Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

R&D Project Proposal

Evaluation of Active Learning for Short Answer Grading

Mohandass Muthuraja, Jeeveeswaran Kishaan

Supervised by

Prof. Paul G. Ploeger

Deebul Nair

Month 20XX

1. Introduction

Assessment of the knowledge acquired by the students is one of the most important aspects of the learning process. Different forms of assessments that exist today include multiple choice questions, fill-in-the-blanks, essay questions and short answer questions. Prior works have shown that multiple choice questions and fill-in-the-blanks fail to capture the vital aspects of the acquired knowledge such as reasoning and self-explanation [25]. In contrast, questions which require the students to construct responses in natural language have been found to be more effective in assessing their grasp on the subject matter [21]. Essay questions and short answer questions belong to this category.

Limited availability of teachers, online learning platforms, and individual or group study sessions done outside classrooms necessitated quick and efficient assessment of free text responses. Computer assisted assessment / automatic grading evolved as a solution to this problem and a lot of research has been done on automating the grading of essay[8] and short answer responses[12, 19, 17]. The focus of this work would be on improving the existing solutions to grade short answer questions. Short answer questions are characterized by the following aspects[4]:

- the question must require a response that recalls external knowledge instead of requiring the answer to be recognized from within the question
- the question must require a response given in natural language
- the answer length should be roughly between one phrase and one paragraph
- the assessment of the responses should focus on the content instead of writing style
- the level of openness in open-ended versus close-ended responses should be restricted with an objective question design

Automatic short answer grading essentially deals with using computational methods to predict the grades for students' answers, thus cutting down the effort

and time of teachers / professors. Many automated approaches have been proposed in the past for grading short answer questions. These methods compare how similar the students' answers are to the one provided by the teacher or professor and assign a grade proportional to the magnitude of similarity [16]. Different approaches of computing these similarity measures include handcrafted pattern matching, automatic pattern matching, lexical similarity (how particular words effect other words), semantic similarity (deals with the meaning of sentences), and entailments (whether one sentence leads to another without any contradiction). Fig 1.1 shows a general workflow of automatic short answer grading as a pipeline. After creating a dataset of all the students' answers and model answers written by the teachers, useful features are extracted from the answers using natural language processing techniques. A model is developed based on these features and it is evaluated on new datasets.

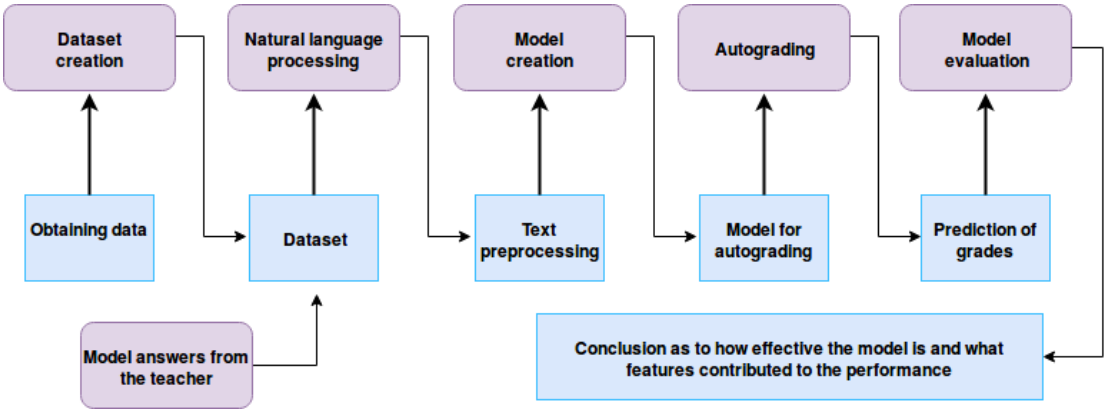


Figure 1.1: Workflow of automatic short answer grading [4].

