Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review: 16th Ibero-American Conference on AI, Trujillo, Peru, November 13-16, 2018, Proceedings

Chapter · November 2018

DOI: 10.1007/978-3-030-03928-8_31

2 authors:

Lucas Busatta Galhardi
Universidade Estadual de Londrina
4 PUBLICATIONS **1** CITATION
SEE PROFILE

Jacques Brancher
Universidade Estadual de Londrina
81 PUBLICATIONS **129** CITATIONS
SEE PROFILE

Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review

Lucas Busatta Galhardi(✉) and Jacques Duílio Brancher

Computer Science Department, State University of Londrina, Londrina, PR, Brazil
{lucasbgalhardi,jacques}@uel.br

Abstract. In this systematic review, we investigate the automatic short answer grading (ASAG) field, which focuses on assessing short natural language responses to questions in an automatic way. Short answers have been recognized as a tool to perform a deeper assessment of the student's knowledge than, for example, multiple choice questions. Automatically scoring short responses can be used as an important resource to the educational field, where the student's answers can be easily, fairly and quickly evaluated for feedback purposes in, for instance, massive open online courses, in which precision and agility are required. We conducted the research by including only works that employed machine learning methods in order to solve the problem. The final selection considering all criteria selected 44 papers reporting different ASAG systems. Those studies were analyzed by answering the proposed research questions, extracting: the nature of datasets, used natural language processing and machine learning techniques, features selected to create the models and the results obtained from their systems' evaluation.

Keywords: Automatic grading · Short answers · Machine learning
Systematic review

1 Introduction

Different tools as VLEs (Virtual Learning Environments), MOOCs (Massive Open Online Courses) and CBA (Computer-Based Assessment) have recently improved their popularity as they provide a new resource for teachers so they can define and manage their didactic resources and expose multimedia contents to help students in their learning process. In addition, these environments can have assessment systems that can support teachers in evaluating many students.

In their learning process, students will have to experience evaluations to demonstrate their acquired knowledge. However, teachers usually find the task of assessing respondents' answers very time-consuming. Also, students may have to wait for a long time to receive feedback on their responses and, when they finally get it, the grade can be different from another classmate's, who has given a very similar answer [13,16].

G. R. Simari et al. (Eds.): IBERAMIA 2018, LNAI 11238, pp. 380–391, 2018.
https://doi.org/10.1007/978-3-030-03928-8_31

Computer-based assessment came to address these issues and improve other aspects of learning by automating the evaluation process. Some of the benefits of automatic assessments are: criteria is formalized, can provide faster feedback to both teacher and student, can save professors' time so they can use it to work better, and allows teachers to easily follow the class performance [1,7].

There are many different types of questions used to evaluate students. In this work, the focus of interest is in short answers. In addition to the answer length, short answers are written in some natural language and recalls to external knowledge outside the question statement. Important to state that short answers differ from essays as the last can have from two paragraphs to several pages, the evaluation focus is on the writing style and it has a more open scope [1].

To the best of our knowledge, there are six literature reviews in the automatic short answer field. However, two of them [14,18] have essay systems along with the short answer ones. [3] reviews only studies that used an Information Extraction approach for ASAG. The other three [1,15,20] reviews only automatic short answer grading systems and without restrictions on the approach, especially [1] that has the most recent and comprehensive review. Despite all these reviews, this work proposes a systematic approach for conducting the literature review based on [4] guidelines. Also, this paper is limited to review only studies that used machine learning (ML) approaches to solve ASAG, as ML is indicated to be the new trend for ASAG [1].

This paper is organized following the steps of a systematic review [4]. In Sect. 2 the methodology of the research is presented, alongside with its planning and conduction. Section 3 shows the results obtained, answering the research questions. Finally, the conclusions of the work can be seen in Sect. 4.

2 Methodology

According to [4], systematic reviews can identify, evaluate and interpret all available research concerning a specific research question or topic area. It can present a fair review of the research topic by using a rigorous, trustworthy and auditable methodology. The process defines a research protocol which researchers have to follow when conducting the research. The detailed and replicable aspects of systematic reviews are their main advantage since other researchers can follow the conducted process and even repeat the research obtaining the same results (considering same period). Common steps in this kind of reviews are: Planning, Conducting and Reporting.

