International Journal of Artificial Intelligence in Education 25 (2015) 60 – 117
IOS Press

The Eras and Trends of Automatic Short Answer Grading

Steven Burrows[1], **Iryna Gurevych**[1], **and Benno Stein**[2]

[1] *Ubiquitous Knowledge Processing Lab*
German Institute for International Educational Research, 60486 Frankfurt, Germany, and
Technical University of Darmstadt, 64289 Darmstadt, Germany
burrows@dipf.de, gurevych@dipf.de

[2] *Web Technology and Information Systems*
Bauhaus-Universität Weimar, 99421 Weimar, Germany
benno.stein@uni-weimar.de

Abstract. Automatic short answer grading (ASAG) is the task of assessing short natural language responses to objective questions using computational methods. The active research in this field has increased enormously of late with over 80 papers fitting a definition of ASAG. However, the past efforts have generally been ad-hoc and non-comparable until recently, hence the need for a unified view of the whole field. The goal of this paper is to address this aim with a comprehensive review of ASAG research and systems according to history and components. Our historical analysis identifies 35 ASAG systems within 5 temporal themes that mark advancement in methodology or evaluation. In contrast, our component analysis reviews 6 common dimensions from preprocessing to effectiveness. A key conclusion is that an era of evaluation is the newest trend in ASAG research, which is paving the way for the consolidation of the field.

Keywords. Short Answer, Automatic Grading, Natural Language Processing.

INTRODUCTION

The assessment of learning outcomes with tests and examinations can be facilitated by many question types and grading methods. The specific question types may be designed as anything from simple multiple-choice questions, to questions requiring natural language responses such as short answers or essays. The grading method may be either manual grading by hand or automatic grading by computational methods. In this paper we focus on the *short answer* question type and the *automatic grading* method. We refer to this field as *automatic short answer grading*, or ASAG.

The difference between say multiple choice and short answer questions is easy to comprehend, but the difference between other question types such as short answers and essays can become blurred. Therefore we say that a *short answer* question is one that can be considered as meeting at least five specific criteria. First, the question must require a response that recalls external knowledge instead of requiring the answer to be recognized from within the question. Second, the question must require a response given in natural language. Third, the answer length should be roughly between one phrase and

DOI 10.1007/s40593-014-0026-8

one paragraph. Fourth, the assessment of the responses should focus on the content instead of writing style. Fifth, the level of openness in open-ended versus close-ended responses should be restricted with an objective question design.

Concerning grading methods, some questions are more difficult to grade manually than others. Indeed much variation is present when technology is applied for *automatic grading*. A multiple-choice question can be considered easy to grade with computational methods since there is only a single correct response to each question. In contrast, grading natural language responses to short answer questions can be considered much more difficult, as an understanding of the natural language is required.

Research in grading natural language responses with computational methods has a history dating back to the early work of Page (1966). Since then, automatic grading of natural language responses has become a large field. In addition, the techniques have branched depending on the question type, such as short answers versus essays. This is why we choose to focus this article solely on *automatic short answer grading* (ASAG).

There are numerous benefits to be obtained from automatic grading in general, automatic grading of natural language responses, and indeed ASAG. These are themed around summative assessment (for providing grades), formative assessment (for providing feedback), and effectiveness. Concerning summative assessment, the demands of large class sizes and assessment practices (Burrows and D'Souza, 2005) require efficient and cost-effective solutions. In addition, humans make mistakes when grading, and consistency is needed when inter-rater agreement is imperfect that may result from fatigue, bias, or ordering effects (Haley *et al.*, 2007). Another benefit is that the idea of automatic grading in itself may promote the formalization of assessment criteria when not performed otherwise (Williamson *et al.*, 2012). One must also consider the immediacy that automatic grading systems can provide, where test takers would otherwise need to wait for the human marker to complete the grading (Hirschman *et al.*, 2000). Concerning formative assessment, automatic grading is of interest in broader applications such as e-learning and intelligent tutoring systems. Finally concerning effectiveness, automatic grading is becoming very competitive with human grading for both ASAG (Butcher and Jordan, 2010) and AEG (automatic essay grading) (Shermis *et al.*, 2008).

The technology of interest is still subject to open research issues. The ongoing question concerns the quality of the scores (Williamson *et al.*, 2012) and faith in the process. Indeed, some of the aforementioned advantages do not come without problems. For example, the work needed to create an automated solution often requires much development time, the consistency benefit can be a liability for poorer parts of a model when the poor parts make consistent errors, and care must be taken that patterns in system behavior are not gamed during assessment with unnatural language (Williamson *et al.*, 2012).

When considering ASAG, one must not only consider the algorithms and technology, but also the data sets and evaluation techniques that are used to measure effectiveness. All of these components can be considered a "pipeline" where each artifact or process feeds the next. The notion of a pipeline is well supported by several fields of natural language processing research including relation extraction and template filling (Wachsmuth *et al.*, 2013) and efficient information extraction (Wachsmuth *et al.*, 2011).

The general form of an ASAG system development pipeline is given in Figure 1. This pipeline has 11 components comprising 6 artifacts and 5 processes, which we now summarize. First, test or exam settings (1) with appropriate materials must be identified. Then one or more data sets are created (2) by gathering the questions, teacher answers, and student answers together. The data sets (3) are stored on disk in a flat file, XML, or similar format. Natural language processing (NLP) techniques (4/5) are applied to generate post-processed text and statistics comprising of normalized word forms, annotations, numerical measurements, and similar. Some amount of the data or domain knowledge is used for model building (6) based on a grading method using machine learning, concept mapping, corpus-based meth-

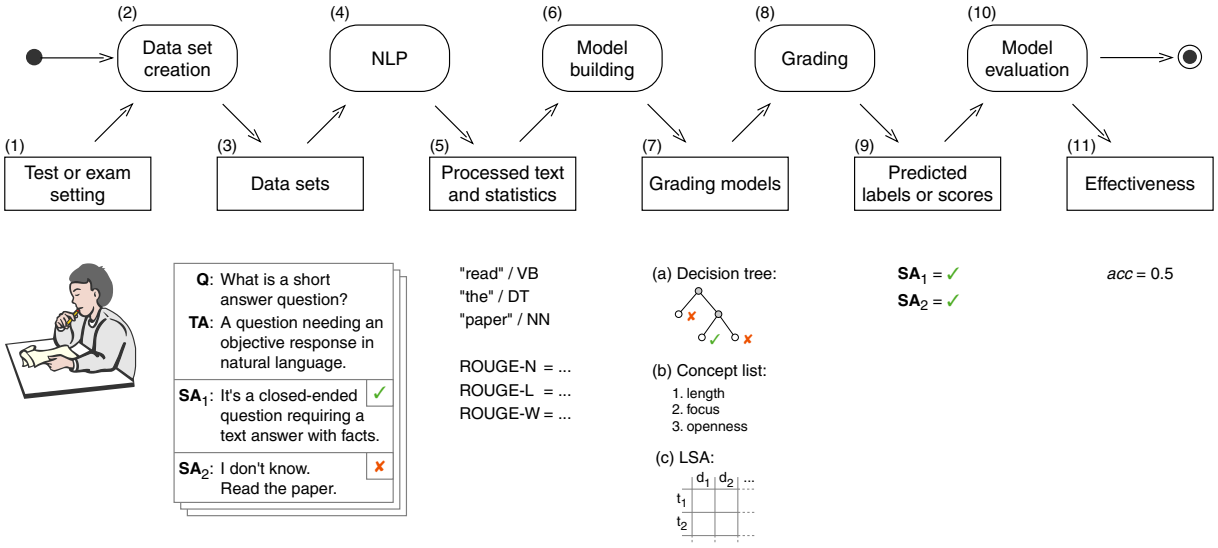

Figure 1. An ASAG system development pipeline represented by 6 artifacts (rectangles) and 5 processes (ovals).

ods, or information extraction techniques. The remainder of the data is then automatically graded (8) to produce a series of predictions (9) based on assigned labels or scores. These predictions are considered during model evaluation (10) where the outcome is the calculation of one or more measurements of effectiveness (11).

In putting the discussion together, there is much to say about the definitions, history, and components in ASAG. The field of ASAG is large and unique, but no review that is unified, comprehensive, and timely is available. Therefore the goal of this survey article is to address this shortcoming along the dimensions of definitions, history, and components. We do so with the corresponding three contributions:

- We review and define many common question types that can be automatically graded, paying particular attention to short answer questions. This contribution defines how short answer questions fit into a bigger picture described by depth of learning, broad question categories, and specific question types.
- We review 35 ASAG systems as a historical analysis. The organization comprises our 5 "eras" of ASAG, comprising 4 methodology eras for concept mapping, information extraction, corpus-based methods, and machine learning, plus a fifth era for initiative in evaluation. This contribution demonstrates the longitudinal trends in ASAG.
- We provide a review of the components of all systems across 6 common dimensions. In reference to the numbering of Figure 1, these are data sets (3), natural language processing (4), model building (6), grading models (7), model evaluation (10), and effectiveness (11). This contribution illustrates the trends across all of these dimensions including the recent and meaningful effectiveness comparisons that can be made.

More broadly, there should also be interest in this article for related communities that work with semantic textual similarity and notions of paraphrasing. Examples are the work by Bär *et al.* (2011, 2012a, 2013), Burrows *et al.* (2013) and Potthast (2011) on text similarity and paraphrasing, and evaluation competition work by Bär *et al.* (2012b) on computing text similarity. This body of work complemented a competitive submission (Zesch *et al.*, 2013) in the SemEval '13 Task 7 competition for

ASAG (Dzikovska *et al.*, 2013). Another example is that research in ASAG has also been cast as a paraphrase recognition problem (Leacock and Chodorow, 2003). Therefore, the comparison of teacher and student answers in ASAG could be supported by the semantic textual similarity and paraphrasing communities. In addition to semantic textual similarity and paraphrasing, the field of intelligent tutoring systems can also be considered as related as a more interactive form of ASAG systems. Example intelligent tutoring systems are AutoTutor (Graesser *et al.*, 2005), CIRCSIM-Tutor (Evens *et al.*, 2001), Geometry Explanation Tutor (Aleven *et al.*, 2004), and Why2-Atlas (VanLehn *et al.*, 2002).

The next three sections of this article address the contributions listed above for definitions (p. 4), history (p. 7), and components (p. 18) respectively. Lessons learned are given in the final section (p. 33).

AN OVERVIEW OF AUTOMATIC ASSESSMENT

The literature on ASAG is vast, and there have been many publications in the last decade in particular. We find it necessary to precisely define the type of question we are dealing with in order to proceed. Therefore the purpose of this section is to show how short answer questions can be distinguished from other types of questions in automated assessment.

The Educational Testing Service (ETS)[1] is one of the largest players in the field of automatic assessment. Their website contains a typology of their research in automated scoring and natural language processing[2] including writing content (i.e.: short answers), writing quality (i.e.: essays), mathematics, and speech. Further typologies include those of Bejar (2011) and Zenisky and Sireci (2002), providing much additional detail. In contrast, György and Vajda (2007) offer a hierarchy providing a grouping for active and passive questions, and a sub-grouping of active questions that require answers as numbers or text. In summarizing these existing bodies of work, Figure 2 provides the highlights under three "swim lanes": "depth of learning", "question category", and "question type". The figure is not intended to be exhaustive, but the goal is to simply show sufficient and common examples to differentiate ASAG questions from others. We now review the three swim lanes emphasizing the parts relevant to ASAG, which are highlighted in Figure 2.

Depth of Learning

The first level of organization concerns the depth of learning between "recognition" and "recall" questions, which is terminology supported by the literature (Gay, 1980; Jordan, 2009a). Alternatively, we may say closed versus open questions (Gonzalez-Barbone and Llamas-Nistal, 2008). Yet another distinction is passive versus active questions as mentioned above (György and Vajda, 2007). For recognition questions, the respondents usually only need to organize or identify some key information. In contrast, recall questions have the benefit of requiring the respondents to come up with original answers expressed in their own way. With respect to pedagogy, recall methods represent a higher level in Bloom's taxonomy of learning objectives (Krathwohl, 2002). In comparison, recognition questions can be considered as representing low-level factual knowledge (Martinez and Bennett, 1992). More practically, recall questions are less susceptible to test taking strategies (Hirschman *et al.*, 2000) and guessing (Conole and Warburton, 2005) compared with recognition questions.

For recognition questions, automatic grading is a solved problem, as the answer is always among a set of options. This is emphasized with the "Recognition" part of Figure 2. Therefore, the momentum in

[1] http://www.ets.org

[2] http://www.ets.org/research/topics/as_nlp

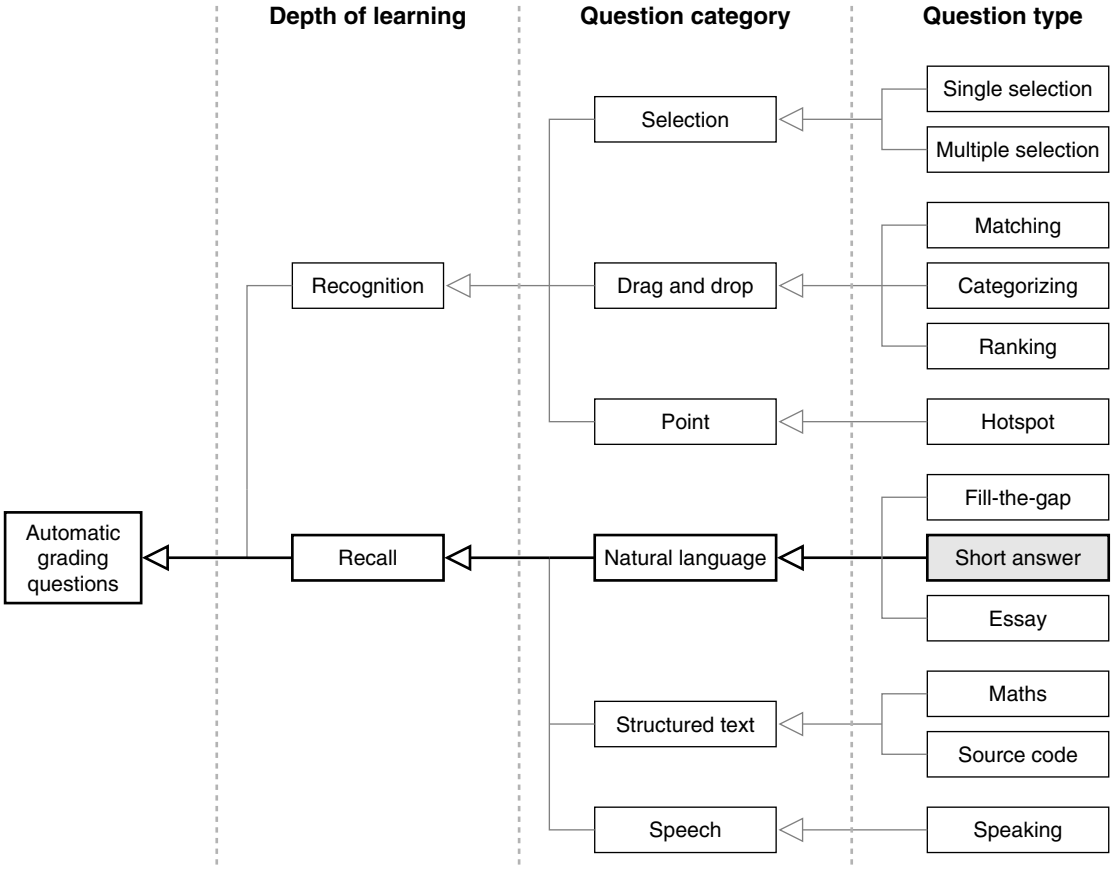

Figure 2. A hierarchical view of common question types where automatic grading methods can be applied.

automatic grading is for recall questions due to this reason and the others above. Short answer questions fall within this category.

Question Category

The second level of organization has several broad groupings for specific question types, from which we only consider the bottom half (recall) as relevant to this article. The first of these is the appropriate label for "short answers"; that of "natural language". This explains the absence of notation-based math questions in our literature review: maths notation can be considered structured text, not natural language. As an additional example of structured text, the specialized study of source code as structured text has received attention in areas such as plagiarism detection (Burrows *et al.*, 2007) and authorship attribution (Burrows *et al.*, 2014). Finally, that leaves us with speech: some overlap can be considered with natural language after transcription (Wang *et al.*, 2013), however notions of pronunciation and enunciation are of interest too. We choose to omit types of graphical questions (Csink *et al.*, 2003) from Figure 2, as our interest in recall questions only extends to those that can be modeled in a text-based format.

Question Type

For the third level of organization, we list several specific question types. For the natural language question types, we need to separate short answer questions from fill-the-gap and essay questions. The

Table 1
Properties that distinguish types of natural language questions.

Property	Question type		
	Fill-the-gap	Short answer	Essay
Length	One word to a few words.	One phrase to one paragraph.	Two paragraphs to several pages.
Focus	Words.	Content.	Style.
Openness	Fixed.	Closed.	Open.

difference between these types can be fuzzy, particularly for short answers versus essays when other terminology is used such as "free-text answer" (Sargeant *et al.*, 2004) and "constructed response" (Bennett, 2011). Our three key dimensions to distinguish natural language question types are *length*, *focus*, and *openness*. Table 1 summarizes these dimensions, which we now discuss.

The first key dimension to separate the natural language question types is answer *length*. For both short answers and essays, the answers must be sufficiently long such that a wide variety of unique answers and wordings can be expressed. This is not true for fill-the-gap questions, since the solutions comprise no more than a few words. For short answers, the range in length should be from about one phrase (several words) up to one paragraph to be consistent with the existing literature. The examples we find state that the length of short answers are "phrases to three to four sentences" (Siddiqi *et al.*, 2010) or "a few words to approximately 100 words" (Sukkarieh and Stoyanchev, 2009). This leaves essays as defined as two or more paragraphs up to several pages.

The second key dimension is the *focus* of the grading technique. Here, ASAG systems tend to focus more on content, whilst automatic essay grading (AEG) systems (Shermis and Burstein, 2003, 2013) tend to focus more on style (Gütl, 2007; Pérez-Marín, 2004). This observation is supported by two ETS systems as examples of ASAG and AEG systems called c-rater (Leacock and Chodorow, 2003) and e-rater (Attali and Burstein, 2006) respectively. Specifically, Attali *et al.* (2008, pp. 1–2) state that the goal of c-rater is to "map student responses onto the experts' models in order to determine their correctness or adequacy" whilst the e-rater system is "based on a generic model of writing that is applied to any prompt that belongs to an assessment". Put another way, Jordan and Mitchell (2009, p. 372) state that AEG systems "focus on metrics that broadly correlate with writing style, augmented with aggregate measures of vocabulary usage" whilst ASAG systems are "concerned with marking for content above all else". Yet another comparison is content versus expression and fluency (Williamson *et al.*, 2012). An exception can be made for systems that claim to do both essay and short answer grading (Pearson Education, 2010). For fill-the-gap questions, we simply say that the focus is on specific words.

