A Perspective on Computer Assisted Assessment Techniques for Short Free-Text Answers

Shourya Roy[1(✉)], Y. Narahari[2], and Om D. Deshmukh[1]

[1] Xerox Research Centre India, Bangalore, India
{shourya.roy,om.deshmukh}@xerox.com
[2] Indian Institute of Science, Bangalore, India
hari@csa.iisc.ernet.in

Abstract. Computer Assisted Assessment (CAA) has been existing for several years now. While some forms of CAA do not require sophisticated text understanding (e.g., multiple choice questions), there are also student answers that consist of free text and require analysis of text in the answer. Research towards the latter till date has concentrated on two main sub-tasks: (i) grading of essays, which is done mainly by checking the style, correctness of grammar, and coherence of the essay and (ii) assessment of short free-text answers. In this paper, we present a structured view of relevant research in automated assessment techniques for short free-text answers. We review papers spanning the last 15 years of research with emphasis on recent papers. Our main objectives are two folds. First we present the survey in a structured way by segregating information on dataset, problem formulation, techniques, and evaluation measures. Second we present a discussion on some of the potential future directions in this domain which we hope would be helpful for researchers.

Keywords: Automatic scoring · Short answer grading · Assessment

1 Introduction

Assessing students' acquired knowledge is one of the key aspects of teachers' job. It is typically achieved by evaluating and scoring students' responses in classroom assessments such as quizzes, examinations, and worksheets. Assessments are important for teachers as these provide them insights on how effective their teaching has been. However, assessment is a monotonous, repetitive and time consuming job and often seen as an overhead and non-rewarding[1]. In addition, it seldom helps teachers to improve their knowledge of subject matters.

Computer Assisted Assessment(CAA) has been prevalent in schools and colleges for many years now albeit for questions with constrained answers such as multiple choice questions (MCQs). There have been several studies on MCQs

[1] https://www.experience.com/alumnus/article?channel_id=education\&source_page=editor_picks\&article_id=article_1133291019105 Assessment (or grading) takes of the order 20 % time for teachers.

E. Ras and D. Joosten-ten Brinke (Eds.): CAA 2015, CCIS 571, pp. 96–109, 2015.
DOI: 10.1007/978-3-319-27704-2_10

which brought out important aspects such as high degree of correlations with constructed response items [37]. While assessment of answers to MCQs are easier for computers, they have been reported to suffer from multiple shortcomings compared to questions requiring free-text answers. Firstly, they are less reliable owing to pure guessing paying some dividends. Techniques which do not account for influence of guessing strategies used by students do not lead to reliable assessment [42]. Secondly, presence of alternative responses provide inadvertent hints which may change nature of problem-solving and reasoning. Finally, in many cases MCQs are not appropriate to measure acquired knowledge such as hypothetical reasoning and self-explanation in Science courses [51]. Consequently, use of open-ended questions that seek students' constructed responses is more commonly found in educational institutions. They reveal students' ability to integrate, synthesize, design, and communicate their ideas in natural language. We call them *free-text* answers. In this paper we consider *short* free-text answers which are at least a sentence long but less than 100 words in length (broadening from the definition of up to 20-word answers from previous work [19]).

Assessment of free-text answers is more laborious and subjective for humans as well as much harder to automate. Research towards the same started multiple decades ago (with publications first appearing in 1960s [34]) and till date has concentrated on two main tasks: (i)grading of essays, which is done mainly by checking style, correctness of grammar, fluency etc. of essays and (ii) assessment of short free-text answers. While the former has seen a vast amount of research work, Jordan had observed that short-answer free-text e-assessment has remained an underused technology [20]. Multiple surveys have been written about automatic grading of essays [7,10,49]. In this paper, we present a structured survey of techniques developed for assessment of short free-text answers which to date is the first attempt to the best of our knowledge.

CAA[2] techniques for short free-text ingest student answers and assign scores usually by comparing to one or more correct answers. Developing a general solution to this is a hard problem owing to multiple reasons viz. linguistic variations in student answers (multiple ways of expressing the same answer), subjectivity of questions (multiple correct answers) and topical variations (Science vs Literature). At a broad level, two types of automatic approaches for scoring have been followed by researchers. Knowledge based approaches involve experts creating all possible model answers for a question and representing them in computer understandable manner. Computer systems then use these model answers to automatically score student responses. On the other hand, machine learning based approaches develop statistical *models* based on a collection of expert graded answers. Loosely speaking, these techniques attempt to learn characteristics (or features) of answers which make them correct. Knowledge-based approaches are useful if variations possible in student answers are limited and can be enumerated. However, considering reasons described above such as linguistic diversity and subjectivity of questions, it could be laborious and ineffective in many cases.