All the above-mentioned methods belong to a learning paradigm called supervised learning where the right answer for every question is available [10]. Though these approaches were able to produce decent results, they suffer from many shortcomings such as;

- the failure to capture the different wordings/phrasing of the students while trying to answer the short answer questions. It is obvious that anticipating all different ways of answering his questions is practically impossible for the professor.

- lack of sufficient amount of labeled training data in the domain to learn the models. Reasons include privacy, availability, and the quality of the correct answers.
- inability to capture consistent patterns of misunderstandings among students. Ability to recognize such patterns would enable the automated systems to provide useful feedback to students as to why there was a reduction in marks awarded.
- accounting for small deviations in the answers which might affect the whole meaning of the sentence (for ex. in mathematical terms, though each and every word of the student's answer align with that of the professor, a small negation or inverse operation would change the whole meaning).
- finding a way to understand the underlying concept of various students' answers and bagging the similar ones (or the right and wrong ones separately) is also a very tedious task.
- being a passive learner, these models learn the rules once and apply them on new input answers. Thus, it would be very difficult to achieve robustness when applied to new data over time.

This work proposes an approach where a generic scoring model tries to learn this task of measuring the correctness of each answer continuously with a human in the loop. During the learning stage, the system selects the best samples for the human to grade which would eventually contribute to its knowledge base. Thus, it tries to reduce the human effort of going through all the answers while improving it's understanding of the problem on a cyclical and iterative basis.

Active learning seems to be the best choice for this task as it actively queries the human for grades of the samples it is most uncertain of. Fig 1.2 illustrates a typical workflow of active learning. Active learning is a subfield of machine learning which works under the hypothesis that if the learning algorithm is allowed to choose the data from which it learns it will perform better with less labeled data and training

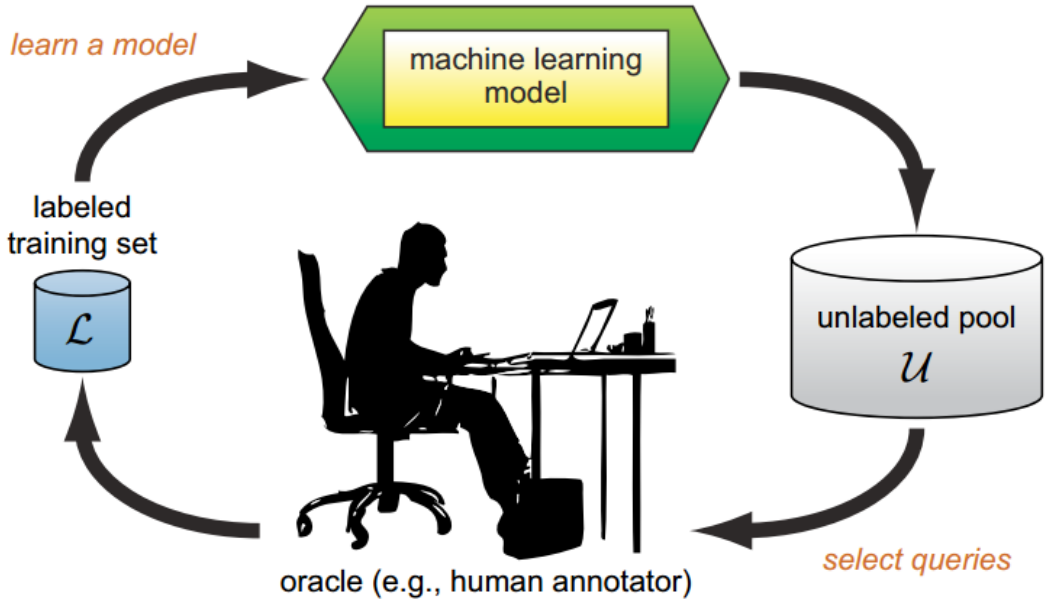


Figure 1.2: Workflow of active learning [22].

[22]. Such a model queries a user / human expert for the labels of certain data samples in such a way that it can learn to produce the desired outputs with higher accuracy. By actively selecting the data samples to label, it reduces a considerable amount of labeled training samples, thus, alleviating the problem of insufficient labeled data which is prevalent in supervised learning approaches. In addition, it would be a solution to deal with the diversity in answers as there is generally no single best response for an open-ended question.