2.1 Planning

The planning stage of a systematic review creates the review protocol, which specifies the methods that will be used before starting the review. Such early definition helps researchers avoid a biased process [4].

This systematic review seeks to study, explore and understand the current state-of-the-art of automatic short answer grading with focus on works that used

machine learning approaches to handle ASAG. The elaborated research questions to address the review objective are: **RQ1:** "What is the nature of datasets?", **RQ2:** "Which natural language processing and machine learning techniques are used?", **RQ3:** "What are the selected features?" and **RQ4:** "What are the achieved results?".

The sources of this systematic review are the following nine online databases: *LearnTechLib*, *Microsoft Research*, *ScienceDirect*, *IEEE Xplore Digital Library*, *ACM Digital Library*, *Scopus*, *Springer*, *Semantic Scholar* and *Keele University Library*.

Some preliminary research was made to determine the most used words for the subject matter. Similar words were grouped and a search string using boolean operators was created and refined using one of the online databases until it was considered good despite all the possibilities that those keywords creates. The search string composed by the keywords and synonyms is:

("automatic assessment" OR "automatic scoring" OR "automatic marking" OR "automatic grading") AND (short OR "short answer" OR "free text" OR free OR text) AND (response OR question OR answer)

The Inclusion (I), Exclusion (E) and Quality (Q) criteria applied to each reviewed work is detailed in Table 1. Filtering of papers were performed strictly considering these criteria.

Table 1. Criteria

Type	Criteria	Type	Criteria
I	Studies written in English	E	Studies written in another language than English
I	Article, Conference or Methodology papers	E	Studies that do not match the research questions
I	Studies relevant to the subject matter	E	Papers about the same study or system
E	Secondary studies	Q	Are most of the research questions answered?
E	Semi automatic approaches	Q	Is the research methodology properly exposed?
E	Studies that assess essay length answers	Q	Are all the used techniques properly described?

2.2 Conducting

In this stage, the defined search string was used to perform the search in the nine online databases. The results were exported from the databases in some bibliography reference format like bibtex, RIS or CSV. The sum of the retrieved results from the nine databases was 6789. From those papers, 1562 consisted of duplicated papers due to papers that are in more than one online database. The large initial number of results is due to the broad range that the string creates. The search string fits good for getting wanted results, but it also gets other areas of research like medicine. For instance, one possible form that the search string can assume is "automatic scoring short response" which can refer to analyses about the performance of medical tools for measuring short body responses like stimulus or impulses.

In sight of the 5227 remaining papers, we applied the inclusion, exclusion and quality criteria in three levels. First, the inclusion and exclusion criteria were applied only in the title and keywords (and abstract if necessary), which returned 182 papers. These were related with the research area but not necessarily addressing all criteria. In the second stage, the title, keywords and abstract were carefully read to identify relevant studies. That stage left 112 remaining papers, which were all downloaded with the exception of papers that we did not have access to (five). These 107 papers were read using a skimming approach, passing by the title, abstract, introduction and section and sub-section headings. The results, figures and references were also glanced through to determine if the paper would pass to the next step, which resulted in 75 remaining papers.

Among the 75 studies, we selected only the papers that used a machine learning approach to handle ASAG. We did so by looking at each paper's abstract and introduction and searching for keywords like *"machine learning"*, *feature*, *classifier*, *regression* and similar, which left us with 18 papers.

Knowing that due to the nature of the field a variety of keywords could be used in the studies even though not being present in those 18 papers, we looked up for studies using the machine learning keywords in the six identified review papers mentioned in the introduction (Sect. 1). We gathered 26 more papers by looking in their references and 14 more looking at papers that cite those reviews in Google Scholar or in the references of the 14 recently acquired papers (from 2014 to 2016), which left us with 58 papers. Then, the remaining 58 papers were fully read in order to apply the quality criteria and at the same time do the data extraction step. After the quality filter, the number of papers was finally established in 44 as seen in Table 2. From these 44 studies, the answers of the research questions were obtained, the data was summarized and the results created.