[2] We use the terms "computer assisted assessment(CAA)" and "automated assessment" interchangeably in this paper.

Table 1. Summarized view of data and tasks mentioned in relevant prior work

Ref.	Topic	Level	Nature of answers	Scoring scale
[35,45–47]	GCSE Biology Examinations	14–16 year old pupils	up to 5 lines (about 200 answers for 9 questions)	0–2
[23,48]	Reading comprehension and mathematics	7^{th} and 8^{th} Graders	Short answers up to 100 words	0–2
[28]	1999 Science National Test Paper A and B	11 year old pupils	Single word, single value; short explanatory sentence (120 answers; 4 questions)	0–2
[13,53]	Introductory course in computer literacy	100 College students	short answers with multiple (3–9) correct concepts associated (192 answers; 36 questions)	
[32,33]	Science	3^{rd} to 6^{th} Graders	moderately short verb phrases to several sentences (15,400 answers; 287 questions)	8-point scale
[27]	Assessment of summaries based on reading comprehension	75 undergraduate students	75–100 word long	
[9]	High school Physics	Undergraduate students	at most 1–2 sentences (8000 responses)	4-point scale
[41]	300 Middle school virtual environment scenarios	Middle school Science students	short answers of usually 50–60 words	0–4
[51]	Creative problem solving in Earth Sciences	226 High school students	short-text in Chinese	
[24]	Summary writing for reading comprehension	6^{th} to 9^{th} graders	summaries of about 4 sentences	0–4
[3]	United States Citizenship Exam	Crowd workers on AMT	Up to a couple of sentences (698 respondents; 20 questions)	Boolean Correct and incorrect
[22]	Critical thinking tasks GRE Analytical Writing Prompts	Students from 14 colleges and universities	short answers (5–10 open ended questions)	5-point scale
[11,29,30]	80 questions from Introductory Data Structure Course	Undergraduate students	Short answers of about 1–2 lines	0–5
[40]	87 questions on Object Oriented Programming	Undergraduate students	heterogeneous; about 1–2 lines maximum	
[1]	Essays on a variety of topics	10^{th} grade students	50 words (17,000 responses)	

Oragnization of the paper: We start by presenting a structured view of prior research in automatic assessment of short free-text answers in an organized manner. Starting with types of data and domains researchers looked at, we follow up with technical problem formulations and solutions developed before leading to evaluation metrics used. In Sect. 5, we provide insights obtained from prior work leading to new research directions in this topic. We feel that such a structured view of research would be more useful to researchers than merely describing all

prior work in some order. Finally, we do feel that this work is timely considering the recent trend of large scale democratization of education through Massive Online Open Courses(MOOCs). While content from leading colleges are being floated around to students all over the world, assessments in MOOCs have been quite primitive with usage of multiple choice and 1–2 word/number/formulae questions. Research in automatic assessment has to take leaps over the next few years, supported by advances in machine learning and natural language processing techniques, to enable MOOCs to have assessment of equivalent quality to traditional pedagogical ecosystem.

2 Nature of Data and Tasks

In this section, we provide a summarized view of the wide variety of tasks, subject matter, student population level and size as well as scoring scale used in prior work towards automated assessment of short free-text answers in Table 1(blank cells indicate information not found in respective papers). It is evident that a wide variety of data was used in prior research in computer aided assessment of short free-text answers with little standardization in scoring scheme. While such heterogeneity implies possible wide applicability of techniques, generalizability of developed techniques is not proven. One of the reasons being these datasets were seldom shared and hence tried out by subsequent research to move the state of the art forward. We will come back to these issues while discussing future research directions in Sect. 5.

3 Techniques

In this section, we review techniques which have been used for automatic assessment in prior art. Again, we observe that a wide variety of techniques have been used which we group under key themes and present in an organized manner.

3.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a well-established field of research focusing on developing techniques for computers to understand and generate natural language text. Natural Language Understanding (NLU), a sub-field of NLP, is the process of disassembling, parsing and canonicalizing natural language text.