1.1. Problem Statement

- One of the biggest challenge involved in supervised learning in the need for large labeled datasets. Generating a large data for short answer grading is also an obstacle. Many automatic short answer grading systems rely on supervised learning techniques. This, in turn, leads to high annotation cost of labeling the large training data. Also, the workload for labeling the data is more.
- As new data are generated on a regular basis, retraining the model for every new data also have a caveat of high computation cost. Developing a generic model to overcome the deficit of retraining the model for every new answer can solve this issue.
- Short answer grading typically focuses on the content rather than the style of writing. Short answers written by students have a lot of lexical diversity. Active learning can be used to handle the lexical diversity by involving human in training the model.

2. Related Work

2.1. Concept mapping based:

Concept mapping based techniques refers to splitting student answer into various concepts and graded based on the existence or non-existence of individual concepts[4].

- Callear et al.(2001) [5] developed a computer-assisted assessment (CAA) termed as Automated Text Marker (ATM) in which the answers are split into their smallest viable unit of concepts. All the atomic concepts are given some weight for the purpose of grading
- Leacock et al. (2003)[12] developed an automated scoring system called C-rater for ETS(Educational Testing Service) technologies to grade responses to content-based short answers based on the syntactical matching(subject, object, and verb). This uses deep natural language processing to determine the correctness of student responses. The preprocessing steps done by the c-rater systems include spelling correction, determining the grammatical structure of each sentence, resolving pronoun reference and analyzing paraphrases.
- Brill et al. (2002) [3] developed a question answering system called AskMSR. It uses the techniques such as query-reformulation, n-gram mining, filtering, and n-gram tiling. This system reformulates queries as declarative sentence segments to help query-answer matching. The shortcoming of this approach is that it works only when the (exact) content words appearing in a query appears also in the answer.

2.2. Information extraction:

Information extraction system refers to extracting patterns from student answer followed by a series of pattern matching operations that includes regular expression

or parse trees [4].

- Bachman et al. (2002) [1] developed a short answer scoring system called WebLAS. This system extracts regular expressions from a model answer to generate a scoring key. This grading system finds important segments of teacher answers through parsing and prompt the teacher to confirm the weights. It also prompts the teacher to confirm or decline the semantically similar alternatives.

- Mitchelle et al. (2002) [15] discuss a software called AutoMark, which is developed to achieve robust grading of short answers for open-ended questions. This approach employs information extraction techniques to provide computerized marking of short free-text responses. Student answers are first parsed, and then intelligently matched against each mark scheme template, and a mark for each answer is computed.

- Oxford UCLES is an information extraction short answer scoring system that was developed at the Oxford University. which uses hand-crafted patterns by human experts to compare the answers with the model answer by Pulman et al. (2005) [19]. This work compares information extraction with both hand-crafted and machine learning assisted pattern matching. The results of attempting to use machine learning strategies to extract patterns were not satisfactory. It was concluded that hand-crafted patterns perform better than machine learned patterns.

- Jordan and Mitchell (2009) [11] developed a graphical user interface(FreeText Author) for short answer grading whereby the teacher's model answer is converted into syntactic-semantic templates for the student answers to be matched against. The question authors don't have to be well versed in Natural Processing Techniques as the model creates the patterns from the correct answers by its own.

- Raheel Siddiqi (2010)[23] proposes a grading system that works on the structure of students' answer. It simply uses question answer markup language (QAML) to represent the required answer structures. The evaluation process

starts with spell checking and some basic linguistic analysis, then the system matches the student's answer text structure with the required saved structure to compute the final mark.