Table 2. Selected Papers (IDs and References)

ID	Reference	ID	Reference	ID	Reference
1	[Rosé et al. 2003]	16	[Peters and Jankiewicz 2012]	31	[Higgins et al. 2014]
2	[Pulman and Sukkarieh 2005]	17	[Sil et al. 2012]	32	[Aldabe et al. 2015]
3	[Makatchev and VanLehn 2007]	18	[Dzikovska et al. 2012]	33	[Sakaguchi et al. 2015]
4	[Nielsen et al. 2008]	19	[Madnani et al. 2013]	34	[Nye et al. 2015]
5	[Wang et al. 2008]	20	[Levy et al. 2013]	35	[Luo et al. 2015]
6	[Lee et al. 2009]	21	[Heilman and Madnani 2013]	36	[Sorour et al. 2015]
7	[Sukkarieh 2010]	22	[Jimenez et al. 2013]	37	[Ramachandran et al. 2015]
8	[HOU and TSAO 2011]	23	[Bicici and van Genabith 2013]	38	[Zesch and Heilman 2015]
9	[Mohler et al. 2011]	24	[Gleize and Grau 2013]	39	[Zhang et al. 2016]
10	[Meurers et al. 2011a]	25	[Ott et al. 2013]	40	[Magooda et al. 2016]
11	[Meurers et al. 2011b]	26	[Kouylekov et al. 2013]	41	[Sultan et al. 2016b]
12	[Zbontar 2012]	27	[Horbach et al. 2013]	42	[Roy et al. 2016]
13	[Tandalla 2012]	28	[Leeman-Munk et al. 2014]	43	[Liu et al. 2016]
14	[Conort 2012]	29	[Gomaa and Fahmy 2014]	44	[Sultan et al. 2016a]
15	[Jesensky 2012]	30	[Moharreri et al. 2014]		

3 Results

In this section, the results of the systematic review are presented. The next subsections answers to the research questions defined in Sect. 2.1. In the remainder of this work, references to works from Table 2 are made by the following way: (ID) (e.g. (16) or (20, 21, 22)). When referring to the number of papers that used a specific technique, they are presented in this way: technique {Number}.

3.1 RQ1: Nature of Datasets

There is a great variety in the nature of datasets used by each reviewed paper. They vary in many aspects such as in the topic of the questions, language, student characteristics, grading scale, answers average size and the number of questions, answers and reference answers.

We identified 28 different datasets among the 44 reviewed papers. Some works evaluates in more than one dataset and some evaluates in one of the six publicly available datasets, released to stimulate new researches in the ASAG field. In 2011, three datasets were published: the Texas [12][1] and the CREE and CREG from the CoMic project [9,10][2]. In 2012 and 2013, two ASAG competitions took place, highly increasing the number of researches on the field. They were the 2012 Automated Student Assessment Prize (ASAP) from the Kaggle website[3] and the SemEval 2013 Task 7, the Joint Student Response Analysis and Eighth Recognizing Textual Entailment Challenge, who released two datasets, Beetle and SciEntsBank [2][4].

Regarding the subject matter of the questions, science related topics are the most common (57%) in the studies. Some studies only reports generic science (D6) whereas some specify like Scientific Inquiry (43), Biology (30), Physics (39) and Electronics (D5). Another greatly used kind of question is the reading comprehension type, present in 21% of the datasets (19, 33, D2). Computer Science related topics are also present (11%) in some works dealing with programming basic concepts, introductory and formal language content (8, 35, D1). Other topics comprise of Philosophy (29), US citizenship test (38) and interdisciplinary content (D4).

Concerning the language of the datasets, 75% of them are in English. The other 25% are distributed between Chinese (5, 6), Arabic (29, 40), Japanese (35), Hindi (42) and German (D3).

The respondents' educational level is reported in 89% of the papers. From those, 56% are in school as some report being in "middle school", "high school", "grade x to y" or the students' age. The other large group (36%) is in college, usually without specified age or year. Two works also deal with Second (Foreign) Language Studies (in English and German, (D2, D3)).

[1] http://web.eecs.umich.edu/~mihalcea/downloads.html.
[2] www.uni-tuebingen.de/en/research/core-research/collaborative-research-centers/sfb-833/section-a-context/a4-meurers/software-resources-and-corpora.html.
[3] www.kaggle.com/c/asap-sas.
[4] www.cs.york.ac.uk/semeval-2013/task7/index.php%3Fid=data.html.