The third key dimension concerns the *openness* of the question. Specifically, ASAG systems require answers to objective or close-ended questions. In contrast, AEG systems require answers to subjective or open-ended questions (Leacock and Chodorow, 2003; Siddiqi and Harrison, 2008b; Wood *et al.*, 2006). Put another way, the difference is facts and statements versus examples and opinions (Leacock and Chodorow, 2003). For fill-the-gap questions, we say that the responses are fixed since there is essentially no novelty to be expressed.

Exception: Reading Comprehension

Questions on reading comprehension do not fully comply with our pathway in Figure 2. Reading comprehension fits our definition of "natural language" and "short answer" for the second and third swim lanes, but not "recall" for "depth of learning". In reading comprehension, the student is given sample text from which to formulate an answer to a question. For example, a student might be asked why a character

from a story given in a short passage performed a certain action. In this case, the student must recognize the answer from the passage given, and does not need to recall existing knowledge.

Despite the definition mismatch, we have included a few key papers with reading comprehension due to their relevance to ASAG. Specifically, four systems (CAM, CoMiC-DE, CoMiC-EN, CoSeC-DE, introduced in the next section) are linked through the common authorship of Detmar Meurers and stem from his group's work that aims to link otherwise quite separate strands of ASAG research together (Ziai *et al.*, 2012). Another inclusion is the paper by Horbach *et al.* (2013) that is specifically advertised as "short answer scoring". The final paper by Madnani *et al.* (2013) has components common to ASAG research including many features, a well-defined scoring scale, and a familiar evaluation style.

HISTORICAL ANALYSIS

To the best of our knowledge, two relevant survey papers are available, namely the work by Valenti *et al.* (2003) and Pérez-Marín *et al.* (2009). However, these have ASAG and AEG systems mixed together. Aside from this, Ziai *et al.* (2012) devote over 4 pages of their workshop paper to reviewing 12 ASAG systems, but this review is not complete as we demonstrate. In comparison, our historical review is intended to be comprehensive and comprises 35 identified ASAG systems and 2 competitions.

We observe the existing ASAG systems as falling into broad themes and time periods, from which we model the organization of our literature review. Here, we state that each category is an "era" in the field of ASAG, to emphasize the historical organization. Therefore we define an "era" as a thematically consistent set of activities with a particular time period. The era time periods may overlap with others, but we otherwise keep the activities as disjoint as possible. The ASAG systems themselves may sometimes overlap with multiple eras, in which case we refer to the dominant era.

For each era, we first define and explain the key ideas as an introduction. The five eras are concept mapping, information extraction, corpus-based methods, machine learning, and evaluation, as listed in Figure 3. Based on this list, we point out that the first four eras are method-based, but the fifth is evaluation-based. Given that there is a current and big movement towards reproducibility, standardized corpora, and permanent evaluation, an "era of evaluation" is important to emphasize this movement.

Following the introductions for each era, each corresponding system is then given its own section. The naming convention for the systems considers both named and unnamed systems from the literature. For the named systems, we state the name given in the publication (such as "c-rater"). For the unnamed systems, we give the name of the first author and starting year (such as "Mohler '09").

Each section heading for the reviewed systems is also immediately followed by an ID number in parentheses. The ID number refers to the numbering scheme in Table 2 and allows for simple cross-referencing within this article between the system descriptions in the historical analysis and numerous tables in the component analysis.

Finally, when reviewing each system, we reference the main publication at the start, and reference some secondary publications in the text as necessary. The full set of main and secondary references is also given in Table 2. The review for each system then generally proceeds with a description of the key ideas and methods.

It is clear that our historical analysis as described above creates a lengthy historical review. An alternative structure would be to review each era as a whole. However, we find that our organization is very helpful for the component analysis that follows. Here, the historical analysis presents the systems with a temporal continuity that shows how the field has developed over time. This then allows the component analysis to provide a view that cuts across time and reveal the underlying structure of the

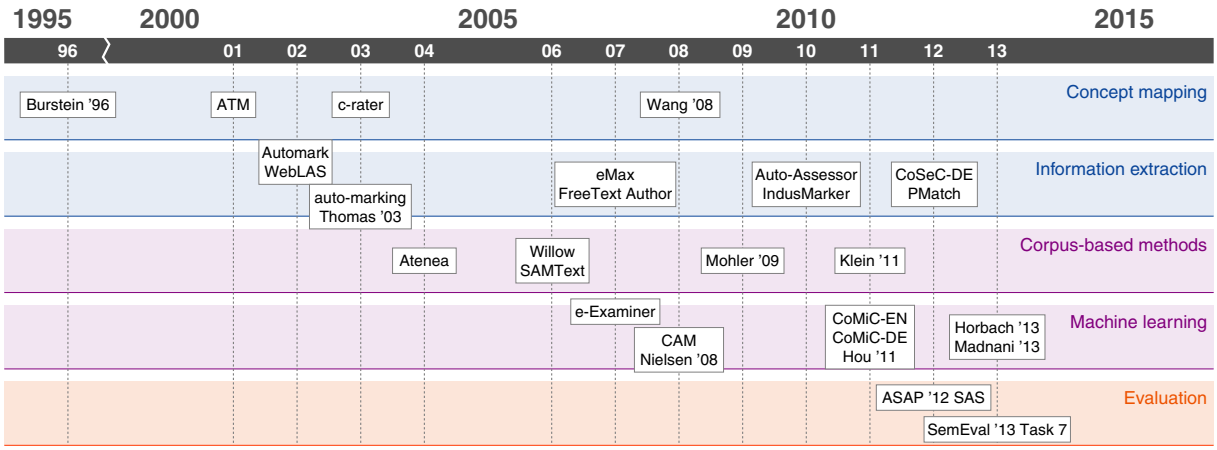

Figure 3. Historical organization for the literature review. The five eras are (top-down) Concept mapping, Information extraction, Corpus-based methods, Machine learning, and Evaluation. Each system is recorded against the year published ignoring month-offsets. The systems developed for the evaluation competitions are grouped together and represented by the competition name.

various systems. Therefore the following sections aim to develop an understanding of the systems for the component analysis that follows.

All eras and systems reviewed are organized by Figure 3. We now begin the historical analysis according to this organization.

Era of Concept Mapping

The idea of concept mapping is to consider the student answers as made up of several concepts, and to detect the presence or absence of each concept when grading. Suitable questions must therefore facilitate this idea, such as a question that asks for a solution to a problem plus a justification, or a question that asks for multiple explanations to the same problem. To cite an example from the literature, Burstein *et al.* (1996) have a question where students are expected to provide multiple reasons for decreases in deaths in the police force over time. Three sample concepts by Burstein *et al.* (1996) are: (1) "Better economic circumstances mean less crime", (2) "Advanced medical technology has made it possible to save more lives", and (3) "Crooks now have a decreased ability to purchase guns".

Note that the concept mapping is expressed at the *sentence level*. It is possible to delve into a finer level of detail concerning individual fragments (typically word pairs and triples), but this problem is typically known as *facet mapping* instead. For example, Nielsen *et al.* (2008a) conduct "facet-based classification" and have a question where students are asked about the sounds produced by string instruments, and the reference answer is: "A long string produces a low pitch". Again referring to Nielsen *et al.* (2008a), this sentence-level answer can be broken down into four facets: (1) string/long: "There is a long string", (2) produces/string: "The string is producing something", (3) produces/pitch: "A pitch is being produced", and (4) pitch/low: "The pitch is low". Based on this process, essentially any concept can be broken down into facets.

Consider also that there is some relation of concept mapping (and facet mapping) to textual entailment (and partial textual entailment) (Levy *et al.*, 2013). In textual entailment research, the nomenclature does not describe answers as correct or incorrect, preferring to state that concepts (or facets) have been either expressed or unaddressed (Dzikovska *et al.*, 2013). This link is demonstrated in the c-rater liter-

Table 2
List of main and secondary references for all 35 systems and the 2 competitions.

ID	System	Reference	Page	Secondary References
1	Atenea	Alfonseca and Pérez (2004)	12	Alfonseca *et al.* (2005); Pérez and Alfonseca (2005); Pérez *et al.* (2004a,b, 2005a,b,c); Pérez-Marín (2004).
2	ATM	Callear *et al.* (2001)	9	
3	Auto-Assessor	Cutrone *et al.* (2011)	11	Formerly "Automarking" (Cutrone and Chang, 2010).
4	auto-marking	Sukkarieh *et al.* (2003)	11	Pulman and Sukkarieh (2005); Sukkarieh *et al.* (2004); Sukkarieh and Pulman (2005).
5	AutoMark	Mitchell *et al.* (2002)	11	Mitchell *et al.* (2003a,b).
6	Burstein '96	Burstein *et al.* (1996)	9	
7	c-rater	Leacock and Chodorow (2003)	10	Attali *et al.* (2008); Sukkarieh (2010); Sukkarieh and Blackmore (2009); Sukkarieh and Bolge (2008, 2010); Sukkarieh and Kamal (2009); Sukkarieh and Stoyanchev (2009).
8	CAM	Bailey and Meurers (2008)	13	Bailey (2008).
9	CoMiC-DE	Meurers *et al.* (2011a)	14	Meurers *et al.* (2010); Ott *et al.* (2012).
10	CoMiC-EN	Meurers *et al.* (2011b)	14	Ziai *et al.* (2012).
11	Conort '12	Conort (2012)	16	
12	CoSeC-DE	Hahn and Meurers (2012)	12	
13	Dzikovska '12	Dzikovska *et al.* (2012)	17	
14	e-Examiner	Gütl (2007)	13	Gütl (2008). Latterly "Electronic Assessor" (Moser, 2009).
15	eMax	Sima *et al.* (2009)	11	György and Vajda (2007); Sima *et al.* (2007). Formerly "EVITA" (Csink *et al.*, 2003).
16	ETS	Heilman and Madnani (2013)	17	
17	FreeText Author	Jordan and Mitchell (2009)	11	Intelligent Assessment Technologies (2009); Jordan (2007, 2008, 2009a,b); Jordan *et al.* (2007); Swithenby and Jordan (2008); Willis (2010).
18	Horbach '13	Horbach *et al.* (2013)	14	
19	Hou '11	Hou and Tsao (2011)	14	Hou *et al.* (2010, 2011, 2012).
20	IndusMarker	Siddiqi and Harrison (2008b)	12	Siddiqi and Harrison (2008a).
21	Klein '11	Klein *et al.* (2011)	13	
22	Levy '13	Levy *et al.* (2013)	17	
23	Madnani '13	Madnani *et al.* (2013)	15	
24	Mohler '09	Mohler and Mihalcea (2009)	13	Mohler *et al.* (2011).
25	Nielsen '08	Nielsen *et al.* (2008b)	14	Nielsen *et al.* (2009).
26	PMatch	Jordan (2012a)	12	Jordan (2012b).
27	SAMText	Bukai *et al.* (2006)	13	
28	SoftCardinality	Jimenez *et al.* (2013)	17	
29	Tandella '12	Tandalla (2012)	15	
30	Thomas '03	Thomas (2003)	11	
31	UKP-BIU	Zesch *et al.* (2013)	17	
32	Wang '08	Wang *et al.* (2008)	10	
33	WebLAS	Bachman *et al.* (2002)	11	
34	Willow	Pérez-Marín and Pascual-Nieto (2011)	12	Pascual-Nieto *et al.* (2008, 2011); Pérez-Marín (2007); Pérez-Marín *et al.* (2006a,b,c,d, 2007).
35	Zbontar '12	Zbontar (2012)	16	
36	ASAP '12 SAS	Hewlett Foundation (2012)	15	
37	SemEval '13 Task 7	Dzikovska *et al.* (2013)	16	

ature, where Leacock and Chodorow (2003) specify that "the scoring engine must be able to recognize when a concept is *expressed* [emphasis added] and when it is not". The link is further demonstrated by Levy *et al.* (2013) whom translate their entailment expertise to ASAG grading.

Burstein '96 (6) Burstein *et al.* (1996) consider hypothesis-style questions where multiple explanations must be given for a given hypothesis, each of which may or may not match one of the teacher answers. Each answer can be considered a separate concept. The applied technique is the Lexical Conceptual Structure representation (Dorr *et al.*, 1995) whereby a concept-based lexicon and a concept grammar must be developed from a training set before grading the hypotheses in the student answers.

ATM (2) ATM (Automatic Text Marker) (Callear *et al.*, 2001) breaks down teacher and student answers into lists of minimal concepts comprising no more than a few words each, and counts the number of concepts in common to provide an assessment score. Each concept is essentially the smallest possible unit in an answer that can be assigned a weight for the purposes of grading. The weights are summed to produce the overall score.

c-rater (7) The Concept Rater (c-rater) (Leacock and Chodorow, 2003) aims at matching as many sentence-level concepts as possible between teacher and student answers. The matching is based on a set of rules and a canonical representation of the texts using syntactic variation, anaphora, morphological variation, synonyms, and spelling correction. Specifically, the teacher answers are entered as a separate sentence for each concept. This simplifies the assessment since only one concept is considered at a time when grading. This technique avoids the need for an indirect solution, such as dividing the question into multiple parts (Jordan, 2009b) and it is argued that this can lead to higher accuracy (Sukkarieh and Blackmore, 2009). Furthermore, the natural language input format is advantageous compared with other systems that require expertise and use of a markup language (Sukkarieh and Stoyanchev, 2009).

An important development that follows is the use of automated concept-based scoring for model building, to replace manual holistic scoring, that is described as taking 12 hours of human time per question (Sukkarieh and Stoyanchev, 2009). The manual holistic scoring required a user to manually express equivalent sentences and the lexicon as the basis of a model. The automated method instead only requires manual concept-based scoring, but then the lexicon is automatically generated. The automatic generation of the lexicon is performed by creating a stratified sampling of the sentences, and selecting the lexicon based on one of several selection strategies that are compared empirically. Results indicate that the unweighted kappa values for the automatically built models are "comparable" to the manually built models for 11/12 scenarios. The remaining scenario had seven concepts, which was the highest number of concepts among all scenarios, so these results suggest that further experimentation may be warranted for questions with many concepts.

Later, the c-rater work regards the grading problem as textual entailment (Sukkarieh and Bolge, 2008). Here, the "GoldMap" concept mapping algorithm (Sukkarieh and Stoyanchev, 2009) uses calculations of maximum entropy (Sukkarieh, 2010) between teacher and student answers for grading.

Wang '08 (32) Wang *et al.* (2008) compare three methods for grading earth science questions in secondary education, which are based on concept mapping, machine learning, or both. The first concept mapping method is cosine on *tf.idf* (term frequency multiplied by inverse document frequency) vectors of bag-of-words features. The second concept mapping method is a support vector machine (SVM) with bag-of-words features. Note that the second concept mapping method is remarkable as it is implemented with machine learning, and can be considered a blend of concept mapping and machine learning. The third and final method is a pure machine learning method employing SVM regression with unigrams, bigrams, and part-of-speech bigrams. Unlike the first two methods, the pure machine learning method grades holistically, and all concepts are considered together as a single answer.

Era of Information Extraction

In the context of this article, information extraction (Cowie and Wilks, 2000) is concerned with fact finding in student answers. Given that short answers are usually expected to include specific ideas, these can be searched for and modeled by templates. Simply, information extraction methods in this article can be considered as a series of pattern matching operations such as regular expressions or parse trees.

More generally, information extraction techniques can extract structured data from unstructured sources, such as free text, and represent the structured data as tuples for use in numerous applications.

AutoMark (5) AutoMark (Mitchell *et al.*, 2002) performs pattern matching as a form of information extraction on parse tree representations of teacher and student answers for grading. Two approaches are described, namely the "blind" and "moderated" approaches. The blind approach represents the best definition of ASAG in that it is fully automated. In contrast, the moderated approach includes a human-driven step that allows the model to be revised after grading has been performed. Therefore, the overall approach allows for optional improvement when human resources are available.

WebLAS (33) WebLAS (Web-based Language Assessment System) (Bachman *et al.*, 2002) identifies important segments of the teacher answers through parsed representations, and asks the teacher to confirm each and assign weights. The teacher is also prompted to accept or reject semantically similar alternatives. Regular expression matching is performed to detect the presence or absence of each segment in the student answers. Partial grading is possible as each segment is accounted for separately.

auto-marking (4) Auto-marking (Sukkarieh *et al.*, 2003) uses hand-crafted patterns that are fitted to a training set for model building. Two patterns are formed for each question as each question is worth two marks. Empirical evaluation shows that the approach is more effective than a k-nearest neighbor baseline with bag-of-words features weighted by *tf.idf*. Sukkarieh *et al.* (2004) also explore the idea of forming the patterns using bootstrapping. However the amount of data is not reported, which makes it difficult to compare this approach to the hand-crafted approach.

Thomas '03 (30) Thomas (2003) addresses ASAG as a boolean pattern matching problem with thesauri support. That is, the required phrases are defined as boolean-AND expressions, and acceptable alternatives are added as boolean-OR expressions. Awarding credit to correct solutions therefore requires a perfect match.

eMax (15) eMax (Sima *et al.*, 2009) requires the teacher to mark-up required semantic elements[3] of the teacher answers, accept or reject synonyms to these elements as prompted, and assign weights to each element for calculating the final score (Sima *et al.*, 2007). The approach to grading is a combinatoric one, where all possible formulations are considered when pattern matching is performed. The assigned scores are also given a confidence rating, so that difficult cases can be forwarded for manual review.

FreeText Author (17) FreeText Author (Jordan and Mitchell, 2009) (formerly AutoMark as above) provides a graphical user interface for teacher answer input and student answer grading. The teacher answers are composed as syntactic-semantic templates for the student answers to be matched against. These templates are automatically generated from the natural language input of teacher answers, therefore no user expertise in natural language processing is required. Through the interface, the teacher can specify mandatory keywords from the teacher answers and select from synonyms provided by thesauri support. Both acceptable and unacceptable answers can be defined, and student answers are awarded credit according to template matches.