- Meurers et al. (2011) [13] and Hahn et al. (2012) [7] used semantic analysis to align student and target answers, including functional roles such as subject/object, gives better performance over similarity measures.
- Higgins et al. (2014) [9] work describe the importance and the efficiency of syntactically-informed features like n-gram features, language model features, dependency features, k-nearest neighbor features and discourse segment features in short answer grading.
- Ramachandran et al. (2015) [20] developed a short answer scoring system to provide effective scoring using automatic pattern extraction. Word-order graphs and semantic metrics(Lexico-semantic matching technique) were used to identify important patterns automatically from human-provided rubric texts and top-scoring student answers. Patterns were automatically extracted using two algorithms namely, generating patterns containing unordered content tokens and generating patterns containing sentence structure or phrase pattern information.

2.3. Corpus based:

Corpus-based methods are the statiscal methods that uses large document corpora(wikipedia, google etc.,) to obtain informations like synonyms, degree of similarity,frequency of term pairs etc. [4],

- Mihalcea et al.,(2006) [14] found comparable results to corpus-based measures by using word-to-word similarity measures.
- Nielsen et al., (2009) [18]proposed a dependency-based classification component called Intelligent Tutoring System. In this approach, the instructor

answers are parsed, enhanced, and manually converted into a set of content-bearing dependency triples or facets. The system uses a decision tree trained on part-of-speech tags, dependency types, word count, and other features to attempt to learn how best to classify an answer/facet pair.

- Mohler and Mihalcea (2009) [17] use the similarity between the students' answers and the teacher's answers for automatic short answer grading. This work addresses topics such as comparison of knowledge-based and corpus-based measures of text similarity, evaluation of the effect of domain and size on the corpus-based measures, and a novel technique to improve the performance of the system by integrating automatic feedback from the student answers. Eight knowledge-based measures of semantic similarity and two corpus-based semantic similarities for short answer grading were successfully compared. They also created a dataset of questions, students answers and the grades for those answers given by two human annotators. Making the dataset open-source contributed a lot to other researchers to benchmark their implementations on it.

- Gomaa and Fahmy (2012) [6] use string similarity and corpus-based similarity for automatic short answer grading. In addition to comparing different similarity measures, they compared the usefulness of two different corpora as well and this gave some insight into the useful features of a corpus. This work also reinforced the fact that short answer grading could be formulated as a similarity task.

2.4. Machine learning:

Machine learning based approaches refers to training a model utilizing the features extracted using natural language processing techniques [4].

- Mohler et al. (2011) [16] like his previous approach uses the similarity between the students' answers and the teacher's answers for grading but incorporated

several graph alignment features along with lexical semantic similarity measures This approach uses machine learning techniques and concludes that the student answers can be more accurately graded by incorporating the graph alignment features along with semantic measures. An attempt was also made to align the dependency graphs of the student and the instructor answers in order to make use of a structural component in the automatic grading of student answers.

- Fast, simple, and high-performance short answer grading system was proposed by Sultan et al. (2016) [24] that uses text similarity features such as word alignment, embeddings, question demoting, term weighting and length ratio was proposed. A supervised learning model is trained using this features to predict the grades for short answer grading.

- Basu et al. (2013) [2] developed an approach termed as Powergrading. This approach of divide and conquer the short answers proved to be useful in reducing the number of actions required for grading them by extending the impact of a small number of user actions when grading resources are limited. Their work incorporates clustering in short answer grading to cluster the similar answers together. This work gave an efficient solution to one of the main deficits of automated short answer grading, namely, capturing the modes of misunderstanding among the students' answers.

Evaluation of Active Learning for Short Answer Grading

The research in automatic short answer grading started as matching the concepts in students' answers to that of teacher's model answer and then moved on to match patterns such as regular expressions or parse tree which were written by experts (information extraction). Both of these methods could be categorized under rule-based methods which is best for repeated assessment. The latter part of the research focused into more statistical models such as corpus-based techniques (harnessing information from large corpora to calculate the degree of similarity) and machine learning techniques (features extracted form answers using Natural Language Processing techniques and machine learning models are used to compute the final grade). These models were found to work well for unseen questions and new domains [4].