The number of questions present in the 28 datasets varies between 1 (1, 6) and 482 (39) prompts. Usually works with less questions have more answers and works with more questions have less answers. The number of answers per question is distributed in this way: 18% have up to 12 answers, 32% have between 12 and 99 (dozens), the greater part (36%) have hundreds of answers and 14% of the works have more than 1000 answers (the maximum being 2295 (D4)).

The number of reference answers is not reported in one third of the works. Most studies used 1 (sometimes more) reference answers for comparisons with the students' ones. Some works does not make use of reference answers and some describe the use of concepts (a very short sentence (1, 3, 5, 7, 33)).

The grading scale is reported in three possible formats: number of matches, a range of points or the number of classes. Some datasets have more than one grading scale for the same questions. Two, three, four and five points or classes correspond to the majority of works (D1, D2, D3, D4, D5, D6). Some have a 10 point scale (6,29) and one isolated work has a 30 point range scale (5).

Only half the studies reports the responses' length. They are presented in terms of average number of words, sentences, tokens, lines or a written estimate (like "from short verb phrases to several sentences", "from a couple of words to several sentences", "up to around five lines" and "one to few sentences" (2, 4, 28, D3)). The number of sentences varies from 1 to 7 and the number of words from 7 to 63 in average.

3.2 RQ2: Natural Language Processing Techniques

In order to model answers in easier ways for the computer to interpret them, some Natural Language Processing (NLP) techniques are used. They are employed in the prepossessing stage to prepare the answers for the feature extraction. A reasonable number of different techniques are used in the reviewed studies to perform this step.

Not all works describe using NLP for preprocessing and we can assume that either they did not use these techniques or considered them not sufficiently relevant to report. We found the use of more than 10 different techniques among the 44 works. The basic and self-explained are: punctuation, numbers and other symbols removal, acronym expansion, sentence segmentation, case normalization and tokenization.

Besides these essential techniques, we have some more that aggregates value to the answer by acting in the lexical, syntactical or semantic level. Techniques applied considering only the words by themselves are stopword removal (to not account for too common words {12}), spelling correction (to increase the chances of matches with another words {14}) and stemming and lemmatization (two processes to shorten words by reducing their morphological variance in order to increase matches between words {17 and 10}).

Syntactically, part-of-speech is usually {18} performed to account for the syntactic role of each word in the sentence. To improve this technique, entities with more than one word can be tagged together as a single word, using chuncking, present in {6} works. Also, chuncking helps this process by identifying structures

inside the sentence that can be grouped. Simple part-of-speech tagging would tag "the man" as an article and a noun. Chuncking could identify them as a single entity to be tagged. Also, the reviewed works performs syntactic parsing {16} to identify important structural aspects of sentences.

Finally, in the semantic level, the reviewed works uses mainly two techniques. The first is Semantic Role Labeling (17, 33), a process that assigns labels to the roles that words represent in sentences considering their semantic aspect. The another technique is to use WordNet (9, 25, 27, 37), a semantic network, to retrieve semantic synonyms for the words in the answers, improving their capabilities to match with other answers.

3.3 RQ2: Machine Learning Algorithms

A specific kind of task addressed by machine learning algorithms is the supervised learning, where the computer is presented with a set of example inputs and outputs and the challenge of the algorithm is to discover a general rule that models the maximum number of world samples. Real world problems are not easily classified and thus the goal is to reach as close to 100% accuracy as possible.

In ASAG, machine learning is used to solve a classification or regression problem, usually using supervised learning. The model is built upon the answers of students and the correspondent grades assigned by a teacher. The objective is to predict which score should be assigned to a new answer. By collecting the data and building and evaluating the model, an automatic short answer grading system can be built and be ready to use with some specific degree of confidence.

Four different approaches were identified in the selected papers. Firstly, three works used Artificial Neural Networks (15, 35, 39) and one used Deep Belief Networks (40), algorithms inspired by the neural connexions of the human brain. Secondly, one work (36) approached the ASAG task as an unsupervised learning, using K-Means to find clusters of answers.