Auto-Assessor (3) Auto-Assessor (Cutrone *et al.*, 2011) focuses on grading canonicalized single-sentence student answers based on bag-of-words coordinate matching and synonyms with WordNet (Pedersen *et al.*, 2004). Coordinate matching in ASAG simply refers to matching individual terms between teacher and student answers. In Auto-Assessor, each word that matches exactly is given one point, related words from WordNet are given partial credit, and the rest are given no credit.

[3] A "semantic element" (Sima *et al.*, 2009) refers to a fragment of the answer and other meanings that can be extrapolated.

IndusMarker (20) IndusMarker (Siddiqi and Harrison, 2008a) is used to perform word- and phrase-level pattern matching to grade student answers. This is referred to as "structure matching". The credit-worthy phrases are defined using an XML markup language called the Question Answer Markup Language. Using the "structure editor", the text and number of points can be input for each phrase.

CoSeC-DE (12) CoSeC-DE (Comparing Semantics in Context) (Hahn and Meurers, 2012) uses the Lexical Resource Semantics (LRS) method (Richter and Sailer, 2003) to create abstract representations of texts. The idea is exemplified by comparing the following three sentences:

(1) "The hare beats the tortoise." (2) "The tortoise beats the hare." (3) "The tortoise was beaten by the hare."

Here, (1) and (2) are equivalent according to a bag-of-words model, but (1) and (3) are equivalent under a LRS model. Specifically, LRS representations of teacher and student answers are modeled as graphs, and a threshold-based alignment is performed to detect equivalent meanings.

PMatch (26) PMatch (Jordan, 2012a) is considered a successor to FreeText Author (above) at the Open University. This system is capable of grading very short answers of up to one sentence in length. The system performs word-level pattern matching where all required words, word stems, and allowed synonyms for correct answers are matched by regular expressions against the teacher answers.

Era of Corpus-Based Methods

Corpus-based methods exploit statistical properties of large document corpora. Although such methods are often used for applications with longer texts (Bukai *et al.*, 2006), these methods can also be useful when interpreting synonyms in short answers, as using only the original teacher answer vocabulary will limit the correct answers that can be identified. A typical technique to increase the vocabulary is to use bilingual parallel corpora to analyze the frequency of term pairs being resolved to the same common second-language translation. Then synonyms with particularly common translations can be incorporated into the teacher answers.

Atenea (1) Atenea (Alfonseca and Pérez, 2004) initially uses the BLEU (BiLingual Evaluation Understudy) metric (Papineni *et al.*, 2002) for scoring. This metric is based on n-gram overlap and normalized sample length. Then the [0,1] interval value is scaled to the appropriate point range. Importantly, Alfonseca and Pérez (2004) argue that BLEU should be both precision and recall accommodating, as the original BLEU only considers precision. The extension is referred to as ERB (Evaluating Responses with Bleu) (Pérez-Marín, 2004). Atenea is shown to be more effective than coordinate matching and vector space model baselines (Pérez *et al.*, 2004a; Pérez-Marín, 2004).

Later, Latent Semantic Analysis (LSA) (Landauer *et al.*, 1998) is added and a weighted combination of BLEU and LSA scores is taken instead (Pérez *et al.*, 2005a). LSA is a corpus-based approach akin to a vector space model that accommodates for lexical variability. The approach to combine BLEU and LSA offers a consistent improvement compared to the previous work (Pérez *et al.*, 2005a,c). Since a weighted combination of BLEU and LSA is taken instead of the individual features as part of a machine learning solution, we consider Atenea as a corpus-based method.

Willow (34) Willow (Pérez-Marín and Pascual-Nieto, 2011) is the successor to Atenea (described above). However, the research on ASAG is only incremental, as much of the new work takes on a pedagogic flavor instead. For example, the current performance of the students is consulted to select the difficulty of new questions (Pérez-Marín *et al.*, 2006c), topic suggestions are given to the students for continued study (Pascual-Nieto *et al.*, 2011), and self-assessment functionality is introduced (Pascual-Nieto *et al.*, 2008).

SAMText (27) SAMText (Short Answer Measurement of TEXT) (Bukai *et al.*, 2006) applies a variant of LSA based on an inverted index data structure, which is seeded by content from a web crawl using topically-relevant documents. In contrast, LSA normally uses a matrix data structure based on large corpora for modeling semantic relatedness. Bukai *et al.* (2006) argue that the inverted index and crawling idea is more suitable for short answers compared with long answers because web crawls can be tailored to each topic instead of trying to model all language at once.

Mohler '09 (24) Mohler and Mihalcea (2009) develop several systems to investigate unsupervised grading methods by individually comparing eight knowledge-based and two corpus-based semantic similarity measures. The knowledge-based measures are Hirst and St-Onge (1998), Jiang and Conrath (1997), Leacock and Chodorow (1998), Lesk (1986), Lin (1998), Resnik (1995), shortest path (Mohler and Mihalcea, 2009), and Wu and Palmer (1994). The two corpus-based measures are Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2006) and LSA. Apart from comparing these measures, Mohler and Mihalcea (2009) also consider incorporating the best student answers with the teacher answer to expand the teacher answer vocabulary, which they find to be effective.

Klein '11 (21) Klein *et al.* (2011) implement an LSA system where the key idea is to use the student answers as the LSA model instead of general texts from another source. Some of the texts are then marked manually and this forms the model for automatically grading the remainder. The key problems are to select specific texts to mark manually and determine the overall quantity. Concerning the quantity, the process is repeated until a specific correlation threshold is achieved. Concerning the selection, three approaches are considered: random selection, clustering, or selecting the least similar text to those marked already. This third approach is shown to be the most effective.

The benefit of the Klein '11 approach is that the set of submissions to be marked manually is chosen automatically and is minimized. The disadvantage of the approach is apparent in the evaluation, whereby the desired effectiveness level is only achieved after manually marking the majority of the students' answers. This amounts to 83% when all presented scenarios are summed. Another problem is that the method is parameter-dependent, in that the semantic space dimensionality and the similarity threshold parameter must be determined.

Era of Machine Learning

Machine learning systems typically utilize some number of measurements extracted from natural language processing techniques and similar, which are then combined into a single grade or score using a classification or regression model. This can be supported by a machine learning toolkit such as Weka (Hall *et al.*, 2009). Features involving bag-of-words and n-grams are typical of this category, as are decision trees and support vector machines as representative learning algorithms.

e-Examiner (14) e-Examiner (Gütl, 2007) uses ROUGE metrics (Lin, 2004) as machine learning features. These are combined as a linear regression. Much of the remainder of this work is focused on system architecture, where the flexible design allows the service to be used in a stand-alone fashion, or as a component in an existing system such as an e-learning platform.

CAM (8) CAM (Content Assessment Module) (Bailey and Meurers, 2008) uses a k-nearest neighbor classifier and features that measure the percentage overlap of content on various linguistic levels between the teacher and student answers. The types of overlap include word unigrams and trigrams, noun-phrase chunks, text similarity thresholds, parts of speech, lemmas, and synonyms. It is also interesting to note the unusual terminology used to describe the two evaluation tasks. First, "semantic error detection"

represents a 2-way test of correctness (i.e.: binary classification). Second, "semantic error diagnosis" represents a 5-way test against an expanded set of class labels for the negative class (i.e.: 5-class classification). Therefore the experiments represent summative and formative grading schemes respectively.

Nielsen '08 (25) Nielsen *et al.* (2008b) evaluate their machine learning system on the SciEntsBank data that later became part of the SemEval '13 Task 7 competition (Dzikovska *et al.*, 2013). The classification task is 5-way based on primary school science questions for grades 3–6, with labels "understood", "contradicted", "self contradicted", "different argument", and "unaddressed" (these are later remapped for SemEval '13 Task 7). In the system, the choice of features includes both lexicalized features (parts of speech, stem matches, and entailment probabilities) and syntactic features (dependency relation type and edit distance). A C4.5 decision tree is used for classification.

CoMiC-EN (10) CoMiC-EN (Meurers *et al.*, 2011b) and CoMiC-DE (next system) come from the Comparing Meaning in Context project (CoMiC).[4] CoMiC-EN is an iteration of CAM and the implementation is similar. The main goal of CoMiC-EN is not to necessarily be more effective than CAM, but to switch to an architecture and toolset with sufficient flexibility for integration in intelligent tutoring systems. The evaluation is also on the CAM data (Bailey and Meurers, 2008), now called CREE (Corpus of Reading comprehension Exercises in English). The 2-way and 5-way evaluation performed with CAM is also the same, now called "binary classification" and "detailed classification".

CoMiC-DE (9) CoMiC-DE (Meurers *et al.*, 2011a) is essentially the German-language counterpart to CoMiC-EN. The preprocessing, feature selection, and classification steps are all the same, with necessary changes in the toolsets for the German language (e.g.: using GermaNet (Hamp and Feldweg, 1997) instead of WordNet). The evaluation corpus is changed from CREE to CREG (Corpus of Reading comprehension Exercises in German) (Meurers *et al.*, 2010).

Hou '11 (19) Hou and Tsao (2011) implement a system used for providing teachers with an indicator of student progress, but there is obvious extension for use as a typical ASAG system. Four classes of features are extracted comprising POS tags, term frequency, *tf.idf*, and entropy, which are combined with an SVM classifier. We say that the experimental setup is suited to providing teachers with a progress indicator because the 10-point marking scale has only been explored coarsely in the experiments. That is, for the 2-way experiment the buckets 0-5 and 6-10 are predicted, and for the 3-way experiment the upper bucket is split as 6-7 and 8-10. So regression may be a good option to extend the work for ASAG in this setting.

Horbach '13 (18) Horbach *et al.* (2013) include the reading texts from reading comprehension questions in their data sets as their key idea. For all other types of ASAG questions, the data comprises three components: (1) the questions, (2) the teacher answers, and (3) the student answers. However, in reading comprehension questions, another component is available: (4) the reading texts. Horbach *et al.* (2013) describe this as helpful because the student answers may only refer to one part of the reading texts. So normally, ASAG systems exploit the relationship between (2) and (3), however in this paper the pairs (2)/(4) and (3)/(4) are also exploited.

Much of the remainder of the work is actually based on CoMiC-DE (above). That is, the CREG data set is re-used, and the methodology is based on global alignment (with sentence alignment features comprising simple agreement, entropy, and alignment error in number of sentences) and the k-nearest neighbor classifier. The new work required additional annotation on the CREG data set to mark the sentence alignments between (2) and (4) to assist with feature extraction. However, the alignments are

[4] `http://purl.org/icall/comic`

also automated as an alternative approach, and these results indicate that the effectiveness is essentially the same or marginally better than CoMiC-DE.

Madnani '13 (23) Madnani *et al.* (2013) implement a system for grading reading comprehension questions about living standards. Each text has three paragraphs, and the student answers specifically require one sentence giving an overall summary and three more sentences giving a summary of each paragraph. The machine learning approach comprises eight features (BLEU, ROUGE, measurements concerning different dimensions of text copying, number of sentences, and counts of commonly used discourse connector words) as input to a logistic regression classifier.

Era of Evaluation

Unlike the preceding four eras that describe methods, the era of evaluation is method-independent. In particular, this means the use of shared corpora, so that advancements in the field can be compared meaningfully. This also refers to competitions and evaluation forums whereby research groups from all around the world compete against one another on a particular problem for money or prestige.

ASAP ASAP (Automated Student Assessment Prize) is an automatic grading competition series organized by the commercial competition hosting company Kaggle.[5] The Kaggle community is made up of client companies and participating data scientists. The client companies pay a fee to host and get support for their particular competition, and the data scientists participate freely and compete to create the most desired solution and possibly win a monetary prize. In return, the client companies benefit from having custom-created solutions created by world-leading data scientists. The three ASAP competitions comprise of AEG from January to April in 2012,[6] ASAG from June to October also in 2012,[7] and symbolic mathematical and logic reasoning for charts and graphs in the future.[8]

ASAP '12 SAS (36) For the ASAG offering of ASAP,[9] the ten questions comprised of varied subject matter at the high school level from arts to science. The participants were given 1,800 student answers for training, which were randomly taken from a pool of 3,000. Each student answer is associated with the score to predict and a confidence score. Then 6,000 student answers were used for the testing phase. Quadratic weighted kappa is used to evaluate agreement between the predicted scores and the resolved scores from 2 human judges. The top methodology papers are also available, however a few participants chose to keep their code and methodology private. This resulted in a modified top-5 ranking that excludes good submissions that ranked 1st, 5th, and 6th from the original leaderboard. We review the top three performing systems from the modified ranking: Tandella '12, Zbontar '12, and Conort '12.

ASAP '12 SAS: Tandella '12 (29) Tandalla (2012) uses a machine learning solution with regression. Features comprised of a set of hand-crafted expressions that give binary measurements as to whether an important pattern is present in the answer. This implies that the system is highly fitted to the questions. An interesting idea is to include the assessments for both judges in the model even when there is disagreement, which would create a model that naturally favors the cases where there is agreement, whilst also taking the disagreement into account. The overall regression model comprised predictions of two random forest and two gradient boosting models.

[5] `http://www.kaggle.com`
[6] `http://www.kaggle.com/c/asap-aes`
[7] `http://www.kaggle.com/c/asap-sas`
[8] `http://www.kaggle.com/c/asap-sas/forums/t/4266/phase-3`
[9] The organizers say "short answer scoring" (SAS) instead of ASAG.

ASAP '12 SAS: Zbontar '12 (35) Zbontar (2012) uses a stacking method to combine several models into a final ridge regression (Marquardt and Snee, 1975) model. Several bag-of-words representations are formulated based on character n-grams that comprised just the character n-grams themselves, or in combination with some of the natural language processing strategies or latent semantic indexing (Papadimitriou *et al.*, 1998). The base learners that formed the combined model are ridge regression, support vector regression, gradient boosting, random forests, and k-nearest-neighbor. Zbontar (2012) observes that the stacking method has been successful in other competitions, hence the decision to implement it for the ASAP '12 SAS competition.

ASAP '12 SAS: Conort '12 (11) Conort (2012) is another to use stacking. The stacking model uses 81 different models as features, and ordinary least squares (Hayashi, 2000) is used to create the final combined model. Original features include n-grams plus counts and ratios such as characters, words, word length, verb occurrences, transition words, spelling errors, and some types of punctuation. The machine learning algorithms that are used to produce the final model were regularized generalized linear models, support vector machines, random forests, and gradient boosting machines.

RTE RTE is a series of competitions on recognizing textual entailment. RTE began in 2005 with a corpus of 1,367 pairs of texts where the task is to determine if the hypothesis text can be inferred from a second given text (Dagan *et al.*, 2006). Judgments are binary and the corpus is class-balanced. Evaluation is based on accuracy and average precision of confidence-ranked submissions. Since then, variations of the competition ran annually for the next six consecutive years (Bar-Haim *et al.*, 2006; Giampiccolo *et al.*, 2007, 2008; Bentivogli *et al.*, 2009, 2010, 2011). New data sets were introduced as well as new sub-tasks or pilot-tasks such as differentiating unaddressed and contradicting entailment, providing justifications, entailment search, detecting novel information, and knowledge-base population.

RTE took a break in 2012, but returned in 2013 as a shared RTE and ASAG task. This time the RTE task is based on the notion of *partial* textual entailment, where not one but many hypotheses must be inferred or otherwise from a text. This broad idea has similarity to concept mapping in ASAG as mentioned above. This task only received one submission, and we do not review it as it is not an ASAG system by definition. However, the SemEval '13 Task 7 ASAG task is reviewed extensively as follows.

SemEval '13 Task 7 (37) SemEval '13 Task 7 is the Joint Student Response Analysis and Eighth Recognizing Textual Entailment Challenge (Dzikovska *et al.*, 2013), which was a part of the semantic evaluation (SemEval 2013) workshop series. This competition was the first large-scale and non-commercial ASAG competition. The corpora comprised data from a tutorial dialog system for high school physics (Beetle) and primary school science questions from grades 3–6 (SciEntsBank). Approximately 8,000 student answers are included across all questions. A 5-way categorical grading scheme is defined with labels "correct", "partially correct incomplete", "contradictory", "irrelevant", and "non domain". In addition, 3-way and 2-way grading schemes are included based upon collapsed versions of the above. Yet another dimension to the data is the degree of domain adaptation (Prettenhofer and Stein, 2011) required in the solutions. That is, some of the test data is for unseen answers to the same questions, some is for unseens questions in the same domain, and the rest is for questions in unseen domains. Therefore a significant advancement for ASAG research is the notion of unseen domains that provides a framework to pursue solutions that are genuinely generalizable.

We now turn to some of the specific systems, including Dzikovska '12 (Dzikovska *et al.*, 2012) as the strongest baseline, three of the top entries, and Levy '13 (Levy *et al.*, 2013), which was created outside of the competition. In determining three strongly performing systems to report as top entries, we consider the most difficult and novel dimensions to the competition, since there is no notion of an

overall winner. Specifically, we concentrate on the 5-way task as being the most difficult, the unseen domains as being the most novel and generalizable part of the competition, and the macro-averaged F_1 measure since accuracy and micro-averaged F_1 do not have an in-built mechanism for accommodating class imbalance. The best performing systems by this definition are SoftCardinality$_1$, UKP-BIU$_1$, and ETS$_2$ respectively.[10]

SemEval '13 Task 7: Dzikovska '12 (13) Dzikovska *et al.* (2012) provide simple baseline systems for the competition. The most advanced baseline is a lexical similarity system based on four features computed from the Text::Similarity package:[11] count of overlapping words, F1, Lesk, and cosine scores. These four features are combined with a C4.5 decision tree.

SemEval '13 Task 7: SoftCardinality (28) SoftCardinality (Jimenez *et al.*, 2013) is based on the idea of soft cardinality as an extension to classical cardinality. Using this idea, the system utilizes measurements of textual overlap based on the questions, teacher answers, and student answers. The measure is effectively recursive, whereby the overlap of words based on character n-grams is the smallest unit, which is then combined to the sentence level for words, then the passage level. Altogether 42 soft cardinality features are extracted from the text. Classification is made with a J48 graft tree, and the models are improved by bagging (Breiman, 1996).

SemEval '13 Task 7: UKP-BIU (31) UKP-BIU (Zesch *et al.*, 2013) is based on combining multiple textual similarity measures together using DKPro Similarity (Bär *et al.*, 2013) and BIUTEE (Stern and Dagan, 2011) as established technology for textual similarity and entailment respectively. Six families of features are used comprising of bag-of-words features, syntactic features, basic similarity features, semantic similarity features, spelling features, and entailment features. The most effective model utilizes all six feature families with a naive Bayes classifier from Weka (Hall *et al.*, 2009).