Requiring large amount of annotated dataset, inconsistent performance, lack of generic scoring models for new dataset, depending on a single best answer, and the need of a human expert's hand to write down the patters are some of the common drawbacks that were inherent in these works. In light of the aforementioned problems, evaluating active learnign for short answer grading would be the next best logical step. In [10], the authors have investigated the applicability of active learning for this task and have presented a brief study on different methods of item selection, seed selection and the influence of different numbers of samples to be labeled at each iteration. In this research and development project, we strive to extend this step of applying active learning to short answer grading by exploring different techniques and algorithms in natural language processing and active learning.

3. Project Plan

3.1. Work Packages

The bare minimum will include the following packages:

WP1 Problem formulation and proposal

- Problem formulation
- Proposal

WP2 Natural language processing and active learning

- Understanding concepts
- Hands on Natural language processing tools and libraries
- Understanding active learning algorithms

WP3 Literature survey

- Collect most relevant papers
- Shortlist papers that include computerized short answer grading
- Shortlist papers that include active learning in the context of NLP

WP4 Working on datasets

- Reviewing existing datasets
- Compile new dataset

WP5 Evaluating various NLP features

- Evaluating features based on semantic similarity
- Evaluating features based on entailment
- Evaluating features based on pattern matching

- Evaluating features based on lexical similarity

WP6 Evaluating various active learning strategies

- Evaluating various query strategy frameworks
- Empirical analysis of active learning for short answer grading

WP7 Implementation

- Minimum working model
- Testing and evaluating performance
- Improvements

WP8 Developing graphical user interface(GUI)

- User research
- Design and protyping
- Evaluation

WP9 Final Report

- Writing report
- First draft of the report
- Corrections in report
- Final report

3.2. Milestones

M1 Comprehensive state of art

M2 Determining the best NLP features for short answer grading

M3 Determining the best active learning strategy

M4 Delivering a working model

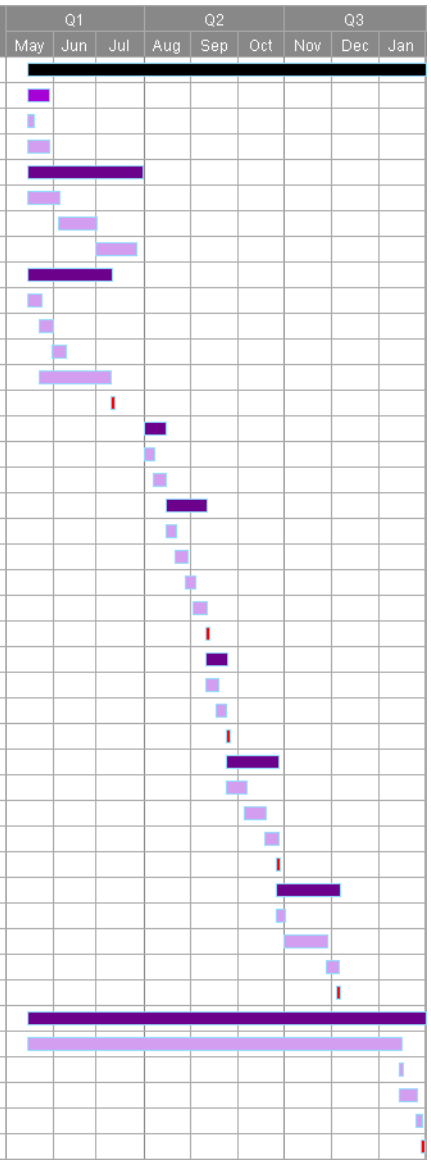
M5 Incorporating model into an interface