Then, the most common approach found was to used a classification or regression algorithm to build the model. The ones used were the following: Support Vector Machine {24}, Decision Tree {9}, Logistic Regression {7}, Ridge Regression {6}, Naive Bayes {5}, K-Nearest Neighbors {5} and Linear Regression {2}. Finally, in order to boost the results, several papers reported the use of ensemble learning algorithms: Stacked Generalization {7}, Random Forests {6}, Gradient Boosting Machine {6}, Bagging {3} and Adaptive Boosting {1}.

3.4 RQ3: Features

In ASAG, a large number of different features have been used in the literature to improve the results. They can be grouped in three categories: Lexical, Syntactic and Semantic. Each subsequent subsection will explore each of them and present some of the most used and representative features.

Lexical. The most common model used in the reviewed works is Ngrams, with the basic model present in more than 70% of the reviewed studies. This model takes the input text and groups n words together. Considering the phrase "The man fishes on the lake", a bigram would produce: "The man", "man fishes", "fishes on", "on the" and "the lake". Ngrams are not restricted to words but can also be used with letters. In the literature the n of ngrams varies from one to six. A special case of the word ngrams is when n equals 1, which is commonly known as Bag-of-Words (BoW).

Ngrams can use three approaches to be represented as document-term matrices. One will use the simple binary presence (1) or absence (0) of the ngram. Another approach is to count the number of times that the word appears (the term frequency). Finally, the weight of a word can be given by the Term Frequency-Inverse Document Frequency (TF-IDF), a metric of frequency used to penalize general words by decreasing their value.

Some studies use established metrics that have ngrams underneath like BLEU (19, 21, 31, 33) and ROUGE (19). Another growing representation of words as vectors in recent studies is Word2Vec (33, 35, 42, 44). One of its main difference from simple ngrams is in the concept of word embeddings, employed as a dense alternative to ngrams sparse models.

Another group of features are those who uses some metric to measure the lexical similarity between answers. These string similarities measures are created using Cosine, Overlap, Sorensen, Levenshtein, Hamming and similar. Another similarities used in the reviewed works can be seen in [19] survey.

Other greatly used features are text statistics like response's length, count of words, count of unique words, verb counts, number of characters, sentences, word average length and similar.

Syntactic. Representing the syntactical features, reviewed studies uses phrase and dependency ngrams. Phrase ngrams are the combination of the main verb and their noun phrase. Dependency ngrams are made of the syntactical relations between words. These dependencies can be obtained from a natural language parser like Stanford Parser [5]. The usual format is a triple containing two words and their relation dependency. These triples are used as features in 23% of the reviewed works.

Another important syntactic feature is the similarity between student's and reference's answer's part-of-speech tags. PoS tags represent the words' class and what is the behavior of that group in syntactic terms. Therefore, if two answers share many PoS tags they have a similar structure and more likely to be meaning the same.

Semantic. A greatly used approach in the semantic level are knowledge-based features. Present in 25% of the reviewed studies, it is used to calculate similarity between words using a knowledge source. The similarity in this case means the

[5] https://nlp.stanford.edu/software/lex-parser.shtml.

semantic properties of the words [8]. The most used source of knowledge similarity is WordNet [11]. WordNet is a lexical database of English words and their part-of-speeches. WordNet also models semantics by grouping and linking words with similar meanings in Synsets. Some similarities measures that can be used in WordNet are Resnik, Lin, Jiang-Conrath, Leacock-Chodorow, Hirst-St.Onge, Wu-Palmer, Banerjee-Pedersen and Patwardhan-Pedersen.

Another group of semantic features is composed by textual entailment (TE). TE consists of judging if one text can be inferred by another text. Some of the reviewed works interpret the ASAG problem as a textual entailment recognition problem (like [17]). One of them is the study of [6] that uses a TE recognition engine: BIUTEE. This tool tries to convert one text to another by applying a series of transformations. This is used in ASAG by making the student answers the test instance and the reference answer the hypothesis. As output, BIUTEE will return numerical entailment confidence values that are used as features. In [5], the EDITS system is used to generate these features.

The last group of semantic information features are the corpus-based similarity measures. They use large corpus to obtain statistical information that can later be used to calculate a relation value between words and documents. Three different similarity measures of this kind were identified in the reviewed works: Latent Semantic Analysis (LSA) (34, 36, 39, 42), Explicit Semantic Analysis (ESA) (9, 20) and "Extracting DIStributionally similar words using CO-occurrences" (DISCO) (29, 40).