SemEval '13 Task 7: ETS (16) ETS (Heilman and Madnani, 2013) employs stacking (also seen in some top ASAP '12 SAS systems) and domain adaptation as a technique to apply non-uniform weights to the features in any model. Four classes of features are considered that include the lexical similarity features from the competition baseline system (Dzikovska *et al.*, 2012), an "intercept feature" used for modeling class distribution, word and character n-gram features, and text similarity features. The final model is created with a logistic regression classifier.

SemEval '13 Task 7: Levy '13 (22) Levy *et al.* (2013) implement several competing solutions to partial and full textual entailment on the SemEval '13 Task 7 data set. This is done immediately after and outside of the competition, so this shows the recent impact of the competition for promoting comparable solutions. Here, the partial textual entailment work operates on the level of facets. In the evaluation, the work investigates if each facet is expressed or unaddressed in the student answers, and how the knowledge can be combined for full textual entailment.

Five competing systems are implemented where partial textual entailment is analyzed in terms of a bag-of-words model, lexical inference with semantically related words, syntactic inference with dependency trees, disjunction of all three, and majority voting of all three. The majority voting implementation is most effective. This model is then adapted to full textual entailment, where the answer is marked as correct if all facets are expressed. It should be noted that the work assumes the existence of a manual facet annotation process.

[10] Each team is allowed to submit three runs, so the subscripts here refer to the run numbers.

[11] `http://search.cpan.org/dist/Text-Similarity`

COMPONENT ANALYSIS

Having completed our historical review of the literature, we now conduct a higher-level analysis according to common components. That is, we consider dimensions across systems, instead of focusing on one system at a time. Therefore we now review the six dimensions of data sets, natural language processing, model building, grading models, model evaluation, and effectiveness respectively in the subsections that follow. These map directly to the aforementioned artifacts and processes from Figure 1.

Before reviewing these dimensions, we must begin by highlighting some of the general organization. First, we omit some systems that have no or insufficient empirical contributions when reviewing the data sets, model building, model evaluation, and effectiveness. This comprises ATM, Auto-Assessor and WebLAS as having no empirical contributions, and Thomas '03 as using a maximum of 20 student answers in experiments. Second, we also group the systems in the era of evaluation under their respective competition names when discussing the data sets, model building, and model evaluation, as these dimensions are common at the competition level. Third, it is apparent that not all data is available for the properties we wish to analyze, and missing data are marked as "??" in our tables. Noting missing data is interesting to indicate trends of underreporting in the original work.

This section focuses on the details and trends of the components. We revisit the highlights when concluding with "lessons learned" in the last section of this article.

Data Sets

For the data sets dimension, we focus on qualitative properties comprising the cohort, year level, language, and topic as summarized in Table 3. Concerning the cohort first and only counting each system name once, we find that university data sets dominate school data sets by about double. This can be explained by the fact that many authors of these publications are academics using data sets from their own teaching experiences. We also have one cohort marked as "industry" for the SAMText system where the authors created their data set by requesting participation from their colleagues.

Concerning the year level, the data varies greatly. One trend is that all non-reported year levels are from the university cohort. Here we hypothesize that academics using data sets from their own teaching experiences assume that the year level is self-evident through teaching listings on university websites and similar. Given that this historical data is not so easy to acquire, we recommend that more background about the university course is supplied in describing data sets. We also have identified some foreign-language data sets for ESL (English as a Second Language) and GSL (German as a Second Language) students. Here, the progress is measured using the notion of language units instead of year level or age.

Concerning the language, the reviewed body of work is represented by four languages: Chinese, English, German, and Spanish, with English dominating. Chinese poses an additional challenge due to the requirement of word segmentation.

Concerning the topic, the data varies greatly again. There are many topics from computer science disciplines, since many developers of the systems are also academics from computer science and related areas. Reading comprehension and sciences are also popular.

The other property that we examined is data availability, and from what we can see only the data sets connected to the Mohler '09 system, CoMiC project, and SemEval '13 Task 7 competition are public. We note that the work to develop these open data sets is fairly recent, therefore the era of evaluation in ASAG has a lot of room to develop. Supporting software can help too in terms of a software framework capable of providing data confidentiality so that private data sets can form part of the era of evaluation. An example is the TIRA evaluation framework for experiments in information retrieval and related top-

Table 3
List of data sets. The abbreviations OOP and RC are short-hand for object-oriented programming and reading comprehension. The ID numbers with suffixes ('a', 'b', 'c', etc.) in Tables 3, 4, 5, 7, and 10 form part of our cross-referencing mechanism between these tables for system versions beyond our historical analysis; refer to the cited papers for details of these systems.

ID	System	Reference	Cohort	Year level	Lang.	Topic
1	Atenea	Alfonseca and Pérez (2004)	University	??	EN, ES	Operating systems
1a	Atenea	Pérez and Alfonseca (2005)	University	??	ES	Operating systems
1b	Atenea	Pérez *et al.* (2005a)	University	??	EN, ES	Operating systems, OOP
1c	Atenea	Pérez *et al.* (2005b)	University	??	ES	Operating systems
4	auto-marking	Sukkarieh *et al.* (2003)	School	Age 15-16	EN	Biology
4a	auto-marking	Sukkarieh *et al.* (2004)	School	Age 15-16	EN	Biology
4b	auto-marking	Pulman and Sukkarieh (2005)	School	Age 15-16	EN	Biology
5	AutoMark	Mitchell *et al.* (2002)	School	Age 11	EN	Science
6	Burstein '96	Burstein *et al.* (1996)	University	Postgraduate	EN	Police training
7	c-rater	Leacock and Chodorow (2003)	School	Grade 4, 8, 11	EN	Maths, RC
7a	c-rater	Attali *et al.* (2008)	University	Postgraduate	EN	Biology, psychology
7b	c-rater	Sukkarieh and Bolge (2008)	??	??	EN	Biology, RC
7c	c-rater	Sukkarieh and Stoyanchev (2009)	School	Grade 7-8	EN	Maths, RC
7d	c-rater	Sukkarieh (2010)	School	Grade 7-8	EN	Maths, RC
8	CAM	Bailey and Meurers (2008)	University	ESL intermediate	EN	RC
9	CoMiC-DE	Meurers *et al.* (2011a)	University	GSL all levels	DE	RC
10	CoMiC-EN	Meurers *et al.* (2011b)	University	ESL intermediate	EN	RC
12	CoSeC-DE	Hahn and Meurers (2012)	University	GSL all levels	DE	RC
14	e-Examiner	Gütl (2007)	University	??	EN	Computer science
15	eMax	Sima *et al.* (2009)	University	??	EN	Computer architecture
17	FreeText Author	Jordan and Mitchell (2009)	University	??	EN	Science
18	Horbach '13	Horbach *et al.* (2013)	University	GSL all levels	DE	RC
19	Hou '11	Hou and Tsao (2011)	University	??	EN	Automata, formal languages
20	IndusMarker	Siddiqi and Harrison (2008a)	University	Undergraduate	EN	Biology
20a	IndusMarker	Siddiqi *et al.* (2010)	University	Undergraduate	EN	OOP
21	Klein '11	Klein *et al.* (2011)	University	??	??	Algorithms
23	Madnani '13	Madnani *et al.* (2013)	School	Grade 6-9	EN	RC
24	Mohler '09	Mohler and Mihalcea (2009)	University	Undergraduate	EN	Data structures
25	Nielsen '08	Nielsen *et al.* (2008b)	School	Grade 3-6	EN	Science
26	PMatch	Jordan (2012a)	University	??	EN	Science
27	SAMText	Bukai *et al.* (2006)	Industry	Employees	EN	Botany
32	Wang '08	Wang *et al.* (2008)	School	High school	CN	Earth science
34	Willow	Pérez-Marín and Pascual-Nieto (2011)	University	??	EN, ES	Operating systems, OOP
36	ASAP '12 SAS	Hewlett Foundation (2012)	School	High school	EN	Interdisciplinary
37	SemEval '13 Task 7	Dzikovska *et al.* (2013)	School	High school	EN	Electronics
37	SemEval '13 Task 7	Dzikovska *et al.* (2013)	School	Grade 3-6	EN	Science

ics (Gollub *et al.*, 2012a,b,c), which has allowed the organizers of the PAN competition series (Potthast, 2011) to maintain control over data assets, whilst still providing a stimulating competition.

Natural Language Processing

Natural Language Processing (NLP) techniques are required to analyze the language in student answers. The techniques are either linguistic processing techniques that perform textual manipulation, or statistical techniques based on the features extracted from them. In this section we review both categories, beginning with linguistic processing techniques.

From reviewing all systems, we find 17 different linguistic processing techniques as shown in Figure 4. We stress that the data is only indicative of the linguistic processing techniques, since we find that 8 of 35 sets of literature did not document the linguistic processing techniques at all. In some cases, we assume that the authors deliberately chose to focus on other dimensions of the work. In some other cases, we presume that the authors consider some linguistic processing techniques as obvious or trivial, and hence not worth documenting. The best example of this is perhaps tokenization, where the texts are

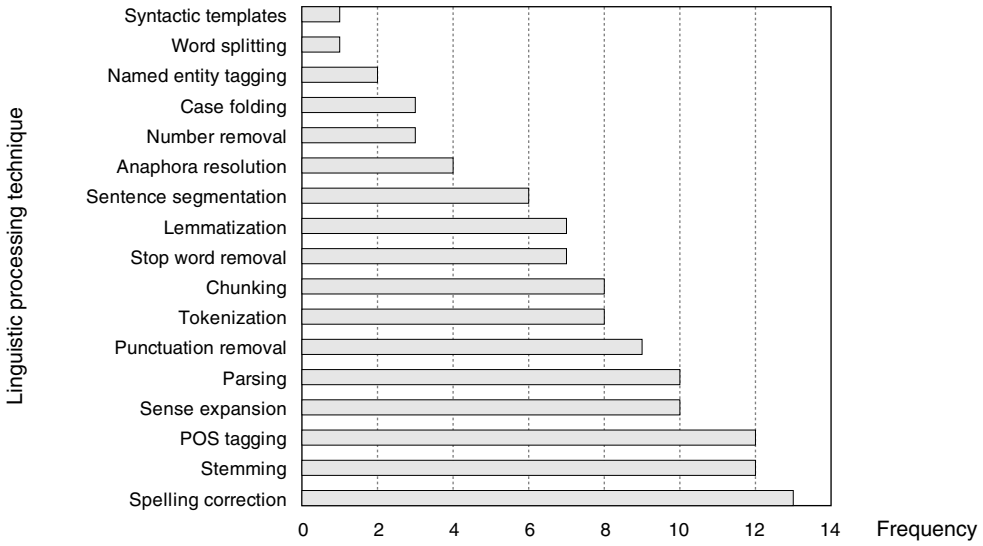

Figure 4. Frequency of linguistic processing techniques from the literature.

segmented into words. Other examples may be case folding and punctuation removal. We expect linguistic processing techniques like these to be more common in practice than shown in Figure 4. Nevertheless, we noted 3.3 linguistic processing techniques per system on average.

We also note that the common terminology from the field of NLP is adopted in Figure 4, and there are some small differences to the ASAG literature. Here, we presume some basic understanding of NLP to make these transitions. One uncommon case is "syntactic templates", whereby we refer to syntactic templates (Szpektor and Dagan, 2007) up to the sentence level being used to canonicalize expressions with equivalent meaning, such as the hare-tortoise example from earlier. Another uncommon case is "word splitting" for segmenting Chinese data (Wang *et al.*, 2008).

In summary, many types of linguistic processing techniques may be needed depending on the era or other dimension of the problem. In order to organize this work, we categorize the 17 linguistic processing techniques we found as falling into one of five broad categories: lexical, morphological, semantic, syntactic, and surface. We represent this organization as the hierarchy in Figure 5.

Unlike linguistic processing techniques, statistical techniques result in singular measurements or "features" that typically only apply to machine learning systems. The 15 systems we consider are those from the era of machine learning and all machine learning systems from the era of evaluation.

Two bodies of work were particularly interesting in the choice of features. First, the CAM, CoMiC-EN, CoMiC-DE, and Horbach '13 systems all use the same or very similar features. The specific features are for alignment "at different levels and using different types of linguistic abstraction" (Meurers *et al.*, 2011b) when comparing teacher and student answers. Here, the continuity allowed the researchers to focus on other avenues such as multiple languages, architecture, and creating standards for data sets. A second interesting example is the UKP-BIU system from SemEval '13 Task 7 that used a very large feature space. Here, the textual similarity and entailment technology is significant due to public availability and reuse potential.

Many of the features can be also categorized by the five broad headings from Figure 5. The features that we found are as follows. Lexical: Bag-of-words, spelling errors, and stop word overlap. Morphological: Stem matches. Semantic: Lesk and LSA. Syntactic: Dependency parse tree features, POS tags, and verb occurrences. Surface: Character count, word count, sentence count, punctuation, and word length.

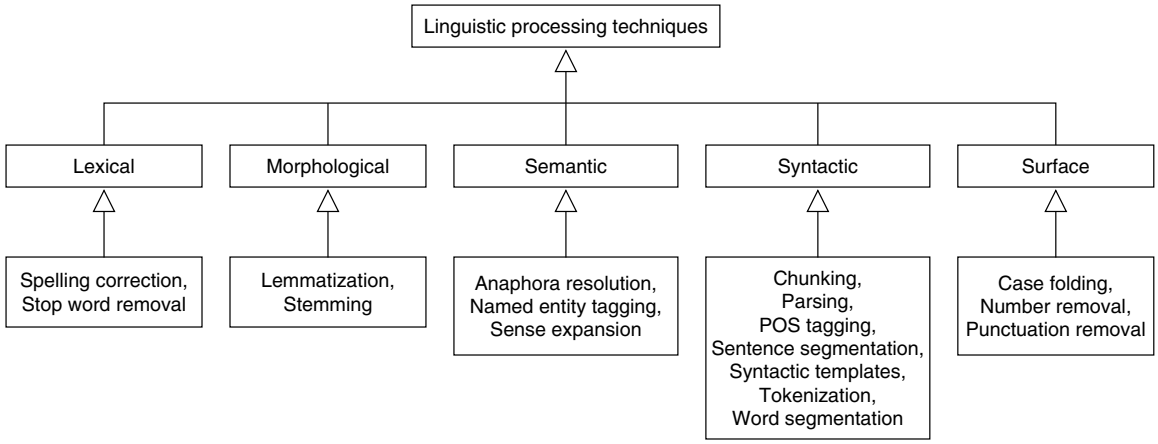

Figure 5. Taxonomy of linguistic processing techniques from the literature.

Considering other trends, we also found the use of n-grams common whether they are on the character level, word level, or another representation created by a linguistic processing technique. Other mini-themes for features were information retrieval (term frequency, cosine, F_1), machine translation (ROUGE), textual similarity (edit distance), overlaps (greedy string tiling (Wise, 1993), longest common subsequence, overlap at start of sentences), entailment, entropy, and presence of specific phrases.

Model Building

We now turn to the quantitative dimensions of the data sets and how this data is organized for model building. Here, a model is defined as any representation of the student answers that allows a mapping between the student answers and the correct score with reasonable accuracy. A quantitative summary of the original data and the organization for model building is given in Table 4. Here, we first list the number of questions (Q), number of teacher answers (TA), and number of student answers (SA). Then we describe how this data is divided between the tasks of model building (M.Dat) and model evaluation (E.Dat) since these mappings vary greatly.

We first note that we have multiple rows for some references in Table 4. In these cases, the data covers multiple topics often with a different amount of data for each topic. In addition, there are often different splits of the data for model building and evaluation, hence we list these separately too. This trend is particularly true in the c-rater work.

From the Table 4 data, we first see that the number of teacher answers is frequently not reported. The teacher answers usually exist, but we assume only one teacher answer per question in many cases, and hence some authors may have neglected to mention this specifically. Alternatively, the authors may acknowledge the teacher answers in a way that cannot be quantified such as "*the* teacher answers". In addition, we found it cumbersome to describe the teacher answers for concept mapping systems, since it would be necessary to report additional data including the number of concepts and the number of teacher answers per concept. Here, we just say "concepts" and invite the interested reader to follow the references for the details. Another special case is the ASAP '12 SAS competition, where the teacher answers did not form part of the competition, hence we say zero teacher answers here. The SemEval '13 Task 7 competition is also an unusual case, as there is exactly one teacher answer per question for the SciEntsBank part, but one or more teacher answers per question for the Beetle part.

Now considering how the data is divided for model building and evaluation, we see a great deal

Table 4
Data set sizes and model building organization. The columns are number of questions (Q), number of teacher answers (TA), number of student answers (SA), the model data (M.Dat), and the evaluation data (E.Dat). Acronyms are used to describe cross-validation experiment designs such as leave-one-out cross valiation (LOOCV) and five-fold cross validation (5FCV).