M5 Report submission

3.3. Project Schedule

Figure 3.1: Gantt chart

	TASK	Start	Finish
1	Research and Development project	05/15/18	01/30/19
2	WP1 : Problem formulation and proposal	05/15/18	05/28/18
3	Problem formulation	05/15/18	05/18/18
4	Proposal	05/15/18	05/28/18
5	WP 2: Natural Language Processing and Active learning	05/15/18	07/30/18
6	Understanding concepts	05/15/18	06/04/18
7	Hands on Natural language processing tools and libraries.	06/04/18	06/29/18
8	Understanding active learning algorithms	06/29/18	07/26/18
9	WP 3: Literature Survey	05/15/18	07/10/18
10	Collect most relevant papers	05/15/18	05/23/18
11	Shortlist papers that include computerized short answer grading	05/23/18	05/31/18
12	Shortlist papers that include active learning in the context of NLP	05/31/18	06/08/18
13	Finalize review of top 12 papers	05/23/18	07/10/18
14	Milestone 1: Comprehensive state of art	07/10/18	07/10/18
15	WP 4: Working on Datasets	08/01/18	08/14/18
16	Reviewing existing datasets	08/01/18	08/07/18
17	Compile new dataset	08/07/18	08/14/18
18	WP 5: Evaluating various features	08/15/18	09/11/18
19	Evaluating features based on Semantic similarity	08/15/18	08/21/18
20	Evaluating features based on Entailment	08/21/18	08/28/18
21	Evaluating features based on Pattern matching	08/28/18	09/04/18
22	Evaluating features based on Lexical similarity	09/03/18	09/11/18
23	Milestone 2: Determining the best NLP features for short anser grading	09/11/18	09/11/18
24	WP 6: Evaluating various active learning strategies	09/11/18	09/24/18
25	Evaluating various query strategy frameworks	09/11/18	09/18/18
26	Empirical analysis of active learning for short answer grading	09/18/18	09/24/18
27	Milestone 3: Determining the best active learning strategy	09/24/18	09/24/18
28	WP 7: Implementation	09/24/18	10/26/18
29	Minimum working model	09/24/18	10/05/18
30	Testing and evaluating the performance	10/05/18	10/18/18
31	Improvements	10/18/18	10/26/18
32	Milestone 4: Delivering a working model	10/26/18	10/26/18
33	WP 8: Developing Graphical User Interface(GUI)	10/26/18	12/06/18
34	User research	10/26/18	11/01/18
35	Design and prototyping	11/01/18	11/28/18
36	Evaluation	11/28/18	12/04/18
37	Milestone 5: Incorporating the model into an interface	12/04/18	12/04/18
38	WP 9: Final Report	05/15/18	01/30/19
39	Writing report	05/15/18	01/14/19
40	First draft of report	01/14/19	01/14/19
41	Corrections in report	01/14/19	01/25/19
42	Final report	01/25/19	01/28/19
43	Milestone 6: Report submission	01/28/19	01/28/19



3.4. Deliverables

3.4.1. Minimum Viable

- Survey the existing approaches for computerized short answer grading
- Review existing datasets for automatic short answer grading.
- Compile new datasets from different in-house domain

3.4.2. Expected

- Evaluate the quality of different features for interactive short answer grading
- Evaluate Active Learning strategies for Short-Answer grading
- Implement a working model of 100 clicks for 1000 grades.

3.4.3. Desired

- Integrate the best model with a graphical user interface(GUI)

3. References

[1] Lyle F Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Michael J Pan, Chris Salvador, and Yasuyo Sawaki. A reliable approach to automatic assessment of short answer free responses. *Proceedings of the 19th international conference on Computational linguistics* -, 2:1–4, 2002. doi: 10.3115/1071884.1071907. URL `http://portal.acm.org/citation.cfm?doid=1071884.1071907`.

[2] Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402, 2013.

[3] Eric Brill, Susan Dumais, and Michele Banko. An analysis of the AskMSR question-answering system. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, 10(July):257–264, 2002. doi: 10.3115/1118693.1118726. URL `http://portal.acm.org/citation.cfm?doid=1118693.1118726`.

[4] Steven Burrows, Iryna Gurevych, and Benno Stein. *The eras and trends of automatic short answer grading*, volume 25. 2015. ISBN 4059301400268. doi: 10.1007/s40593-014-0026-8.

[5] David Callear, Jenny Jerrams-Smith, and Victor Soh. Caa of short non-mcq answers. 2001.