NLU techniques have been applied to extract syntactic and semantic structures from short free-text answers. These techniques typically require certain amount of data cleaning owing to the noisy nature of the free-text answers. Spelling and punctuation correction, lemmatization, etc. are commonly applied to clean surface form of the text. Stopword removal and stemming are two other commonly used NLU pre-processing steps towards eliminating non-indicative features and reducing variation of words. Researchers have also developed custom parsing methods to handle language errors in student responses to provide

accurate assessments [15]. Parse trees obtained from parsers not only show shallow structure of free-text answers but also can be used to extract higher level features indicating clause structure, negation etc. [48]. Siddiqi et al. developed a system, IndusMarker, based on syntactic structure of student responses using freely available linguistic tools such JOrtho[3] and Stanford Parser[4] to compare extracted structures with examiner specified grammatical structures to arrive at a score [39,40]. They also developed a XML like Question Answer Markup Language (QAML) to capture structure extracted from text. Lexicons and dictionaries play important role in NLU techniques. Given high degree of domain specificity of different assessment tasks, it is quite common to develop domain specific lexicons to include relevant keywords and variations thereof [41]. Towards developing an assessment system for Biology domain, Sukkarieh et al. observed that many relevant terms were missing in the training data which they had to add manually to the lexicon [47]. In webLAS [2], regular expressions are created out of the model answers and given answers are evaluated against these regular expressions to get a grade. The WebLAS system, the system presented in [35] and a few other short answer assessments systems are compared and contrasted in [55]. Concepts from theoretical linguistics are also beginning to be used: for example, [17] uses under-specified semantic formalism *Lexical Resource Semantics (LRS)* to evaluate the meaning of the answers to content-based reading comprehension tasks.

ETS 'e-rater' is a rating engine to evaluate responses to short-answer questions [4]. They argue that domain specific NLP techniques need to be used for these evaluation tasks and motivate the use of *metonyms*: words or multiword terms that can be substituted for each other in a given domain. Authors in [22] compared how the hand-assigned scores compare with machine-assigned scores under a variety of circumstances (difficulty of task, nature of task, gender-bias, ethnicity-bias etc.) where the machine-assignment was done using the 'e-rater' system.

3.2 Information Extraction and Pattern Matching

Information extraction (IE) techniques pull out pertinent information from syntactically analysed pieces of text answers by applying a set of *patterns*. Patterns are defined either on surface text (words, phrases) or structural elements such as parts of speech (PoS) tags. In the case of short free-text answers, they are typically created by subject matter experts to indicate important concepts which should be present in answers.

OpenMark system from Open University in United Kingdom compared student responses with model answers using regular expressions based on algorithmic manipulation of keywords [5]. However, most prior work used patterns of higher complexity defined in terms of PoS tags and other structural elements obtained from NLU tools such as parser. Automated Text Marker system was developed on the principle of breaking down student answers as well

[3] http://jortho.sourceforge.net/.
[4] http://nlp.stanford.edu/software/lex-parser.shtml.

as model answers into smallest viable units of *concepts* with linguistic dependencies between concepts. [6]. To make the system adaptable they employed additional thesauri (for synonym, metonym) and other simplification rules such as removing articles and other "unnecessary" words. Similarly c-rater® matched the syntactical features of student responses (subject, object and verb) to those of model answers [23]. It used handcrafted rules to take care of different types of variations (syntactic and inflexional variation, synonyms) that existed in student responses. Dzikovska et al. used a syntactic parser and a set of hand-authored rules to extract semantic representations from student responses which were then matched against semantic representations of expected correct answers supplied by tutors [8]. Sukkarieh et al. used a Hidden Markov Model(HMM) based PoS tagger, and a Noun Phrase (NP) and Verb Group (VG) chunker for developing the Oxford-UCLES system. It bootstraped patterns by starting with a set of keywords and synonyms and searching through windows of text for new patterns [47]. Another popular system AutoMark employed templates to specify expert-written snippets of text which were looked for matches in student answers [28]. The templates were designed in a way that they could handle variations in the input text by listing possible words and phrases, lemmatisation of verbs and sentence structure. A very similar technique was applied in [18] with a differential ability to flag (for human validation) a student response which failed to match a model answer but is recognized being a close one.