3.5 RQ4: Systems' Evaluation

This research question deals with the evaluation of the proposed systems. Each of the 44 reviewed papers describing a specific ASAG methodology was evaluated by the authors on the datasets presented in Subsect. 3.1. Some studies evaluate in only one dataset while others report experiments in more than one. An important difference between researches is the use of private or public datasets.

Different datasets, evaluation metrics and number of classes preclude fair and direct comparisons. In this work, the evaluations of all systems were analyzed, but due to available space, only works that evaluated on public datasets are reported. From the six publicly available datasets presented in this work, only CREE and CREG were left behind due the fact that only one more work evaluates on CREG (with worst results) and none in CREE (only considering the reviewed papers) regarding the original work that introduced the dataset.

In Table 3, the evaluations from systems on public datasets are presented. It is composed with an ID column that refers to the analyzed work from Table 2 and with the evaluation score obtained. The table is divided by dataset and it is in chronological order. Scores of the same dataset uses the same metric, but each dataset has a different metric, presented as follows.

First, we have results for the Texas dataset [12] presented in terms of Pearson's correlation coefficient. The work from Mohler (9) was the first published as the releasing research and in the following years four more works filling our criteria were found and reported comparable results. It is possible to notice that

Table 3. Systems' evaluations

Texas		ASAP-SAS		Beetle UA 5-way		SciEntsBank UA 5-way	
ID	Score	ID	Score	ID	Score	ID	Score
9-Mohler2011	0,518	12-Zbontar2012	0,7711	20-Levy2013	0,448	21-Heilman2013	0,625
37-Ramachandran2015	0,610	13-Tandalla2012	0,7717	21-Heilman2013	0,705	22-Jimenez2013	0,537
40-Magooda2016	0,550	16-Peters2012	0,7653	22-Jimenez2013	0,558	25-Ott2013	0,598
41-Sultan2016	0,630	31-Higgins2014	0,7680	23-Bicici2013	0,547	40-Magooda2016	0,470
44-Roy2016	0,564	37-Ramachandran2015	0,7800	24-Gleize2013	0,505	41-Sultan2016	0,582
		38-Zesch2015	0,6700	25-Ott2013	0,675	44-Roy2016	0,672

a great improvement has been done, with the work of Sultan (41) exceeding the original work by 0,112.

Then, in 2012 the ASAP-SAS competition from Kaggle released a new dataset and the five winners reported methodology papers. In Table 3, the top three performers from the competition are compared to three more works, using mean quadratically weighted Kappa statistics. Higgins (31) suggests that his work did not perform as good as the winner Tandalla (13) because he did not performed any optimizations, especially of the question-specific type, in order to present a more generalized model. Ramachandran (37) recognizes that Tandalla's performance was greatly helped by manually crafted regular expressions to match simple patterns expected by each question. Thus, Ramachandran proposed a technique to automatically generate these regular expressions in order to fully automate the process and compare results, which showed to be a good strategy as it obtained better results.

In 2013, the SemEval Task 7 competition took place as the Joint Student Response Analysis [2], releasing two more public datasets, Beetle and SciEntsBank. In Table 3 weighted averaged F1 scores achieved in the Beetle dataset are presented, considering the five-way task and the Unseen Answers scenario (more details in [2]). Papers from ID 20 to 25 are from the competition itself and only one more work was found evaluating on the Beetle dataset. However, it did not reported their results in a comparable setting. The best results were obtained by Heilman and Ott, far surpassing the others.

Finally, following the SemEval competition, we report results for the SciEntsBank dataset, also in the five-way task, for the Unseen Answers scenario and with weighted averaged F1 scores. In Table 3, the three top performers in the competition are compared to three more recent works. Magooda (40) did not obtained better results from the three best performers, whereas Sultan (41) obtained similar scores. However, Roy (44) obtained a much higher score, using an ensemble technique that combines bag-of-words modeling with similarity measures extracted from answers, in a similar but improved approach from Heilman (21).

4 Conclusions

This work's objective was to perform a systematic review in the research field of automatic short answer grading with works using a ML approach. We began by exposing the ASAG scenario and its importance on the educational field and specially in virtual environments. Then, the research was planned and conducted following systematic review's guidelines. The final selection resulted in 44 papers and four research questions were answered based on them.