ID	System	Reference	Q	TA	SA	M.Dat	E.Dat
1	Atenea	Alfonseca and Pérez (2004)	7	34	885	TA	SA
1a	Atenea	Pérez and Alfonseca (2005)	9	40	886	TA	SA
1b	Atenea	Pérez *et al.* (2005a)	10	44	924	TA	SA
1c	Atenea	Pérez *et al.* (2005b)	5	27	672	TA	SA
4	auto-marking	Sukkarieh *et al.* (2003)	3	??	798	0.76 SA	0.24 SA
4a	auto-marking	Sukkarieh *et al.* (2004)	2	??	??	??	??
4b	auto-marking	Pulman and Sukkarieh (2005)	9	??	2,340	0.77 SA	0.23 SA
5	AutoMark	Mitchell *et al.* (2002)	2	??	340	TA + 0.29 SA	0.71 SA
6	Burstein '96	Burstein *et al.* (1996)	1	Concepts	378	0.46 SA	0.54 SA
7	c-rater	Leacock and Chodorow (2003)	5	Concepts	1,750	0.29 SA	0.71 SA
7	c-rater	Leacock and Chodorow (2003)	7	Concepts	1,400	0.50 SA	0.50 SA
7a	c-rater	Attali *et al.* (2008)	11	Concepts	389	0.19 SA	0.81 SA
7a	c-rater	Attali *et al.* (2008)	11	Concepts	640	0.14 SA	0.86 SA
7b	c-rater	Sukkarieh and Bolge (2008)	1	Concepts	1,000	0.50 SA	0.50 SA
7b	c-rater	Sukkarieh and Bolge (2008)	1	Concepts	1,000	0.50 SA	0.50 SA
7c	c-rater	Sukkarieh and Stoyanchev (2009)	7	Concepts	990	0.41 SA	0.59 SA
7c	c-rater	Sukkarieh and Stoyanchev (2009)	5	Concepts	650	0.48 SA	0.52 SA
7d	c-rater	Sukkarieh (2010)	8	Concepts	594	??	??
7d	c-rater	Sukkarieh (2010)	10	Concepts	776	??	??
8	CAM	Bailey and Meurers (2008)	75	??	566	0.55 SA	0.45 SA
9	CoMiC-DE	Meurers *et al.* (2011a)	117	136	610	LOOCV SA	LOOCV SA
9	CoMiC-DE	Meurers *et al.* (2011a)	60	87	422	LOOCV SA	LOOCV SA
10	CoMiC-EN	Meurers *et al.* (2011b)	75	??	566	0.55 SA	0.45 SA
12	CoSeC-DE	Hahn and Meurers (2012)	167	223	1,032	LOOCV SA	LOOCV SA
14	e-Examiner	Gütl (2007)	8	8	184	0.48 SA	0.52 SA
15	eMax	Sima *et al.* (2009)	3	??	611	TA + 0.10 SA	0.90 SA
17	FreeText Author	Jordan and Mitchell (2009)	7	??	1,067	TA	SA
18	Horbach '13	Horbach *et al.* (2013)	177	223	1,032	LOOCV SA	LOOCV SA
19	Hou '11	Hou and Tsao (2011)	9	9	342	5FCV SA	5FCV SA
20	IndusMarker	Siddiqi and Harrison (2008a)	5	??	1,396	TA + 0.18 SA	0.82 SA
20a	IndusMarker	Siddiqi *et al.* (2010)	87	87	19,575	TA + 0.11 SA	0.89 SA
21	Klein '11	Klein *et al.* (2011)	7	??	282	0.83 SA	0.17 SA
23	Madnani '13	Madnani *et al.* (2013)	2	??	2,695	5FCV SA	5FCV SA
24	Mohler '09	Mohler and Mihalcea (2009)	7	7	630	TA	SA
24a	Mohler '09	Mohler *et al.* (2011)	80	??	2,273	12FCV SA	12FCV SA
25	Nielsen '08	Nielsen *et al.* (2008b)	54	??	85,481	0.64 SA	0.36 SA
25	Nielsen '08	Nielsen *et al.* (2008b)	22	??	61,666	0.89 SA	0.11 SA
25	Nielsen '08	Nielsen *et al.* (2008b)	211	??	58,126	0.95 SA	0.05 SA
26	PMatch	Jordan (2012a)	11	??	20,114	TA	SA
27	SAMText	Bukai *et al.* (2006)	2	2	129	TA	SA
32	Wang '08	Wang *et al.* (2008)	4	Concepts	2,698	TA	SA
34	Willow	Pérez-Marín and Pascual-Nieto (2011)	10	44	924	TA	SA
36	ASAP '12 SAS	Hewlett Foundation (2012)	10	0	22,950	0.75 SA	0.25 SA
37	SemEval '13 Task 7	Dzikovska *et al.* (2013)	47	47+	5,119	TA + 0.77 SA	0.23 SA
37	SemEval '13 Task 7	Dzikovska *et al.* (2013)	135	135	10,804	TA + 0.46 SA	0.54 SA

of variation. Taking a system such as Atenea/Willow, we see that the mapping from the teacher and student answers to the model and evaluation data is direct. That is, $TA = M.Dat$, and $SA = E.Dat$. In this example, the teacher and student answers are compared using BLEU and LSA to compute the grades. In many other systems, the teacher answers are not used in the training model at all, but are instead used to guide the hand-marking of some student answers that are used as a training model, as in CoMiC-EN for example. This is also done in a cross-validation experiment design such as five-fold cross validation (5FCV) in the Hou '11 system and leave-one-out cross validation (LOOCV) in the Horbach '13 system. Finally, it is also possible to combine the teacher and student answers in model

Table 5
Typical data set size. The number of questions (Q), number of student answers (SA), and number of student answers per question (SA/Q) are listed. The median figures are indicative of the typical amount of data in ASAG publications. Fractional numbers are rounded to the nearest integer.

ID	System	Reference	Q	SA	SA/Q
1	Atenea	Alfonseca and Pérez (2004)	7	885	126
1a	Atenea	Pérez and Alfonseca (2005)	9	886	98
1b	Atenea	Pérez *et al.* (2005a)	10	924	92
1c	Atenea	Pérez *et al.* (2005b)	5	672	134
4	auto-marking	Sukkarieh *et al.* (2003)	3	266	89
4a	auto-marking	Sukkarieh *et al.* (2004)	2	??	??
4b	auto-marking	Pulman and Sukkarieh (2005)	9	260	29
5	AutoMark	Mitchell *et al.* (2002)	2	340	170
6	Burstein '96	Burstein *et al.* (1996)	1	378	378
7	c-rater	Leacock and Chodorow (2003)	12	3,150	263
7a	c-rater	Attali *et al.* (2008)	22	1,029	47
7b	c-rater	Sukkarieh and Bolge (2008)	2	2,000	1,000
7c	c-rater	Sukkarieh and Stoyanchev (2009)	12	1,640	137
7d	c-rater	Sukkarieh (2010)	18	1,370	76
8	CAM	Bailey and Meurers (2008)	75	566	8
9	CoMiC-DE	Meurers *et al.* (2011a)	177	1,032	6
10	CoMiC-EN	Meurers *et al.* (2011b)	75	566	8
12	CoSeC-DE	Hahn and Meurers (2012)	167	1,032	6
14	e-Examiner	Gütl (2007)	8	184	23
15	eMax	Sima *et al.* (2009)	3	611	204
17	FreeText Author	Jordan and Mitchell (2009)	7	1,067	152
18	Horbach '13	Horbach *et al.* (2013)	177	1,032	6
19	Hou '11	Hou and Tsao (2011)	9	342	38
20	IndusMarker	Siddiqi and Harrison (2008a)	5	1,396	279
20a	IndusMarker	Siddiqi *et al.* (2010)	87	19,575	225
21	Klein '11	Klein *et al.* (2011)	7	282	40
23	Madnani '13	Madnani *et al.* (2013)	2	2,695	1,348
24	Mohler '09	Mohler and Mihalcea (2009)	7	630	90
24a	Mohler '09	Mohler *et al.* (2011)	80	2,273	28
25	Nielsen '08	Nielsen *et al.* (2008b)	287	95,339	332
26	PMatch	Jordan (2012a)	11	20,114	1,829
27	SAMText	Bukai *et al.* (2006)	2	129	65
32	Wang '08	Wang *et al.* (2008)	4	2,698	675
34	Willow	Pérez-Marín and Pascual-Nieto (2011)	10	924	92
36	ASAP '12 SAS	Hewlett Foundation (2012)	10	22,950	2,295
37	SemEval '13 Task 7	Dzikovska *et al.* (2013)	182	15,923	87
	Median		9	1,029	92
	Standard deviation		70	16,721	526

building. For example, we see this in eMax whereby the original vocabulary of the teacher answers is expanded by considering the vocabulary in a small subset of good student answers. It is also possible to expand a training set in a machine learning experiment by considering the teacher answers as additional instances with perfect scores, however we never witnessed this.

Related to Table 4, we realized that an extension would allow us to answer the following question: What is the typical size of a data set in ASAG research? To answer this question, we created Table 5 as a variation of Table 4. In this table, we collapse duplicate references and aggregate the affected data. We also introduce a new column to represent the number of student answers per question (SA/Q). Taking the median figure of each column to answer our question, we say that the typical ASAG paper has 9 questions, 1,029 student answers, and 92 student answers per question.

Table 6
List of ASAG grading models sorted alphabetically under our 4-way era classification.

4-way era	Grading model	ID	System	Reference
Concept mapping (CMap)	Concept pattern matching	2	ATM	Callear *et al.* (2001)
	Lexical semantic matching	6	Burstein '96	Burstein *et al.* (1996)
	Maximum entropy	7	c-rater	Leacock and Chodorow (2003)
	Partial textual entailment	22	Levy '13	Levy *et al.* (2013)
	SVM	32	Wang '08	Wang *et al.* (2008)
Information extraction (IE)	Boolean phrase matching	30	Thomas '03	Thomas (2003)
	LRS representation matching	12	CoSeC-DE	Hahn and Meurers (2012)
	Parse tree matching	5	AutoMark	Mitchell *et al.* (2002)
	Regular expression matching	26	PMatch	Jordan (2012a)
	Regular expression matching	33	WebLAS	Bachman *et al.* (2002)
	Semantic word matching	3	Auto-Assessor	Cutrone *et al.* (2011)
	Syntactic pattern matching	4	auto-marking	Sukkarieh *et al.* (2003)
	Syntactic pattern matching	15	eMax	Sima *et al.* (2009)
	Syntactic pattern matching	20	IndusMarker	Siddiqi and Harrison (2008b)
	Syntactic-semantic pattern matching	17	FreeText Author	Jordan and Mitchell (2009)
Corpus-based methods (CBM)	BLEU + LSA	1	Atenea	Alfonseca and Pérez (2004)
	BLEU + LSA	34	Willow	Pérez-Marín and Pascual-Nieto (2011)
	ESA + LSA + knowledge-based measures	24	Mohler '09	Mohler and Mihalcea (2009)
	LSA	21	Klein '11	Klein *et al.* (2011)
	LSA	27	SAMText	Bukai *et al.* (2006)
Machine learning (ML)	Decision tree	13	Dzikovska '12	Dzikovska *et al.* (2012)
	Decision tree	25	Nielsen '08	Nielsen *et al.* (2008b)
	Decision tree	28	SoftCardinality	Jimenez *et al.* (2013)
	k-nearest neighbor	8	CAM	Bailey and Meurers (2008)
	k-nearest neighbor	9	CoMiC-DE	Meurers *et al.* (2011a)
	k-nearest neighbor	10	CoMiC-EN	Meurers *et al.* (2011b)
	k-nearest neighbor	18	Horbach '13	Horbach *et al.* (2013)
	Linear regression	14	e-Examiner	Gütl (2007)
	Linear regression	29	Tandella '12	Tandalla (2012)
	Logistic regression	23	Madnani '13	Madnani *et al.* (2013)
	Naive Bayes	31	UKP-BIU	Zesch *et al.* (2013)
	Stacking	11	Conort '12	Conort (2012)
	Stacking	16	ETS	Heilman and Madnani (2013)
	Stacking	35	Zbontar '12	Zbontar (2012)
	SVM	19	Hou '11	Hou and Tsao (2011)

Grading Models

The models used for grading depend vastly on the era. To organize and visualize this, we must collapse the era of evaluation, which we used to denote community cooperation instead of technology. Most of these systems were machine learning systems. This allows the new organization in Table 6.

This organization helps us highlight a higher organization, that of "rule-based" versus "statistical" grading models. That is, there are similarities between the CMap/IE eras (rule-based) and the CBM/ML eras (statistical). For the CMap/IE pair, we observe that the first four CMap rows in Table 6 are similar to IE as they are based on pattern matching or entailment. The fifth is a special case in that the concepts are given by the respondents separately, and therefore a method to extract the individual concepts from a composite response is no longer necessary. For the CBM/ML pair, we observe that the scores calculated by corpus-based methods have frequently been treated as individual features. As examples, Atenea and Willow use a weighted average of BLEU and LSA that could both be ML features, the latter work of the Mohler '09 system has combined the knowledge- and corpus-based methods into a machine learning system (Mohler *et al.*, 2011), and the UKP-BIU system uses corpus-based features as part of the feature set. Putting this together, a different view is possible as per Figure 6.

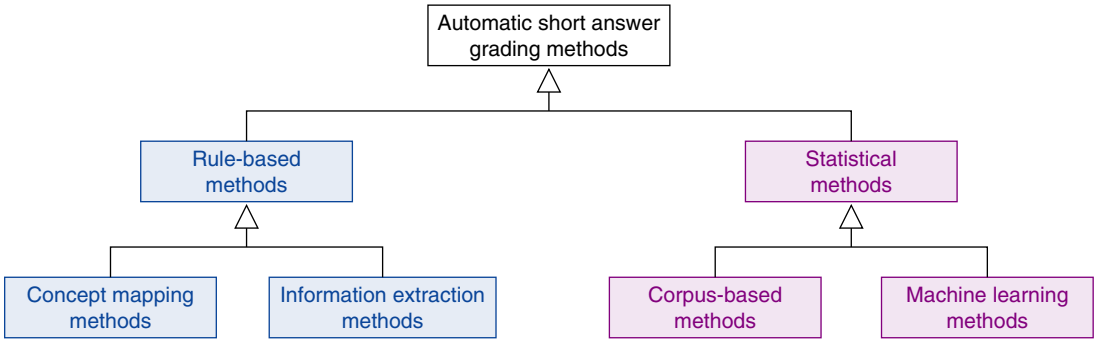

Figure 6. The four method-eras viewed as rule-based or statistical methods.

With this bigger picture, we now consider the trade-offs of these classes of methods. Our hypothesis (visualized in Figure 7) is that rule-based methods are more suitable for repeated assessment (where assessment instruments are reused) while statistical methods are more suitable for unseen questions and domains. That is, we essentially have two key properties: "repetition" and "generalization". Turning to the literature to support our hypothesis, we say that rule-based methods are more suitable for repeated assessment because it is acceptable to make additional investment in specific solutions when the benefits can be realized multiple times. Here, commercial systems can flourish such as c-rater from ETS for repeated testing. In comparison, statistical methods have flourished at the SemEval '13 Task 7 competition that requires solutions to unseen questions and domains. This situation requires a flexible solution as competition participants are only given a few months for development on the competition data sets. We also say that this body of work represents non-repeated assessment as the first and only offering of the ASAG competition at SemEval '13 at the time of writing.

A related note is that in summative, large-scale assessments that have become popular over the last 10 years (e.g.: PISA[12] has been repeated triannually since 2000), we have a rather small number of questions, large numbers of answers, and human ratings. Here, highly accurate models might be built by handcrafting patterns, as is done in many rule-based techniques. In formative assessments, which play an increasingly important role due to the personalization of learning and MOOCs (Massively Open Online Courses), we have to deal with a great variety of questions and noisy answers and ratings, such as those

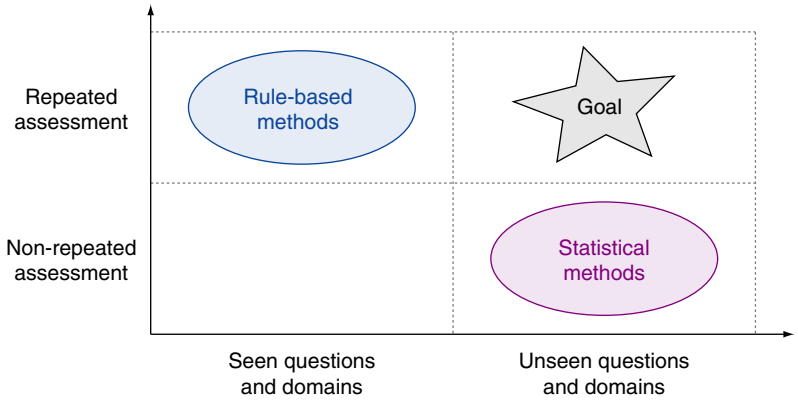

Figure 7. A continuum highlighting trade-offs between rule-based and statistical methods.

[12] http://www.oecd.org/pisa

by peers. Here, effective models might be better created by statistical techniques.

It is evident that for ASAG there is more work on statistical methods than rule-based methods in the recent literature. This is supported by Figure 3 and the fact that most of systems from the era of evaluation are statistical. It must now be questioned whether or not these methods are strong enough to reach the goal in Figure 7. We suggest that this question could be answered with more cooperation with commercial partners in the future and the repeated assessments that they command.

Model Evaluation

The model evaluation and effectiveness data are expressed together in Table 7. Here, following the system names and references, we describe the type of grading scheme as based on categories (X-way) or points (Y-point). For example, Hahn and Meurers (2012) use a categorical grading scheme in CoSeC-DE with five categories (5-way), and Gütl (2007) uses a points-based grading scheme in e-Examiner with questions worth 10 points (10-point). Next, the evaluation of human-system agreement (HSA) is given with the name of the measure (HSA.M) and the score (HSA.S). This convention is repeated for the evaluation of human-human agreement (HHA) again giving the name of the measure (HHA.M) and the score (HHA.S). Table 7 focuses on the most effective methods in each paper, instead of listing additional and inferior methods.

For the evaluation, the grading scheme is important because the types of evaluation metrics that can be applied depend on it. Therefore, it is important to provide this data so the correctness of the evaluation procedure can be verified. As shown in Table 7, this data is sometimes missing. When the data is given, the number of categories or points is almost always capped at 5. One exception is e-Examiner that uses a 10-point scale for 8 computer science questions. Here, the scoring scale is loosely defined as zero (inappropriate) to ten (very good). Having ten points may create difficulty in achieving high agreement between human judges, and may simply create unnecessary complexity for well-defined grading schemes in ASAG. The other exception is the Wang '08 system with questions worth up to 30 points, but this grading scheme is unique in that the concepts are processed in an additive way, and the respondent is rewarded for listing as many concepts as possible.

On a more general level, in ASAG we have data from three of the four levels of measurement represented from the well-known taxonomy of Stevens (1946): nominal data, ordinal data, and ratio data, but not interval data. Nominal data refers to discrete categories where an ordering effect cannot be applied. An example is part of the five-way SemEval '13 Task 7 grading scheme (Dzikovska *et al.*, 2013), where one cannot define if a "contradictory" answer is better or worse than an "irrelevant" answer. Next, ordinal data refers to discrete categories with an inherent ordering effect. An example is yet again the SemEval '13 Task 7 grading scheme (Dzikovska *et al.*, 2013), where the two-way scheme comprises of "correct" and "incorrect", and it is clear that "correct" is better. Ratio data refers to continuous measurements that have a zero-origin. An example is the e-Examiner system (Gütl, 2007), where fractional scores were allowed for 10-point questions. Related to this, in ASAG a ratio scale is often tailored to reduce the number of possible scores that can be awarded to a discrete set. An example is the discrete scoring scheme for the Madnani '13 system (Madnani *et al.*, 2013), where scores are only allowed in the precise set $\{0, 1, 2, 3, 4\}$. Finally, the non-represented interval data category from our literature review is related to ratio data except that it does not have a zero origin. Real world examples are temperatures, dates, and geographic coordinates, that have arbitrary origins.