The primary challenge with information extraction based techniques is to arrive at patterns to cover all possible variations in student answers. In addition, this needs to be done manually for every assessment exercise by subject matter experts which makes the entire exercise an expensive one. On the other hand, as these techniques work on the principle of identifying missing concepts, they have the advantage of crafting feedback for students easily based on knowledge of (un)matched patterns.

3.3 Machine Learning

Classification and regression are the two most popular supervised learning paradigms in machine learning literature. Both techniques attempt to learn unknown functions from which a set of labelled data has been generated and use the estimated functions to predict labels of future unlabeled data. For data in n-dimensional real valued feature space, classification techniques learn functions of type $\mathbb{R}^n \rightarrow \mathcal{A}$ where $\mathcal{A}$ is a set of discreet *class labels*. Regression techniques on the other hand learns real valued functions of type $\mathbb{R}^n \rightarrow \mathbb{R}$. In our context, the data points are answers, scores are labels (or continuous values in regression), scored answers are labelled data and new answers are unlabelled data for prediction. Sukkarieh et al. used statistical text classification techniques which do not require complete and accurate parsing (which is difficult owing to ungrammatical and incomplete sentences). They applied classification techniques such as k-Nearest Neighbor, Inductive Logic Programming, Decision Tree and Naïve Bayes to perform two sets of experiments viz. on raw text answers and annotated answers [35,45]. Annotation involved domain experts highlighting parts of

answers that deserved a score. Machine learning wisdom says that performance of classification techniques depend heavily on the choice and synthesis of features which is evident in prior work for assessment tasks as well. Sukkarieh et al. developed Maximum Entropy classifier using features based on lexical constructs such as presence/absence of concepts, order in which concepts appear, role of a word in a sentence(e.g. active/passive) etc. to predict if a student response is entailed in at least one of the model answers [44]. Nielsen et al. used carefully crafted features using NLP preprocessing obtained from lexical and syntactic forms of text [31]. Dzikovska et al. used lexicial similarity scores (number of overlapping words, F1 score, Lesk score and cosine score) to train a Decision Tree classifier to categorize student responses into one of the 5 categories [9]. For summary assessment Madnani et al. used logistic regression classifier on a 5 point scale [24]. They used interesting features to commonalities between an original passage and a summary such as BLEU score (commonly used for evaluating Machine Translation systems), ROUGE (a recall based metric that measures the lexical and phrasal overlap between two pieces of text), overlap of words and phrases etc. Regression techniques were used for automated assessment to arrive at a real valued score which were later rounded off as per scoring scale. Here again we see use of interesting features with state of the art regression techniques. Sil et al. used Support Vector Machines with Radial Basis Function kernels (RBF-SVM) for learning non-linear regression models of grading with several higher order features derived from free-text answers [41]. Wang et al. applied regression technique for assessing creative answer assessment [51].

3.4 Document Similarity

Large number of techniques have been developed for measuring similarity between a pair of text. Variations exist with respect to representations used for text similarity computation. *Lexical similarity* techniques use surface form text but often give suboptimal results owing to not considering semantics (`automobile` and `car` are considered as distinct as `automobile` and `banana`) and context (`Apple computer` and `apple pie` are considered similar as they share a term). *Corpus based similarity* (or semantic similarity) techniques such as Latent Semantic Analysis (LSA) have shown to perform well by addressing these problems. LSA (and related techniques) project documents to a suitably chosen lower dimensional subspace, where cosine similarity has shown to be a reasonable estimate of semantic similarity. *Knowledge based measures* use background knowledge such as Wordnet[5] or domain specific ontologies to estimate how similar two documents are.

Mohler et al. compared performance of corpus based measures with a number of unsupervised knowledge based measures [30]. Their experiments on a 630 answer dataset did not bring out significant differences in performances of different measures. However, the authors opined that corpus based measures are

[5] http://wordnet.princeton.edu/.

more generalizable as their performance can be improved by improving corpora relevance and increasing corpora size. In a follow up paper, they proposed hybrid techniques using graph alignment on dependency graph (between students' answers and instructor's answer-key) and lexical semantic similarity measures [29]. On the same dataset, Gomaa and Fahmy compared several lexical and corpus based similarity algorithms (13 string based and 4 corpus) and their combinations for grading answers in 0–5 scale [11]. Combination of different string matching and overlap techniques were studied by Gutl on a small scale dataset [16]. Mintz et al. compared different measures such as Word Count, Information Content [36] and Coh-Metrix [26] to score summaries based on features such as *narrativity*, *syntactic simplicity* etc. [27].