[6] Wael H Gomaa and Aly A Fahmy. Short answer grading using string similarity and corpus-based similarity. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 3(11), 2012.

[7] Michael Hahn and Detmar Meurers. Evaluating the meaning of answers to reading comprehension questions a semantics-based approach. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 326–336. Association for Computational Linguistics, 2012.

[8] Derrick Higgins, Jill Burstein, and Daniel Marcu. Evaluating Multiple Aspects of Coherence in Student Essays. *Hlt-Naacl*, pages 185–192, 2004. URL `http://acl.ldc.upenn.edu/hlt-naacl2004/main/pdf/16{_}Paper.pdf`.

[9] Derrick Higgins, Chris Brew, Michael Heilman, Ramon Ziai, Lei Chen, Aoife Cahill, Michael Flor, Nitin Madnani, Joel Tetreault, Daniel Blanchard, Diane Napolitano, Chong Min Lee, and John Blackmore. Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. 2014. URL `http://arxiv.org/abs/1403.0801`.

[10] Andrea Horbach, Alexis Palmer, and Leibniz Sciencecampus. Investigating Active Learning for Short-Answer Scoring. pages 301–311, 2016.

[11] Sally Jordan and Tom Mitchell. e-assessment for learning? the potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology*, 40(2):371–385, 2009.

[12] Claudia Leacock and Martin Chodorow. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405, 2003. ISSN 00104817. doi: 10.1023/A:1025779619903.

[13] Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. Evaluating answers to reading comprehension questions in context: Results for german and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9. Association for Computational Linguistics, 2011.

[14] Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, 2006.

[15] Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. Towards robust computerised marking of free-text responses. 2002.

[16] M. Mohler, R. Bunescu, and R. Mihalcea. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. pages 752–762, 2011. URL `http://ace.cs.ohiou.edu/{~}razvan/papers/acl11.pdf`.

[17] Michael Mohler and Rada Mihalcea. Text-to-text Semantic Similarity for Automatic Short Answer Grading. *Proceedings of the 12th Conference of the*

European Chapter of the Association for Computational Linguistics (EACL '09), (April):567–575, 2009. doi: 10.3115/1609067.1609130. URL http://dl.acm.org/citation.cfm?id=1609130.

[18] Rodney D Nielsen, Wayne Ward, and James H Martin. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(4):479–501, 2009.

[19] Stephen G. Pulman and Jana Z. Sukkarieh. Automatic short answer marking. *EdAppsNLP 05 Proceedings of the second workshop on Building Educational Applications Using NLP*, (June):9–16, 2005. doi: 10.3115/1609829.1609831. URL http://dl.acm.org/citation.cfm?id=1609829.1609831.

[20] Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–106, 2015. doi: 10.3115/v1/W15-0612. URL http://aclweb.org/anthology/W15-0612.

[21] Shourya Roy, Sandipan Dandapat, Ajay Nagesh, and Narahari Y. Wisdom of Students: A Consistent Automatic Short Answer Grading Technique. *Proceedings of the 13th International Conference on Natural Language Processing*, pages 178–187, 2016. URL http://www.aclweb.org/anthology/I/W16/W16-5124.

[22] Burr Settles. Active Learning Literature Survey. *Machine Learning*, 15(2): 201–221, 2010. ISSN 00483931. doi: 10.1.1.167.4245.

[23] Raheel Siddiqi, Christopher J Harrison, and Rosheena Siddiqi. Improving teaching and learning through automated short-answer marking. *IEEE Transactions on Learning Technologies*, 3(3):237–249, 2010.

[24] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. Fast and Easy Short Answer Grading with High Accuracy. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075, 2016. doi: 10.18653/v1/N16-1123. URL http://anthology.aclweb.org/N/N16/N16-1123.pdf.

[25] Hao Chuan Wang, Chun Yen Chang, and Tsai Yen Li. Assessing creative problem-solving with automated text grading. *Computers and Education*, 51 (4):1450–1466, 2008. ISSN 03601315. doi: 10.1016/j.compedu.2008.01.006.