We should also point out that the choice of evaluation metric can sometimes be used for two data types at once. For example, a 2-way ordinal scale of "incorrect" and "correct" could be interpreted as a $\{0, 1\}$ ratio scale. Similarly, ratio data could be considered as discrete or continuous depending on the

Table 7
Model evaluation data. The grading scheme is based on some number of categories (X-way) or points (Y-point). The human-system agreement measure (HSA.M) and score (HSA.S) follow giving an indication of system effectiveness. Then the human-human agreement measure (HHA.M) and score (HHA.S) follow indicating the natural rates of disagreement between humans. Entries marked with an asterisk (*) are reconciled annotations where disagreements were removed.

ID	System	Reference	Grading	HSA.M	HSA.S	HHA.M	HHA.S
1	Atenea	Alfonseca and Pérez (2004)	?-point	r	0.30-0.70	2-judge	??
1a	Atenea	Pérez and Alfonseca (2005)	0.5-1.5-point	r	0.33-0.87	??	??
1b	Atenea	Pérez *et al.* (2005a)	?-point	r	0.33-0.85	??	??
1c	Atenea	Pérez *et al.* (2005b)	?-point	r	0.35-0.77	??	??
4	auto-marking	Sukkarieh *et al.* (2003)	2-point	acc	0.88	??	??
4a	auto-marking	Sukkarieh *et al.* (2004)	1-2-point	acc	0.81	??	??
4b	auto-marking	Pulman and Sukkarieh (2005)	1-4-point	acc	0.84	??	??
5	AutoMark	Mitchell *et al.* (2002)	1-point	acc	0.93	??	??
5	AutoMark	Mitchell *et al.* (2002)	2-point	acc	0.83	??	??
6	Burstein '96	Burstein *et al.* (1996)	?-point	acc	0.81	??	??
7	c-rater	Leacock and Chodorow (2003)	2-5-point	acc	0.81-0.91	agr	0.87-0.94
7	c-rater	Leacock and Chodorow (2003)	2-5-point	κ_u	0.58-0.86	κ_u	0.77-0.90
7	c-rater	Leacock and Chodorow (2003)	2-point	acc	0.84	1-judge	n/a
7	c-rater	Leacock and Chodorow (2003)	2-point	κ_u	0.74	1-judge	n/a
7a	c-rater	Attali *et al.* (2008)	1-point	κ_l	0.68	κ_l	0.84
7a	c-rater	Attali *et al.* (2008)	1-point	κ_l	0.87	κ_l	0.92
7b	c-rater	Sukkarieh and Bolge (2008)	2-point	κ_u	0.54	κ_u	0.69
7b	c-rater	Sukkarieh and Bolge (2008)	3-point	κ_u	0.71	κ_u	0.76
7c	c-rater	Sukkarieh and Stoyanchev (2009)	?-3 point	κ_u	0.63	κ_u	0.42-0.97
7c	c-rater	Sukkarieh and Stoyanchev (2009)	?-3 point	κ_u	0.64	κ_u	0.65-0.82
7d	c-rater	Sukkarieh (2010)	1-2 point	κ_q	0.75	κ_q	0.48-0.98
7d	c-rater	Sukkarieh (2010)	2-3-point	κ_q	0.62	κ_q	0.27-0.83
8	CAM	Bailey and Meurers (2008)	2-way	acc	0.88	agr	*1.00
8	CAM	Bailey and Meurers (2008)	5-way	acc	0.87	agr	*1.00
9	CoMiC-DE	Meurers *et al.* (2011a)	2-way	acc	0.84	agr	*1.00
9	CoMiC-DE	Meurers *et al.* (2011a)	2-way	acc	0.84	agr	*1.00
10	CoMiC-EN	Meurers *et al.* (2011b)	2-way	acc	0.88	agr	*1.00
10	CoMiC-EN	Meurers *et al.* (2011b)	5-way	acc	0.79	agr	*1.00
12	CoSeC-DE	Hahn and Meurers (2012)	2-way	acc	0.86	agr	*1.00
14	e-Examiner	Gütl (2007)	10-point	r	0.81	1-judge	n/a
15	eMax	Sima *et al.* (2009)	??	acc	0.82	??	??
17	FreeText Author	Jordan and Mitchell (2009)	??	acc	0.89-1.00	??	??
18	Horbach '13	Horbach *et al.* (2013)	2-way	acc	0.84	agr	*1.00
19	Hou '11	Hou and Tsao (2011)	2-way	$prec$	0.72	1-judge	n/a
19	Hou '11	Hou and Tsao (2011)	3-way	$prec$	0.72	1-judge	n/a
20	IndusMarker	Siddiqi and Harrison (2008a)	??	acc	0.93-0.95	??	??
20a	IndusMarker	Siddiqi *et al.* (2010)	??	acc	0.84-1.00	??	??
21	Klein '11	Klein *et al.* (2011)	??	r	0.00-1.00	1-judge	n/a
23	Madnani '13	Madnani *et al.* (2013)	4-point	acc	0.52-0.65	1-judge	n/a
24	Mohler '09	Mohler and Mihalcea (2009)	5-point	r	0.47	r	0.64
24a	Mohler '09	Mohler *et al.* (2011)	5-point	r	0.52	agr	0.58
25	Nielsen '08	Nielsen *et al.* (2008b)	5-way	acc	0.61	??	??
25	Nielsen '08	Nielsen *et al.* (2008b)	5-way	acc	0.62	??	??
25	Nielsen '08	Nielsen *et al.* (2008b)	5-way	acc	0.76	??	??
26	PMatch	Jordan (2012a)	??	acc	0.70-0.99	1-judge	n/a
27	SAMText	Bukai *et al.* (2006)	2-way	κ_u	0.54-0.91	??	??
27	SAMText	Bukai *et al.* (2006)	5-point	r	0.74-0.78	r	0.86-0.93
32	Wang '08	Wang *et al.* (2008)	28-30-point	r	0.92	r	0.96
34	Willow	Pérez-Marín and Pascual-Nieto (2011)	?-point	r	0.34-0.83	r	0.02-0.82
36	ASAP '12 SAS	Hewlett Foundation (2012)	2-3-point	κ_q	0.75	2-judge	??
37	SemEval '13 Task 7	Dzikovska *et al.* (2013)	2/3/5-way	acc, F_1	many	??	??

granularity. Therefore, it may not hurt to do conversions in some cases to allow additional comparison. We now discuss all metrics from Table 7 given as the summary in Table 8 according to the taxonomy of Stevens (1946), followed by some general remarks. Note the absence of interval data.

Table 8
Evaluation metrics from the literature.

Nominal and Ordinal Data	Ratio Data (Discrete)	Ratio Data (Continuous)
acc, agr, κ_u, F_1, $prec$	κ_l, κ_q	r

Nominal and Ordinal Data We found that the nominal and ordinal data metrics in our review apply to either data type. For these levels of measurement, the simple notion of "accuracy" or "agreement" is the most common measure of HSA and HHA. Our review of the terminology in the literature shows that accuracy and agreement are used to mean the same thing in different publications. Ideally, the terminology should not overlap. In this respect, one possible solution is to use "accuracy" (acc) for HSA and "agreement" (agr) for HHA. In any case, it is important to be explicit about what metric is used and provide a reference if possible.

Interestingly, some agr scores are perfect (1.00) in Table 7. This happens when ratings with disagreements are discarded to create a "perfect" gold standard. However, this practice may create bias towards answers that are easiest to grade. Instead, it is perfectly reasonable to report HSA.S scores for each judge individually.

The next metric for nominal and ordinal data in Table 8 is Cohen's κ (Cohen, 1960) as a chance-corrected measure of agreement. This metric is unweighted in that all mismatches are treated equally (we use notation κ_u for unweighted κ). This metric is therefore appropriate for nominal and ordinal data where the amount of difference cannot be quantified. The measure is defined as: $\kappa_u = (P_o - P_c)/(1 - P_c)$, where P_o and P_c are the probabilities of observed and chance agreement respectively.

The only other metrics represented for nominal and ordinal data in Table 8 are those based on information retrieval principles: F_1 and precision ($prec$). These appear in the SemEval '13 Task 7 competition and the Hou '11 evaluation respectively. F_1 is chosen for SemEval '13 Task 7 as it is suitable for all combinations of positive and negative class labels in the competition scenarios, but other metrics were provided for the benefit of the participants. Concerning precision, Hou and Tsao (2011) only broadly remark that precision is "widely used to evaluate the systems in the NLP domain" (Hou and Tsao, 2011). Recall, as the counterpart to precision and a component of F_1, is also worth mentioning in passing, as it formed part of the evaluation by Dzikovska *et al.* (2012).

Ratio Data (Discrete) Considering discrete ratio data in Table 8, the weighted variant Cohen's κ is the metric applied (Cohen, 1968). This metric is appropriate for discrete ratio data because contingencies can be made for different amounts of disagreement. So for example, we should penalize a two-point disagreement more than a one-point disagreement on a multiple-point grading scale. The penalty that can be applied however is open. Generally, all weights are expressed in a matrix, and any set of weights may be inserted. However, it can be convenient to apply easily interpretable weights, typically linear or quadratic weights (we use notation κ_l and κ_q respectively). A demonstration of these weighting schemes is provided in Table 9; see Sim and Wright (2005) for calculation details. From this example, the difference between kappas is apparent: $\kappa_u \leq \kappa_l \leq \kappa_q$. When choosing between the weighted kappas, κ_l is the arithmetically logical choice given that a two-point disagreement can be said to be twice as bad as a one-point disagreement. However, Table 7 shows that κ_q has been used more frequently.

Ratio Data (Continuous) Considering continuous ratio data in Table 8, the sample Pearson correlation coefficient (Rodgers and Nicewander, 1988), also known as Pearson's r, is the only metric used as shown in Table 7. Some other examples that *could* be used are Spearman's rank correlation coefficient (Spearman's r) (Spearman, 1904) or Kendall's rank correlation coefficient (Kendall's τ) (Kendall, 1938).

Table 9
Comparison of kappa weights for a question scored up to 4 points.

Kappa metric	0-point difference	1-point difference	2-point difference	3-point difference	4-point difference
κ_u	1.00	0.00	0.00	0.00	0.00
κ_l	1.00	0.75	0.50	0.25	0.00
κ_q	1.00	0.94	0.75	0.44	0.00

General Remarks The literature often provides guidelines for interpreting the values of the evaluation metrics. For example, interpreting any kappa value can be considered as follows: $\kappa < 0.4$ (poor), $0.4 \leq \kappa < 0.75$ (fair to good), and $0.75 \leq \kappa$ (excellent) (Fleiss, 2003). As another example, interpreting r can be considered as follows: $0 \leq r < 0.1$ (none), $0.1 \leq r < 0.3$ (small), $0.3 \leq r < 0.5$ (medium), and $0.5 \leq r \leq 1.0$ (large) (Cohen, 1992). It may be wise however to not interpret the scores so strictly. In particular, Fleiss (2003) apply the same scale to weighted and unweighted kappas, but this does not accommodate the disparity demonstrated as per Table 9. In addition, empirical work by Bakeman *et al.* (1997) also demonstrates flaws in applying broad rules.

In completing Table 7, we also observed that some authors were unspecific about the evaluation metric used. For example, saying "kappa" instead of κ_u, κ_l, or κ_q (Bukai *et al.*, 2006; Leacock and Chodorow, 2003), and saying "correlation" instead of Pearson's r (Alfonseca and Pérez, 2004; Bukai *et al.*, 2006; Gütl, 2007). We contacted the authors about these, and recorded the correct metric in Table 8 to avoid the introduction of additional notation. It is clear that one must clearly define the variant that is used.

Considering all the evaluation metrics in Table 8 and what we know about ASAG, we consider continuous ratio data and the associated metrics as excessive in ASAG. That is, the questions we are dealing with are rarely more than 5-points in weight, and are also rarely graded with fractions of marks. Therefore, we suggest that these data are regarded or collapsed as discrete ratio data where possible and evaluated accordingly.

Finally, we remark that the HHA evaluation is frequently not given in the literature, hence the missing entries in Table 7. On many occasions, we suspect that only one human judge is used for rating the student answers, and mentioning this fact is simply neglected. A few other times, the number of judges is mentioned ("1-judge", "2-judge") without an evaluation of the consistency of the judge ratings.

Effectiveness

Finally, concerning the effectiveness scores in Table 7, the meaningful comparisons that can be performed are limited, as the majority of evaluations have been performed in a bubble. That is, the data sets that are common between two or more publications are relatively few. This means that many of the effectiveness scores in Table 7 only serve an informational purpose. We now highlight the other cases in Table 10 for discussion.

First in Table 10a, we refer to the "Texas" data set that hasn't been highlighted until this point. This data set is publicly available and very similar to that used in the second Mohler '09 publication (Mohler *et al.*, 2011). It is Ziai *et al.* (2012) that use this data to compare the effectiveness of two existing systems: The latest version of Mohler '09 (Meurers *et al.*, 2011b) and a regression-based version CoMiC-EN (Ziai *et al.*, 2012). Therefore, this publication is interesting in that the authors do not propose a new system, but instead give an empirical comparison of existing systems. We found no other paper like this for ASAG in our literature review. The results themselves favor the Mohler '09 approach over the CoMiC-EN

Table 10
Effectiveness comparisons for the Texas, CoMiC, and competition data sets. Refer to the historical analysis for the corresponding features, which are too numerous to repeat here.

ID	System	Reference	Grading model	Grading scheme	HSA.M	HSA.S
(a) Texas. HHA.M = r. HHA.S = 0.59.						
24a	Mohler '09	Meurers *et al.* (2011b)	ESA + LSA + knowledge-based measures	5-point	r	0.52
10a	CoMiC-EN	Ziai *et al.* (2012)	Support vector regression	5-point	r	0.41
(b) CREE. HHA.M = Agree. HHA.S = 1.00.						
8	CAM	Bailey and Meurers (2008)	k-nearest neighbor	2-way	acc	0.88
				5-way	acc	0.87
10	CoMiC-EN	Meurers *et al.* (2011b)	k-nearest neighbor	2-way	acc	0.88
				5-way	acc	0.79
(c) CREG. HHA.M = Agree. HHA.S = 1.00.						
9	CoMiC-DE	Meurers *et al.* (2011a)	k-nearest neighbor	2-way	acc	0.84
12	CoSeC-DE	Hahn and Meurers (2012)	LRS representation matching	2-way	acc	0.86
(d) ASAP '12 SAS. HHA.M = 2-judge. HHA.S = ??.						
29	Tandella '12	Tandalla (2012)	Linear regression	2-3-point	κ_q	0.74717
35	Zbontar '12	Zbontar (2012)	Stacking	2-3-point	κ_q	0.73892
11	Conort '12	Conort (2012)	Stacking	2-3-point	κ_q	0.73662
(e) SemEval '13 Task 7 Beetle. HHA.M = ??. HHA.S = ??.						
28	SoftCardinality	Jimenez *et al.* (2013)	Decision tree	4-way (UD)	F_1	0.375 (1st)
31	UKP-BIU	Zesch *et al.* (2013)	Naive Bayes	4-way (UD)	F_1	0.348 (2nd)
16	ETS	Heilman and Madnani (2013)	Stacking	4-way (UD)	F_1	0.339 (3rd)
13	Dzikovska '12	Dzikovska *et al.* (2012)	Decision tree	4-way (UD)	F_1	0.331
28	SoftCardinality	Jimenez *et al.* (2013)	Decision tree	4-way (UA)	F_1	0.474 (4th)
31	UKP-BIU	Zesch *et al.* (2013)	Naive Bayes	4-way (UA)	F_1	0.560 (2nd)
16	ETS	Heilman and Madnani (2013)	Stacking	4-way (UA)	F_1	0.581 (1st)
13	Dzikovska '12	Dzikovska *et al.* (2012)	Decision tree	4-way (UA)	F_1	0.375
(f) SemEval '13 Task 7 SciEntsBank. HHA.M = ??. HHA.S = ??.						
28	SoftCardinality	Jimenez *et al.* (2013)	Decision tree	5-way (UQ)	F_1	0.436 (3rd)
31	UKP-BIU	Zesch *et al.* (2013)	Naive Bayes	5-way (UQ)	F_1	0.285 (8th)
16	ETS	Heilman and Madnani (2013)	Stacking	5-way (UQ)	F_1	0.552 (1st)
13	Dzikovska '12	Dzikovska *et al.* (2012)	Decision tree	5-way (UQ)	F_1	0.414

approach by a large margin ($r = 0.52$ compared to $r = 0.41$). The Mohler '09 effectiveness is also quite reasonable as it is approaching the HHA.S bound ($r = 0.59$).

Next, Tables 10b and 10c give the comparisons for the CREE and CREG data sets from the CoMiC project. When considering CREE, the CoMiC-EN system as a successor to CAM is not more effective, but this can be explained by the fact that there is more emphasis on architecture for CoMiC-EN. When considering CREG, CoSeC-DE offers a marginal improvement over CoMiC-DE. In all cases, the accuracy is reasonable and is again approaching the HHA.S bound ($agr = 1.00$).

Some of the ASAP '12 SAS competition results follow next in Table 10d. The Tandella '12 system (Tandalla, 2012), Zbontar '12 system (Zbontar, 2012), and Conort '12 system (Conort, 2012) reviewed in this article achieved 2nd to 4th place from the original rankings. The corresponding κ_q scores are all very close, and the system by the top team was only +0.00077 higher again. The full ranking can be obtained from the leaderboard on the competition website.[13]

In general, the information about the ASAP '12 SAS systems is limited, as the intention is for

[13] `http://www.kaggle.com/c/asap-sas/leaderboard`

Table 11
Effectiveness comparisons for SemEval '13 Task 7 subtasks using macro-averaged F_1. A tick (✓) represents a best-performing system or one where the effectiveness is statistically insignificant from the best. The data is aggregated from the SemEval '13 Task 7 overview paper (Dzikovska *et al.*, 2013).