LSA has been extensively used for assessment tasks as researchers observed that capturing semantic similarity is most important (student answer should *mean* the same and not necessarily *read* the same as model answers). One of the early tutoring systems AutoTutor [52] used LSA to compare students' answers to model answers by calculating distance between their corresponding vector projections [13]. If cosine similarity of a student response was greater than a threshold then the answer was considered correct. In addition to correct answers, they also had a list of anticipated bad answers – high similarity with those indicated incorrect student response. In a related work, they studied effect of size and specificity of corpora used for creating LSA space on accuracy of automatic assessment [53]. They reported that performance of automatic assessment improved with corpus size though the increase was not linear. They also reported that the performance improved with specificity and relevance of corpus to the task at hand which is a well accepted wisdom in the field now.

LSA based techniques did not always give good performance due to not considering linguistic characteristics such as negation, attachment, predication, modification etc. Researchers also tried adding higher level NLP features such as POS tags but they did not claim to produce significant improvement over vanilla LSA based techniques [21,54].

3.5 Clustering

Basu et al. used clustering techniques to group responses into a two level hierarchy of clusters based on content similarity [3]. They used human supervision as labeled examples to learn similarity metrics using features such as difference in length, fraction of words with matching base forms etc. They observed that TFIDF[6] was the best similarity metric for performing clustering. Obtained clusters could help teachers efficiently grade a group of responses together. They also provided early results on automatic labeling (correct/wrong) based on content similarity. Not for subjective questions, but a very similar idea for evaluating handwritten answer scripts were proposed by [25].

[6] http://en.wikipedia.org/wiki/Tf%E2%80%93idf.

4 Evaluation

An important aspect of CAA for short free-text answer assessment task is to use appropriate evaluation metrics for judging goodness of developed automated techniques. Typically, performance of automatic assessment techniques is measured in terms of agreement with human assigned scores (often average of multiple human scores). Various measures of correlation such as Pearson's Correlation Coefficient, Cohen's Kappa etc.[7] have been used to quantifiably measure extent of agreement.

Sil et al. used a χ^2 test with a threshold of $p < 0.05$ to determine statistical significance of Pearson's Correlation Coefficient [41]. Graesser et al. introduced the notion of *compatibility percentage* for grading answers which matched ideal answers only partially before applying correlation analysis [13]. Similarly, Kanejiya et al. asked human experts to evaluate answers on the basis of *compatibility score*(between 0 and 1) before applying correlation analysis [21]. Sukkarieh et al. used kappa statistics with respect to percentage agreement between two human annotators for evaluation [44,48]. Mohler et al. reported Root Mean Square Error (RMSE) for the full dataset as well as median RMSE across each individual questions [29]. In a prior work, Mohler also highlighted lack of proper analysis before using Pearson's Correlation Coefficient(e.g. normal distribution, interval measurement level, linear correlational model etc.) as well as abundance of possible measures(e.g. Kendall's tau, Goodman-Kruskal's Gamma).

Performance of supervised classification based techniques is represented as a two dimensional table known as *confusion matrix*[8]. Rows in confusion matrix represent human expert assigned grades and columns are computer assigned grades. A cell c_{ij} represents number of answers which are scored i by human and j by the automated technique. Principal diagonal elements represent number of answers where both have agreed. On the basis of confusion matrix, multiple measures such as *accuracy*, *precision* and *recall*, F_1, *specificity* and *sensitivity* etc. have been used to determine how well predicted scores matched with ground-truth scores [24,28]. Classifiers have parameters using which one can trade off precision for recall or vice versa. One known problem with these measures is that they can grossly misreport in case of uneven class distribution e.g. number of correct responses being much more than number of wrong ones. Dzikovska et al. reported both macro-averaged and micro-averaged measures with the latter taking class size into account (there by favoring techniques doing well on larger classes) [9].

5 Discussion

In this section we identify a few possible future research directions in automatic assessment of short free-text answers:

[7] http://en.wikipedia.org/wiki/Inter-rater_reliability.
[8] http://en.wikipedia.org/wiki/Confusion_matrix.