We first explored the data used in ASAG research, considering many aspects from the language to number of questions, answers, etc. Then, we looked at which natural language processing and machine learning techniques are the most used in the field. After that, we presented the core of the research, how answers are modeled in order to extract features that can predict their scores. And finally, we showed how researchers evaluated their systems and how they can (or can not) be compared to each other.

All presented results shows the essence and evolution of ASAG research using machine learning methods. Public datasets are available for not too long ago, and research in the field is open to new techniques, datasets and specially to deep learning, that has been recently contributing to a lot of different areas and still very underexplored in ASAG.

References

1. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. Int. J. Artif. Intell. Educ. **25**, 60–117 (2015)
2. Dzikovska, M., et al.: SemEval-2013 task 7: the joint student response analysis and 8th recognizing textual entailment challenge. In: Seventh International Workshop on Semantic Evaluation, pp. 263–274 (2013)
3. Hasanah, U., Permanasari, A.E., Kusumawardani, S.S., Pribadi, F.S.: A review of an information extraction technique approach for automatic short answer grading. In: International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 192–196. IEEE (2016)
4. Kitchenham, B.: Procedures for performing systematic reviews. Keele, UK, Keele Univ. **33**(2004), 1–26 (2004)
5. Kouylekov, M., Dini, L., Bosca, A., Trevisan, M.: Celi: EDITS and generic text pair classification. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), vol. 2, pp. 592–597 (2013)
6. Levy, O., Zesch, T., Dagan, I., Gurevych, I.: UKP-BIU: similarity and entailment metrics for student response analysis. In: Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), vol. 2, pp. 285–289 (2013)
7. Liu, O.L., Rios, J.A., Heilman, M., Gerard, L., Linn, M.C.: Validation of automated scoring of science assessments. J. Res. Sci. Teach. **53**(2), 215–233 (2016). https://doi.org/10.1002/tea.21299
8. Magooda, A., Zahran, M.A., Rashwan, M., Raafat, H., Fayek, M.B.: Vector based techniques for short answer grading. In: International Florida Artificial Intelligence Research Society Conference Ahmed, pp. 238–243 (2016)

9. Meurers, D., Ziai, R., Ott, N., Bailey, S.M.: Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. Int. J. Continuing Eng. Educ. Life-Long Learn. **21**(4), 355 (2011). https://doi.org/10.1504/IJCEELL.2011.042793
10. Meurers, D., Ziai, R., Ott, N., Kopp, J.: 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, pp. 1–9 (2011)
11. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
12. Mohler, M., Bunescu, R., Mihalcea, R.: Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 752–762 (2011)
13. Passero, G., Haendchen Filho, A., Dazzi, R.: Avaliação do uso de métodos baseados em lsa e wordnet para correção de questões discursivas. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), vol. 27, p. 1136 (2016)
14. Pérez-Marín, D., Pascual-Nieto, I., Rodríguez, P.: Computer-assisted assessment of free-text answers. Knowl. Eng. Rev. **24**(04), 353–374 (2009)
15. Roy, S., Narahari, Y., Deshmukh, O.D.: A perspective on computer assisted assessment techniques for short free-text answers. In: Ras, E., Joosten-ten Brinke, D. (eds.) CAA 2015. CCIS, vol. 571, pp. 96–109. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27704-2_10
16. Santos, J.C.A.D., et al.: Avaliação automática de questões discursivas usando lsa. Universidade Federal do Pará (2016)
17. Sukkarieh, J.Z.: Using a MaxEnt classifier for the automatic content scoring of free-text responses. In: American Institute of Physics Conference Proceedings, pp. 41–48 (2010). https://doi.org/10.1063/1.3573647
18. Valenti, S., Neri, F., Cucchiarelli, A.: An overview of current research on automated essay grading. J. Inf. Technol. Educ. **2**, 319–330 (2003)
19. Vijaymeena, M., Kavitha, K.: A survey on similarity measures in text mining. Mach. Learn. Appl.: Int. J. **3**(2), 19–28 (2016)
20. Ziai, R., Ott, N., Meurers, D.: Short answer assessment: establishing links between research strands. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 190–200 (2012)