		2-way					3-way					5-way					
		Beetle		SciEntsBank			Beetle		SciEntsBank			Beetle		SciEntsBank			
ID	Run	UA	UQ	UA	UQ	UD	UA	UQ	UA	UQ	UD	UA	UQ	UA	UQ	UD	Count
	$CELI_1$																0
	$CNGL_2$	✓															1
	$CoMeT_1$	✓	✓	✓			✓		✓			✓		✓			7
	$EHUALM_2$		✓												✓		2
16	ETS_1	✓	✓							✓					✓		4
16	ETS_2	✓	✓	✓			✓	✓	✓			✓	✓	✓			9
	$LIMSIILES_1$														✓		1
28	$SoftCardinality_1$				✓	✓				✓	✓				✓	✓	6
31	$UKP\text{-}BIU_1$								✓	✓	✓			✓			4

the winning submissions to inform commercial partners. So for example, the competition organizers only released 5 methodology papers, which is little compared with the 51 teams that achieved a positive evaluation score. In addition, there are 20 teams that achieved $\kappa_q \geq 0.7$, a score considered excellent by Fleiss (2003). In comparison, there were only eight teams at the SemEval '13 Task 7 competition, so ASAP '12 SAS represents the largest participation in any single ASAG initiative.

A further problem with ASAP '12 SAS is that the data cannot be shared outside the context of the competition, limiting its usefulness in the academic community. Summing up, the ASAP '12 SAS competition results are somewhat limited for the broader ASAG community, given the restrictions on the data set and the small set of methodology papers.

Finally, the macro-averaged F_1 results from SemEval '13 Task 7 are given in Tables 10e and 10f. Overall, the full set of evaluation results is challenging to aggregate due to the dimensionality: multiple data sets (Beetle and SciEntsBank), multiple tasks (unseen answers, unseen questions, unseen domains) multiple grading schemes (2-way, 3-way, and 5-way), and multiple evaluation measures (accuracy, micro-averaged F_1, and macro-averaged F_1). As described when reviewing these systems, our approach to prioritizing this literature was to focus on the competition dimensions that are most difficult and novel. This gave us the ranking in the top half of Table 10e based on the overview paper (Dzikovska *et al.*, 2013). However, these rankings do not hold in comparison to the other sample rankings from Tables 10e, 10f, and in general. Indeed, even the Dzikovska '12 baseline (Dzikovska *et al.*, 2012) was a winner for one of the sub-tasks.

On the positive side, the comparisons for SemEval '13 Task 7 are the only with statistical significance tests in contrast to the rest of Table 10. The specific test used was an approximate randomization test with 10,000 iterations and a threshold of $p \leq 0.05$ (Dzikovska *et al.*, 2013). In analyzing this data, multiple effective solutions can be identified for each subtask comprising those with the top score for each subtask and those where the effectiveness was statistically insignificant from the top system. This visualization for the macro-averaged F_1 scores is given in Table 11.

The three systems we reviewed perform well overall based on this view, as does CoMeT (Ott *et al.*, 2013) as another system from the CoMiC project family. However overall, we are missing a single measure to indicate the overall top-performing approach. Indeed this is possible, as it is straight-forward to sum the number of correct classifications across all subtasks, for example. Reporting such a measure would be indicative of the system that performs most effectively over a large range of scenarios. This is a small but important agenda item for future work.

The other missing item for SemEval '13 Task 7 is a measure of HHA. A lot of work has been

done in the creation and running of this competition, so it is very interesting to know whether or not the submitted systems are approaching HHA effectiveness. The original annotation efforts actually have κ_u values (Dzikovska *et al.*, 2012), but these are not of use since the grading schemes were modified for use in the competition. Specifically, the original Beetle annotation effort was performed with 11 categories,[14] and the original SciEntsBank annotation effort was performed with 4 categories at the facet mapping level.[15] Both of these were then remapped to the 5-way scheme for the competition without new work for new measurement of HHA.

In considering overall effectiveness further, the data in Table 10 does not indicate many helpful crossovers between Tables 10a to 10f, apart from the common evaluation between Beetle and SciEntsBank. Perhaps the best example to date is the recent work of Ziai *et al.* (2012) providing a link across Tables 10a and 10b. Another example is the participation of ETS in both ASAP '12 SAS and SemEval '13 Task 7, excluding the fact that ETS chose not to publish their methodology at ASAP '12 SAS.

Effectiveness across Eras

Given that there are few comparisons that can be made across the subtables for Table 10, we also considered comparisons that can be made across eras. Here, we focus on literature where two (or more) eras are represented. Up until this point, our analyses only focused on the most effective method of each paper, and inferior methods were omitted. We identified three publications where cross-era comparisons are represented in the Wang '08, auto-marking, and Mohler '09 literature. Note that parts of the competitions could be considered as comparisons across eras, but we omit a dedicated section here since we have said much about the competitions already.

Wang '08 Wang *et al.* (2008) had the idea to compare two concept mapping methods and a machine learning method by grading all concepts either individually (concept mapping) or together (machine learning). The data comprises of four earth science questions for secondary education that requires the students to give an open number of responses and justifications. There are 2,698 student answers for comparison against the teacher answers. The task is to match as many concepts as possible between the student and teacher answers, and then predict the overall point total, which is calculated by summing the points assigned to each concept. The most effective concept mapping method ($r = 0.92$) was more effective than the machine learning method ($r = 0.86$).

auto-marking Two later publications about the auto-marking system (Pulman and Sukkarieh, 2005; Sukkarieh and Pulman, 2005) provide additional experiments for the information extraction and machine learning methods introduced in the initial work (Sukkarieh *et al.*, 2003, 2004). The data comprises of nine biology questions for secondary education, and the student answers for each question are assigned as 200 for training and 60 for testing. The task is to predict the scores for each question, which ranges between 1 and 4 points in value for full-credit. The results confirm the findings that the information extraction approach is more effective ($acc = 0.84$ versus $acc = 0.68$) (Pulman and Sukkarieh, 2005). When comparing information extraction methods with machine learning methods in general, the authors suggest that it is easier for the grades to be justified with information extraction, because links can be

[14] This comprises three top-level categories each with a number of subcategories: (1) metacognitive – positive and negative; (2) social – positive, negative, and neutral; and (3) content – correct, pc, pc_some_error, pc_some_missing, irrelevant, and incorrect. Acronym "pc" expands to "partially correct". Refer to Dzikovska *et al.* (2012, Section 2.1) for a full description of the categories and also the mapping process.

[15] The four facet mapping categories are "understood", "contradictory", "related", and "unaddressed". Refer to Dzikovska *et al.* (2012, Section 2.2) for a full description of the categories and also the mapping process. See also Levy *et al.* (2013) for other work on this annotation scheme and corpus.

traced back to the model. On the negative side, the information extraction method is more labor-intensive for model development because the patterns are manually engineered (Pulman and Sukkarieh, 2005).

Mohler '09 The idea of the original work is an unsupervised model where the score from a single knowledge- or corpus-based method is taken for grading (Mohler and Mihalcea, 2009). In the follow-on publication, Mohler *et al.* (2011) introduce graph alignment features from dependency graphs to the existing pool of features. This time, all features are combined with a support vector machine, hence implementing a supervised model with machine learning. Overall, the goal of the Mohler *et al.* (2011) was to perform a comparison of the unsupervised and supervised models. To do this, a data set was created comprising of eight computer science questions on data structures for university education. The training/testing split of this data depends on the model: For the unsupervised model, the training portion is the teacher answers only, and for the supervised model, the split is based on a 12-fold cross validation design. The task for the Mohler *et al.* (2011) study is simply to predict the score for each response, and all questions were marked on a 5-point scale. The experiments compare the unsupervised and supervised models, and their effectiveness is evaluated with the population Pearson correlation coefficient, also known as Pearson's ρ. Mohler *et al.* (2011) find that the supervised model (SVMRank, $\rho = 0.518$) is more effective than the unsupervised model (Lesk, $\rho = 0.450$).

General Remarks In summary, we feel that the Wang '08 result emphasizes concept mapping methods as a special case that benefits from the precision that can be obtained in fine-grained as opposed to holistic marking. So there is benefit in employing this type of method when possible, assuming the assessment design supports concept mapping instead of holistic marking. Concerning the other results, and the cross-era comparisons for SemEval '13 Task 7, we feel that they support the grading continuum represented by Figure 7 above. Here, the auto-marking system has been refined over a series of four publications and multiple data sets, so we consider this a type of repeated assessment where rule-based methods perform well. In contrast, the Mohler '09 system and the SemEval '13 Task 7 competition scenario do not have the same level of repetition and SemEval '13 Task 7 specifically targets unseen questions and domains. Given these constraints, we consider Mohler '09 and the SemEval '13 Task 7 competition scenario as representing the other side of the continuum where statistical methods perform well.

LESSONS LEARNED

Research in automatic assessment of natural language questions has moved quickly since the turn of the millennium, and ASAG is no exception. The short answer question type is one of many types of questions requiring deep understanding of material to recall knowledge for free expression, making it a challenging task to grade automatically. In contrast to essay questions, we have defined short answers as typically not exceeding one paragraph in length, they focus on content as opposed to style, and they can be described as objective or close-ended as opposed to subjective or open-ended. This has defined a unique field, from which we have identified over 80 papers that fit the definition dating back to 1996.

Our historical analysis of ASAG indicates five eras in how the research has developed since 1996. Here, we refer to the eras of concept mapping, information extraction, corpus-based methods, machine learning, and evaluation. We found that the trend in the earlier eras is more towards rule-based methods, which either grade answers in parts with concept mapping techniques, or holistically with information extraction techniques. Later, we found that the trend shifted more towards statistical methods, whereby the features are generated with the assistance of corpus-based methods, or NLP methods used as part of a machine learning system. Lastly, we found that the most recent trend is towards evaluation, where com-

petitions and publicly available data sets are finally allowing meaningful comparisons between methods.

We draw many other conclusions when generalizing over all systems simultaneously considering common components comprising data sets, natural language processing, model building, grading models, model evaluation, and effectiveness. Our conclusions for these six dimensions are as follows.

Data Sets We find that most data sets cannot be shared for reasons such as privacy. Frequently, academics are simply adapting data from their own teaching experiences to ASAG projects, but with little consideration that others may want to perform meaningful comparisons to their methodology. We find that the open data sets are the Texas data set (Mohler *et al.*, 2011), and the data sets from the CoMiC project and the SemEval '13 Task 7 competition. In addition, the data is currently represented by four languages, mostly comes from university or school assessments, and can belong to nearly any year level or topic.

Natural Language Processing We find 17 linguistic processing techniques that can serve as a checklist for others. In particular, we created a taxonomy to group these by themes for techniques that are lexical, morphological, semantic, syntactic, or surface for further guidance. For feature extraction, many of the features used fall within these five themes too. Other features are based on n-grams, information retrieval, machine translation, textual similarity, overlaps, entailment, and entropy.

Model Building We observe that the teacher answers play very different roles for model building across systems. Sometimes, the teacher answer effectively *is* the model. For other systems, it is only used to guide a manual marking process of student answers that are themselves used as the model. Other times, the student and teacher answers are combined together to build a model.

Grading Models We describe how the concept mapping and information extraction ASAG methods can more broadly be described as rule-based, and that the corpus-based and machine learning ASAG methods can more broadly be described as statistical. In addition, almost all systems originally marked under the era of evaluation are machine learning systems, suggesting this preference amongst the latest research. We also argue that statistical methods make sense for unseen questions and domains, and that rule-based methods make sense for repeated assessment.

Model Evaluation We find a mixture of nominal, ordinal, and ratio data in evaluation, and questions rarely with more than five categories or worth more than five points. The evaluation metrics themselves are most commonly represented by accuracy, agreement, different variants of kappa, and Pearson correlation. Common mistakes are to neglect reporting inter-rater agreement between human raters, and to omit detail about the variant of an evaluation metric used such as a kappa or correlation.

Effectiveness We observe that meaningful effectiveness comparisons are available for system evaluations with six public or common data sets. The largest of these (SemEval '13 Task 7) would benefit from measures of human-human agreement and overall system effectiveness to advance the research. In comparison, just two bodies of work (Ziai *et al.* (2012) and ETS in ASAP '12 SAS and SemEval '13 Task 7) have comparisons *between* public data sets, so there is scope to consolidate the existing knowledge. Finally, results across eras indicate concept mapping methods as more effective than holistic grading and other methods as influenced by the trade-off between repeated/non-repeated assessment and seen/unseen questions and domains.

Looking Forward

For the future, we see the era of evaluation as having the biggest influence, since this body of work is emerging. Given the corresponding resources, we anticipate more researchers will reuse the publicly available data sets as opposed to using new data from restricted internal sources. We also speculate that there will be more participation from commercial groups above and beyond submitting runs to competitions and publishing stand-alone papers, to keep up with the new openness in the field. A recent alternative is the Open edX project as an open source platform with significant ASAG components.[16] Regardless of forthcoming collaboration, it may be possible to advance evaluation efforts with the question-answering research community as an alternative or addition. For example, the TREC (Text REtrieval Conference) forum[17] has included question-answering evaluation tracks from 1999 to 2007.[18] Hirschman *et al.* (2000) noted this interesting link many years ago, and we wonder when this idea will be realized.

ACKNOWLEDGEMENTS

We thank Susanne Neumann and Margot Mieskes for informal discussions that were helpful in developing some parts of this article. We additionally thank Kostadin Cholakov and Torsten Zesch for providing feedback on drafts. We also thank the anonymous reviewers for their generous and helpful comments.

REFERENCES

Aleven, V., Ogan, A., Popescu, O., Torrey, C., and Koedinger, K. (2004). Evaluating the Effectiveness of a Tutorial Dialogue System for Self-Explanation. In J. C. Lester, R. M. Vicari, and F. Paraguacu, editors, *Proceedings of the Seventh International Conference on Intelligent Tutoring Systems*, volume 3220 of *Lecture Notes in Computer Science*, pages 443–454, Maceio, Brazil. Springer.

Alfonseca, E. and Pérez, D. (2004). Automatic Assessment of Open Ended Questions with a BLEU-Inspired Algorithm and Shallow NLP. In J. Vicedo, P. Martínez-Barco, R. Muńoz, and M. Saiz Noeda, editors, *Advances in Natural Language Processing*, volume 3230 of *Lecture Notes in Computer Science*, pages 25–35. Springer, Berlin, Germany.

Alfonseca, E., Carro, R. M., Freire, M., Ortigosa, A., Pérez, D., and Rodriguez, P. (2005). Authoring of Adaptive Computer Assisted Assessment of Free-Text Answers. *Educational Technology & Society*, **8**(3), 53–65.

Attali, Y. and Burstein, J. (2006). Automated Essay Scoring with e-rater V.2. *The Journal of Technology, Learning, and Assessment*, **4**(3), 1–31.

Attali, Y., Powers, D., Freedman, M., Harrison, M., and Obetz, S. (2008). Automated Scoring of Short-Answer Open-Ended GRE Subject Test Items. Technical Report RR-08-20, Educational Testing Service, Princeton, New Jersey.

Bachman, L. F., Carr, N., Kamei, G., Kim, M., Pan, M. J., Salvador, C., and Sawaki, Y. (2002). A Reliable Approach to Automatic Assessment of Short Answer Free Responses. In S.-C. Tseng, T.-E. Chen, and Y.-F. Liu, editors, *Proceedings of the Nineteenth International Conference on Computational Linguistics*, volume 2 of *COLING '02*, pages 1–4, Taipei, Taiwan. Association for Computational Linguistics.

[16] `http://code.edx.org/`
[17] `http://trec.nist.gov`
[18] `http://trec.nist.gov/data/qamain.html`

Bailey, S. (2008). *Content Assessment in Intelligent Computer-Aided Language Learning: Meaning Error Diagnosis for English as a Second Language*. Ph.D. thesis, Ohio State University, Columbus, Ohio.

Bailey, S. and Meurers, D. (2008). Diagnosing Meaning Errors in Short Answers to Reading Comprehension Questions. In J. Tetreault, J. Burstein, and R. De Felice, editors, *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 107–115, Columbus, Ohio. Association for Computational Linguistics.

Bakeman, R., McArthur, D., Quera, V., and Robinson, B. F. (1997). Detecting Sequential Patterns and Determining their Reliability with Fallible Observers. *Psychological Methods*, **4**(4), 357–370.

Bär, D., Zesch, T., and Gurevych, I. (2011). A Reflective View on Text Similarity. In R. Mitkov and G. Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 515–520, Hissar, Bulgaria. Association for Computational Linguistics.

Bär, D., Zesch, T., and Gurevych, I. (2012a). Text Reuse Detection Using a Composition of Text Similarity Measures. In P. Bhattacharrya, R. Sangal, M. Kay, and C. Boitet, editors, *Proceedings of the Twenty-Fourth International Conference on Computational Linguistics*, volume 1 of *COLING '12*, pages 167–184, Mumbai, India. Indian Institute of Technology Bombay.

Bär, D., Biemann, C., Gurevych, I., and Zesch, T. (2012b). UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. In S. Manandhar and D. Yuret, editors, *Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 435–440, Montreal, Canada. Association for Computational Linguistics.

Bär, D., Zesch, T., and Gurevych, I. (2013). DKPro Similarity: An Open Source Framework for Text Similarity. In H. Schuetze, P. Fung, and M. Poesio, editors, *Proceedings of the Fifty-First Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 121–126, Sofia, Bulgaria. Association for Computational Linguistics.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., and Szpektor, I. (2006). The Second PASCAL Recognising Textual Entailment Challenge. In B. Magnini and I. Dagan, editors, *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 1–9, Venice, Italy.

Bejar, I. I. (2011). A Validity-Based Approach to Quality Control and Assurance of Automated Scoring. *Assessment in Education: Principles, Policy & Practice*, **18**(3), 319–341.

Bennett, R. E. (2011). Automated Scoring of Constructed-Response Literacy and Mathematics Items. White paper, Educational Testing Service, Princeton, New Jersey.

Bentivogli, L., Dagan, I., Dang, H. T., Giampiccolo, D., and Magnini, B. (2009). The Fifth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Second Text Analysis Conference*, pages 1–15, Gaithersburg, Maryland. National Institute of Standards and Technology.

Bentivogli, L., Clark, P., Dagan, I., and Giampiccolo, D. (2010). The Sixth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Third Text Analysis Conference*, pages 1–18, Gaithersburg, Maryland. National Institute of Standards and Technology.

Bentivogli, L., Clark, P., Dagan, I., and Giampiccolo, D. (2011). The Seventh PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Fourth Text Analysis Conference*, pages 1–16, Gaithersburg, Maryland. National Institute of Standards and Technology.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, **24**(2), 123–140.

Bukai, O., Pokorny, R., and Haynes, J. (2006). An Automated Short-Free-Text Scoring System: Development and Assessment. In *Proceedings of the Twentieth Interservice/Industry Training, Simulation, and Education Conference*, pages 1–11. National Training and Simulation Association.