- We observe that there is a lot of variation in short free-text answers(Refer Table 1). Techniques developed for assessing answers to Science questions for middle school students are not expected to work well for assessing summaries written by undergraduate students. Variations with respect to factors such as subject matter, level of students, length and type of text need to be accounted for in the techniques. *A matchmaking framework providing guidance to choose the most appropriate technique for an assessment use-case would be valuable to practitioners.* On a related note there is a dire need of creating and sharing datasets across researchers as mentioned in Sect. 2. Benchmark datasets in machine learning and natural language processing have enabled researchers to come up with new techniques as well as report quantifiable progress over the years. Similar activity would enable assessment techniques to build on vast amount of existing prior work as reviewed in this paper.
- Almost all prior work have assumed existence of model answers for questions for automated assessment of student answers. *An interesting problem would be to develop techniques which can perform assessment without model answers* leveraging a large relevant knowledge base such as wikipedia[9] and babelnet[10]. Work in automatic question answering from the Web would be a starting point though most of those have focused on factual questions [50].
- Assessment is a long term exercise over months and years. Students undergo a number of quizzes and examinations through out their academic career. Most research described in this paper has considered each assessment independently – ignoring prior knowledge of student performance. If a student performed well in all prior examinations then it is probable that she will perform well in the current assessment as well. *Techniques considering a student model along with free-text answers can overcome limitations of techniques which work only based on answer content.* This is analogous to *prior* in Bayesian framework which is combined with observed data for inferencing.

6 Conclusion

Assessment is important for teachers to understand students' acquired knowledge but it takes up a significant amount of their time and is often seen as an overhead. CAA addressed this problem to some extent by enabling automatic assessment for certain types of questions and letting teachers spend more time in teaching. However the benefit and adoption of CAA in schools and colleges has been marginal owing to their relatively limited applicability. Towards expanding the reach of teachers' pedagogy, computers have been in use for content dissemination over the internet in the form of distance learning and e-learning. Massive Online Open Courses (MOOCs), over the last few years, have expanded the reach of high quality pedagogical materials by orders of magnitude. However, till date certifications and degrees from MOOCs are much less acknowledged and respected than the ones from traditional pedagogical ecosystem. We believe the

[9] http://www.wikipedia.org/.
[10] http://babelnet.org/.

difference in assessment methodologies is one of the key reasons for the same. While classroom-based education system primarily use subjective questions to holistically assess students acquired knowledge, MOOCs have been wanting with their MCQ based and peer assessment practices. Need of the hour is large scale assessment systems capable of handling all types of answers; at least short free-text answers. This is an overwhelming task and consolidated research effort will be needed to bridge the gap over the next few years.

References

1. The hewlett foundation: Short answer scoring. http://www.kaggle.com/c/asap-sas, Accessed on 6 March 2015
2. Bachman, L., Carr, N., Kamei, G., Kim, M., Pan, M., Salvador, C., Sawaki, Y.: A reliable approach to automatic assessment of short answer free responses. In: Proceedings of the 19th International Conference on Computational Linguistics, pp. 1–4 (2002)
3. Basu, S., Jacobs, C., Vanderwende, L.: Powergrading: a clustering approach to amplify human effort for short answer grading. Trans. Assoc. Comput. Linguist. **1**, 391–402 (2013)
4. Burstein, J., Wolff, S., Lu, C.: Using lexical semantic techniques to classify free responses. In: Viegas, E. (ed.) Breadth and Depth of Semantic Lexicons, vol. 10, pp. 227–244. Springer, The Netherlands (1999)
5. Butcher, P.G., Jordan, S.E.: A comparison of human and computer marking of short free-text student responses. Comput. Educ. **55**(2), 489–499 (2010)
6. Callear, D., Jerrams-Smith, J., Soh, V., Dr. Jerrams-smith, J., Ae. H.P.: CAA of short non-MCQ answers. In: Proceedings of the 5th International CAA Conference (2001)
7. Dikli, S.: An overview of automated scoring of essays. J. Technol. Learn. Assess. (JTLA) **5**(1), 36 (2006)
8. Dzikovska, M., Bell, P., Isard, A., Moore, J.D.: Evaluating language understanding accuracy with respect to objective outcomes in a dialogue system. In: EACL, pp. 471–481. The Association for Computer Linguistics (2012)
9. Dzikovska, M., Nielsen, R.D., Brew, C.: Towards effective tutorial feedback for explanation questions: a dataset and baselines. In: HLT-NAACL, pp. 200–210. The Association for Computational Linguistics (2012)
10. Gomaa, W.H., Fahmy, A.A.: Tapping into the power of automatic scoring. In: The Eleventh International Conference on Language Engineering, Egyptian Society of Language Engineering (ESOLEC) (2011)
11. Fahmy, A.A., Gomaa, W.H.: Short answer grading using string similarity and corpus-based similarity. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **3**(11), 210–214 (2012)
12. Graesser, A.C., Person, N.K.: Question asking during tutoring. Am. Edu. Res. J. **31**, 104–137 (1994)
13. Graesser, A.C., Wiemer-Hastings, P.M., Wiemer-Hastings, K., Harter, D., Person, N.K.: Using latent semantic analysis to evaluate the contributions of students in autotutor. Interact. Learn. Environ. **8**(2), 129–147 (2000)
14. Grundspenkis, J.: Development of concept map based adaptive knowledge assessment system. In: IADIS International Conference e-Learning, pp. 395–402 (2008)

