MA%20Weinhofer.pdf

WordNet (2010). ”*WordNet: A lexical database for English”*, Princeton University, http://wordnet.princeton.edu