15. Guest, E., Brown, S.: A new method for parsing student text to support computer-assisted assessment of free text answers. In: 11th CAA International Computer Assisted Assessment Conference, pp. 223–236. Loughborough University, Loughborough, UK, July 2007
16. Gütl, C.: Moving towards a fully automatic knowledge assessment tool. Int. J. Emerg. Technol. Learn. (iJET), 3(1) (2008)
17. Hahn, M., Meurers, D.: Evaluating the meaning of answers to reading comprehension questions a semantics-based approach. In: The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, ACL, pp. 326–336 (2012)
18. Jordan, S., Mitchell, T.: e-assessment for learning? the potential of short-answer free-text questions with tailored feedback. Br. J. Educ. Technol. **40**(2), 371–385 (2009)
19. Jordan, S.: Student engagement with assessment and feedback: some lessons from short-answer free-text e-assessment questions. Comput. Educ. **58**(2), 818–834 (2012)
20. Jordan, S.: Short-answer e-assessment questions: five years on (2012)
21. Kanejiya, D., Kumar, A., Prasad, S.: Automatic evaluation of students' answers using syntactically enhanced LSA. In: Proceedings of the HLT-NAACL Workshop on Building Educational Applications Using Natural Language Processing, vol. 2, pp. 53–60. Association for Computational Linguistics, Stroudsburg, PA, USA (2003)
22. Klein, SP., et al.: Characteristics of hand and machine-assigned scores to college students answers to open-ended tasks. In: Probability and Statistics: Essays in Honor of David A. Freedman, pp. 76–89. Institute of Mathematical Statistics (2008)
23. Leacock, C., Chodorow, M.: C-rater: automated scoring of short-answer questions. Comput. Humanit. **37**(4), 389–405 (2003)
24. Madnani, N., Burstein, J., Sabatini, J., OReilly, T.: Automated scoring of a summary writing task designed to measure reading comprehension. In: Proceedings of the 8th Workshop on Innovative use of NLP for Building Educational Applications, pp. 163–168. Citeseer (2013)
25. Manohar, P., Roy, S.: Crowd, the teaching assistant: educational assessment crowdsourcing. In: HCOMP, AAAI (2013)
26. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press, Cambridge (2014)
27. Mintz, L., DMello, S., Stefanescu, D., Graesser, A.C., Feng, S.: Automatic assessment of student reading comprehension from short summaries. In: Educational Data Mining Conference (2014)
28. Mitchell, T., Russell, T., Broomhead, P., Aldridge, N.: Towards robust computerized marking of free-text responses. In: Proceedings of 6th International Computer Aided Assessment Conference (2002)
29. Mohler, M., Bunescu, R.C., Mihalcea, R.: Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: The Association for Computer Linguistics, ACL, pp. 752–762 (2011)
30. Mohler, M., Mihalcea, R.: Text-to-text semantic similarity for automatic short answer grading. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 567–575. Association for Computational Linguistics (2009)
31. Nielsen, R.D., Buckingham, J., Knoll, G., Marsh, B., Palen, L.: A taxonomy of questions for question generation. In: Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge (2008)

32. Nielsen, R.D., Ward, W., Martin, J.H.: Recognizing entailment in intelligent tutoring systems. Nat. Lang. Eng. **15**(4), 479–501 (2009)
33. Nielsen, R.D., Ward, W., Martin, J.H., Palmer, M.: Annotating students' understanding of science concepts. In: LREC, European Language Resources Association (2008)
34. Page, E.B.: The imminence of grading essays by computer. Phi Delta Kappan, vol. 48, pp. 238–243 (1966)
35. Pulman, S.G., Sukkarieh, J.Z.: Automatic short answer marking. In: Proceedings of the Second Workshop on Building Educational Applications Using NLP, EdAppsNLP 05, pp. 9–16. Association for Computational Linguistics, Stroudsburg, PA, USA (2005)
36. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 1, pp. 448–453 (1995)
37. Rodriguez, M.C.: Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. J. Educ. Meas. **40**(2), 163–184 (2003)
38. Shermis, M.D., Burstein, J., Higgins, D., Zechner, K.: Automated essay scoring: writing assessment and instruction. In: International Encyclopedia of Education, pp. 20–26 (2010)
39. Siddiqi, R., Harrison, C.: A systematic approach to the automated marking of short-answer questions. In: IEEE International Multitopic Conference, 2008, INMIC 2008, pp. 329–332 (2008)
40. Siddiqi, R., Harrison, C.J., Siddiqi, R.: Improving teaching and learning through automated short-answer marking. IEEE Trans. Learn. Technol. **3**(3), 237–249 (2010)
41. Sil, A., Ketelhut, D.J., Shelton, A., Yates, A.: Automatic grading of scientific inquiry. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 22–32. Association for Computational Linguistics (2012)
42. Singley, M.K., Taft, H.L.: Open-ended approaches to science assessment using computers. J. Sci. Educ. Technol. **4**(1), 7–20 (1995)
43. Sukkarieh, J.Z., Bolge, E.: Building a textual entailment suite for the evaluation of automatic content scoring technologies. In: LREC, European Language Resources Association (2010)
44. Sukkarieh, J.Z., Mohammad-Djafari, A., Bercher, J.-F., Bessie're, P.: Using a maxent classifier for the automatic content scoring of free-text responses. In: AIP Conference Proceedings-American Institute of Physics, vol. 1305, p. 41 (2011)
45. Sukkarieh, J.Z., Pulman, S.G.: Information extraction and machine learning: automarking short free text responses to science questions. In: Proceedings of the 2005 Conference on Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology, pp. 629–637. IOS Press (2005)
46. Sukkarieh, J.Z., Pulman, S.G., Raikes, N.: Auto-marking: using computational linguistics to score short, free text responses. In: International Association of Educational Assessment, Philadephia (2004)
47. Sukkarieh, J.Z., Pulman, S.G., Raikes, N.: Auto-marking 2: an update on the ucles-oxford university research into using computational linguistics to score short, free text responses. In: International Association of Educational Assessment, Philadephia (2004)
48. Sukkarieh, J.Z., Blackmore, J.: c-rater: Automatic content scoring for short constructed responses. In: FLAIRS Conference, AAAI Press (2009)

49. Valenti, S., Neri, F., Cucchiarelli, A.: An overview of current research on automated essay grading. J. Inf. Technol. Educ. Res. **2**, 319–330 (2003)
50. Voorhees, E.M., Tice, D.M.: Building a question answering test collection. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 200–207 (2000)
51. Wang, H.-C., Chang, C.-Y., Li, T.-Y.: Assessing creative problem-solving with automated text grading. Comput. Educ. **51**(4), 1450–1466 (2008)
52. Wiemer-Hastings, P., Graesser, A.C., Harter, D.: The foundations and architecture of autotutor. In: Goettl, B.P., Halff, H.M., Redfield, C.L., Shute, V.J. (eds.) ITS 1998. LNCS, vol. 1452, pp. 334–343. Springer, Heidelberg (1998)
53. Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A.: Improving an intelligent tutor's comprehension of students with latent semantic analysis. In: Artificial Intelligence in Education, vol. 99 (1999)
54. Wiemer-Hastings, P., Zipitria, I.: Rules for syntax, vectors for semantics. In: Proceedings of the 23rd Annual Conference of the Cognitive Science Society, NJ (2001)
55. Ziai, R., Ott, N., Meurers, D.: Short answer assessment: establishing links between research strands. In: The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, ACL, pp. 190–200 (2012)