Burrows, S. and D'Souza, D. (2005). Management of Teaching in a Complex Setting. In J. Hurst and J. Sheard, editors, *Proceedings of the Second Melbourne Computing Education Conventicle*, pages 1–8, Melbourne, Australia.

Burrows, S., Tahaghoghi, S. M. M., and Zobel, J. (2007). Efficient Plagiarism Detection for Large Code Repositories. *Software: Practice and Experience*, **37**(2), 151–175.

Burrows, S., Potthast, M., and Stein, B. (2013). Paraphrase Acquisition via Crowdsourcing and Machine Learning. *ACM Transactions on Intelligent Systems and Technology*, **4**(3), 43:1–43:21.

Burrows, S., Uitdenbogerd, A. L., and Turpin, A. (2014). Comparing Techniques for Authorship Attribution of Source Code. *Software: Practice and Experience*, **44**(1), 1–32.

Burstein, J., Kaplan, R., Wolff, S., and Lu, C. (1996). Using Lexical Semantic Techniques to Classify Free-Responses. In E. Viegas, editor, *Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, pages 20–29, Santa Cruz, California. Association for Computational Linguistics.

Butcher, P. G. and Jordan, S. E. (2010). A Comparison of Human and Computer Marking of Short Free-Text Student Responses. *Computers & Education*, **55**(2), 489–499.

Callear, D., Jerrams-Smith, J., and Soh, V. (2001). CAA of Short Non-MCQ Answers. In M. Danson and C. Eabry, editors, *Proceedings of the Fifth Computer Assisted Assessment Conference*, pages 1–14, Loughborough, United Kingdom. Loughborough University.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Cohen, J. (1968). Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement of Partial Credit. *Psychological Bulletin*, **70**(4), 213–220.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, **112**(1), 155–159.

Conole, G. and Warburton, B. (2005). A Review of Computer-Assisted Assessment. *Journal of the Association for Learning Technology*, **13**(1), 17–31.

Conort, X. (2012). Short Answer Scoring – Explanation of "Gxav" Solution. ASAP '12 SAS methodology paper, Gear Analytics.

Cowie, J. and Wilks, Y. (2000). Information Extraction. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 10, pages 241–260. Marcel Dekker, New York City, New York, first edition.

Csink, L., György, A., Raincsák, Z., Schmuck, B., Sima, D., Sziklai, Z., and Szöllösi, S. (2003). Intelligent Assessment Systems for e-Learning. In F. Udrescu, I. G. Rosca, G. M. Sandulescu, and R. Stroe, editors, *Proceedings of the Fourth European Conference on E-Commerce, E-Learning, E-Business, E-Work, E-Health, E-Banking, E-Democracy, E-Government, BB & On-Line Services Applications, New Working Environments, Virtual Institutes, and their Influences on the Economic and Social Environment*, E-COMM-LINE, pages 224–229, Bucharest, Romania. R&D Institute for Automation Bucharest and Academy of Economic Studies Bucharest.

Cutrone, L. and Chang, M. (2010). Automarking: Automatic Assessment of Open Questions. In M. Jemni, D. Sampson, Kinshuk, and J. M. Spector, editors, *Proceedings of the Tenth IEEE International Conference on Advanced Learning Technologies*, pages 143–147, Sousse, Tunisia. IEEE.

Cutrone, L., Chang, M., and Kinshuk (2011). Auto-Assessor: Computerized Assessment System for Marking Student's Short-Answers Automatically. In N. S. Narayanaswamy, M. S. Krishnan, Kinshuk, and R. Srinivasan, editors, *Proceedings of the Third IEEE International Conference on Technology for Education*, pages 81–88, Chennai, India. IEEE.

Dagan, I., Glickman, O., Gan, R., and Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, editors, *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Dorr, B., Hendler, J., Blanksteen, S., and Migdaloff, B. (1995). On Beyond Syntax: Use of Lexical Conceptual Structure for Intelligent Tutoring. In V. M. Holland, J. Kaplan, and M. Sams, editors, *Intelligent Language Tutors*, pages 289–311. Lawrence Erlbaum Publishers, Mahwah, New Jersey, first edition.

Dzikovska, M. O., Nielsen, R. D., and Brew, C. (2012). Towards Effective Tutorial Feedback for Explanation Questions: A Dataset and Baselines. In J. Chu-Carroll, E. Fosler-Lussier, E. Riloff, and S. Bangalore, editors, *Proceedings of the Twelfth Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 200–210, Montreal, Canada. Association for Computational Linguistics.

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., and Dang, H. T. (2013). SemEval-2013 Task 7: The Joint Student Response Analysis and Eighth Recognizing Textual Entailment Challenge. In M. Diab, T. Baldwin, and M. Baroni, editors, *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, pages 1–12, Atlanta, Georgia.

Evens, M. W., Brandle, S., Chang, R.-C., Freedman, R., Glass, M., Lee, Y. H., Shim, L. S., Woo, C. W., Zhang, Y., Zhou, Y., Michael, J. A., and Rovick, A. A. (2001). CIRCSIM-Tutor: An Intelligent Tutoring System Using Natural Language Dialogue. In *Proceedings of the Twelfth Midwest Artificial Intelligence and Cognitive Science Conference*, pages 16–23, Oxford, Ohio.

Fleiss, J. L. (2003). The Measurement of Interrater Agreement. In J. L. Fleiss, B. Levin, and M. C. Paik, editors, *Statistical Methods for Rates and Proportions*, chapter 18, pages 598–626. John Wiley & Sons, third edition.

Gabrilovich, E. and Markovitch, S. (2006). Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In A. Cohn, editor, *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, volume 2, pages 1301–1306, Boston, Massachusetts. AAAI Press.

Gay, L. R. (1980). The Comparative Effects of Multiple-Choice versus Short-Answer Tests on Retention. *Journal of Educational Measurement*, **17**(1), 45–50.

Giampiccolo, D., Magnini, B., Sommarive, V., Gan, R., and Dolan, B. (2007). The Third PASCAL Recognizing Textual Entailment Challenge. In S. Sekine, K. Inui, I. Dagan, B. Dolan, D. Giampiccolo, and B. Magnini, editors, *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, Czech Republic. Association for Computational Linguistics.

Giampiccolo, D., Dang, H. T., Magnini, B., Dagan, I., Cabrio, E., and Dolan, B. (2008). The Fourth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the First Text Analysis Conference*, pages 1–11, Gaithersburg, Maryland. National Institute of Standards and Technology.

Gollub, T., Burrows, S., and Stein, B. (2012a). First Experiences with TIRA for Reproducible Evaluation in Information Retrieval. In A. Trotman, C. L. A. Clarke, I. Ounis, J. S. Culpepper, M.-A. Cartright, and S. Geva, editors, *Proceedings of the First SIGIR Workshop on Open Source Information Retrieval*, pages 52–55, Portland, Oregon. University of Otago.

Gollub, T., Stein, B., and Burrows, S. (2012b). Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In B. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *Proceedings of the Thirty-Fifth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1125–1126, Portland, Oregon. ACM.

Gollub, T., Stein, B., Burrows, S., and Hoppe, D. (2012c). TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In A. M. Tjoa, S. Liddle, K.-D. Schewe, and X. Zhou, editors, *Proceedings of the Ninth International Workshop on Text-based Information Retrieval at DEXA*, pages 151–155, Vienna, Austria. IEEE.

Gonzalez-Barbone, V. and Llamas-Nistal, M. (2008). eAssessment of Open Questions: An Educator's Perspective. In C. Traver, M. Ohland, J. Prey, and T. Mitchell, editors, *Proceedings of the Thirty-Eighth Annual Frontiers in Education Conference*, pages F2B–1–F2B–6, Saratoga Springs, New York. IEEE.

Graesser, A. C., Chipman, P., Haynes, B. C., and Olney, A. (2005). AutoTutor: An Intelligent Tutoring System With Mixed-Initiative Dialogue. *IEEE Transactions on Education*, **48**(4), 612–618.

Gütl, C. (2007). e-Examiner: Towards a Fully-Automatic Knowledge Assessment Tool Applicable in Adaptive E-Learning Systems. In P. H. Ghassib, editor, *Proceedings of the Second International Conference on Interactive Mobile and Computer Aided Learning*, pages 1–10, Amman, Jordan.

Gütl, C. (2008). Moving Towards a Fully Automatic Knowledge Assessment Tool. *International Journal of Emerging Technologies in Learning*, **3**(1), 1–11.

György, A. and Vajda, I. (2007). Intelligent Mathematics Assessment in eMax. In *Proceedings of the Eighth Africon Conference*, pages 1–6, Windhoek, South Africa. IEEE.

Hahn, M. and Meurers, D. (2012). Evaluating the Meaning of Answers to Reading Comprehension Questions: A Semantics-Based Approach. In J. Tetreault, J. Burstein, and C. Leacock, editors, *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 326–336, Montreal, Canada. Association for Computational Linguistics.

Haley, D. T., Thomas, P., Roeck, A. D., and Petre, M. (2007). Measuring Improvement in Latent Semantic Analysis-Based Marking Systems: Using a Computer to Mark Questions about HTML. In S. Mann and Simon, editors, *Proceedings of the Ninth Australasian Conference on Computing Education*, volume 66 of *ACE*, pages 35–42, Ballarat, Australia. Australian Computer Society.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, **11**(1), 10–18.

Hamp, B. and Feldweg, H. (1997). GermaNet—A Lexical-Semantic Net for German. In P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo, and Y. Wilks, editors, *Proceedings of the First ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–16, Madrid, Spain. Association for Computational Linguistics.

Hayashi, F. (2000). Finite-Sample Properties of OLS. In *Econometrics*, chapter 1, pages 3–87. Princeton University Press, Princeton, New Jersey.

Heilman, M. and Madnani, N. (2013). ETS: Domain Adaptation and Stacking for Short Answer Scoring. In M. Diab, T. Baldwin, and M. Baroni, editors, *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 2, pages 275–279, Atlanta, Georgia.

Hewlett Foundation (2012). Automated Student Assessment Prize: Phase Two – Short Answer Scoring. Kaggle Competition.

Hirschman, L., Breck, E., Light, M., Burger, J. D., and Ferro, L. (2000). Automated Grading of Short-Answer Tests. *Intelligent Systems and their Applications*, **15**(5), 31–35.

Hirst, G. and St-Onge, D. (1998). Lexical Chains as Representations of Contexts for the Detection and Correction of Malapropisms. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, Language, Speech, and Communication, chapter 13, pages 305–332. MIT Press.

Horbach, A., Palmer, A., and Pinkal, M. (2013). Using the Text to Evaluate Short Answers for Reading Comprehension Exercises. In M. Diab, T. Baldwin, and M. Baroni, editors, *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 286–295, Atlanta, Georgia. Association for Computational Linguistics.

Hou, W.-J. and Tsao, J.-H. (2011). Automatic Assessment of Students' Free-Text Answers With Different Levels. *International Journal on Artificial Intelligence Tools*, **20**(2), 327–347.

Hou, W.-J., Tsao, J.-H., Li, S.-Y., and Chen, L. (2010). Automatic Assessment of Students' Free-Text Answers with Support Vector Machines. In M. Ali, C. Fyfe, N. García-Pedrajas, and F. Herrera, editors, *Proceedings of the Twenty-Third International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, volume 1, pages 235–243, Cordoba, Spain. Springer.

Hou, W.-J., Tsao, J.-H., Lu, C.-S., and Chen, L.-C. (2011). Free-Text Assessment of Students' Answers with Different Feature Selections. In K. Tang and H. Koguchi, editors, *Proceedings of the Fifth International Conference on e-Commerce, e-Administration, e-Society, e-Education, and e-Technology*, pages 2489–2510, Tokyo, Japan. Knowledge Association of Taiwan and International Business Academics Consortium.

Hou, W.-J., Lu, C.-S., Chang, C.-P., and Chen, H.-Y. (2012). Learning Diagnosis for Students' Free-Text Answers with Feature-Based Approaches. In *Proceedings of the First International Conference on Information and Computer Applications*, volume 24, pages 42–47, Hong Kong, China. International Association of Computer Science and Information Technology.

Intelligent Assessment Technologies (2009). E-Assessment of Short-Answer Questions. White paper, Coatbridge, United Kingdom.

Jiang, J. J. and Conrath, D. W. (1997). Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In L.-S. Lee, K.-J. Chen, C.-R. Huang, and R. Sproat, editors, *Proceedings of the Tenth International Conference on Research in Computational Linguistics*, pages 1–15, Taipei, Taiwan.

Jimenez, S., Becerra, C., Universitaria, C., and Gelbukh, A. (2013). SOFTCARDINALITY: Hierarchical Text Overlap for Student Response Analysis. In M. Diab, T. Baldwin, and M. Baroni, editors, *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 2, pages 280–284, Atlanta, Georgia.

Jordan, S. (2007). Computer Based Assessment with Short Free Responses and Tailored Feedback. In P. Chin, K. Clark, S. Doyle, P. Goodhew, T. Madden, S. Meskin, T. Overton, and J. Wilson, editors, *Proceedings of the Second Science Learning and Teaching Conference*, pages 158–163, Keele, United Kingdom.

Jordan, S. (2008). Online Interactive Assessment with Short Free-Text Questions and Tailored Feedback. *New Directions*, **4**, 17–20.

Jordan, S. (2009a). Assessment for Learning: Pushing the Boundaries of Computer-Based Assessment. *Practitioner Research in Higher Education*, **3**(1), 11–19.

Jordan, S. (2009b). Investigating the Use of Short Free Text Questions in Online Assessment. Final project report, Centre for the Open Learning of Mathematics, Science, Computing and Technology, The Open University, Milton Keynes, United Kingdom.

Jordan, S. (2012a). Short-Answer E-Assessment Questions: Five Years On. In D. Whitelock, G. Wills, and B. Warburton, editors, *Proceedings of the Fifteenth International Computer Assisted Assessment Conference*, pages 1–12, Southampton, United Kingdom.

Jordan, S. (2012b). Student Engagement with Assessment and Feedback: Some Lessons from Short-Answer Free-Text E-Assessment Questions. *Computers & Education*, **58**(2), 818–834.

Jordan, S. and Mitchell, T. (2009). e-Assessment for Learning? The Potential of Short-Answer Free-Text Questions with Tailored Feedback. *British Journal of Educational Technology*, **40**(2), 371–385.

Jordan, S., Brockbank, B., and Butcher, P. (2007). Extending the Pedagogic Role of Online Interactive Assessment: Providing Feedback on Short Free-Text Responses. In *Proceedings of the International Online Conference on Assessment Design for Learner Responsibility*, pages 1–6.

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, **30**(1–2), 81–93.

Klein, R., Kyrilov, A., and Tokman, M. (2011). Automated Assessment of Short Free-Text Responses in Computer Science using Latent Semantic Analysis. In G. Rößling, T. Naps, and C. Spannagel, editors, *Proceedings of the Sixteenth Annual Joint Conference on Innovation and Technology in Computer Science Education*, pages 158–162, Darmstadt, Germany. ACM.

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice*, **41**(4), 212–219.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, **25**(2-3), 259–284.

Leacock, C. and Chodorow, M. (1998). Combining Local Context and WordNet Sense Similarity for Word Sense Identification. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, Language, Speech, and Communication, chapter 11, pages 265–284. MIT Press.

Leacock, C. and Chodorow, M. (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities*, **37**(4), 389–405.

Lesk, M. (1986). Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a Pine Cone from an Ice Cream Cone. In V. D. Buys, editor, *Proceedings of the Fifth Annual International Conference on Systems Documentation*, pages 24–26, Toronto, Canada. ACM.

Levy, O., Zesch, T., Dagan, I., and Gurevych, I. (2013). Recognizing Partial Textual Entailment. In H. Schuetze, P. Fung, and M. Poesio, editors, *Proceedings of the Fifty-First Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 451–455, Sofia, Bulgaria. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In M.-F. Moens and S. Szpakowicz, editors, *Proceedings of the First Text Summarization Branches Out Workshop at ACL*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In J. W. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, Madison, Wisconsin. Morgan Kaufmann Publishers.

Madnani, N., Burstein, J., Sabatini, J., and Reilly, T. O. (2013). Automated Scoring of a Summary Writing Task Designed to Measure Reading Comprehension. In J. Tetreault, J. Burstein, and C. Leacock, editors, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–168, Atlanta, Georgia. Association for Computational Linguistics.

Marquardt, D. W. and Snee, R. D. (1975). Ridge Regression in Practice. *The American Statistician*, **29**(1), 3–20.

Martinez, M. E. and Bennett, R. E. (1992). A Review of Automatically Scorable Constructed-Response Item Types for Large-Scale Assessment. *Applied Measurement in Education*, **5**(2), 151–169.

Meurers, D., Ott, N., and Ziai, R. (2010). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In O. Bott, S. Featherston, I. Steiner, B. Stolterfoht, and Y. Versley, editors, *Proceedings of the Fourth Linguistic Evidence Conference*, pages 214–217, Tübingen, Germany.

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (2011a). Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In P. Clark, I. Dagan, K. Erk, S. Pado, S. Thater, and F. M. Zanzotto, editors, *Proceedings of the Second TextInfer Workshop on Textual Entailment*, pages 1–9, Edinburgh, United Kingdom. Association for Computational Linguistics.

Meurers, D., Ziai, R., Ott, N., and Bailey, S. M. (2011b). Integrating Parallel Analysis Modules to Evaluate the Meaning of Answers to Reading Comprehension Questions. *International Journal of Continuing Engineering Education and Life-Long Learning*, **21**(4), 355–369.

Mitchell, T., Russell, T., Broomhead, P., and Aldridge, N. (2002). Towards Robust Computerised Marking of Free-Text Responses. In *Proceedings of the Sixth Computer Assisted Assessment Conference*, pages 233–249, Loughborough, United Kingdom.

Mitchell, T., Aldridge, N., Williamson, W., and Broomhead, P. (2003a). Computer Based Testing of Medical Knowledge. In *Proceedings of the Seventh Computer Assisted Assessment Conference*, pages 249–267, Loughborough, United Kingdom.

Mitchell, T., Aldridge, N., and Broomhead, P. (2003b). Computerised Marking of Short-Answer Free-Text Responses. In *Proceedings of the Twenty-Ninth Annual Conference of the International Association for Educational Assessment*, pages 1–16, Manchester, United Kingdom.

Mohler, M. and Mihalcea, R. (2009). Text-to-text Semantic Similarity for Automatic Short Answer Grading. In A. Lascarides, C. Gardent, and J. Nivre, editors, *Proceedings of the Twelfth Conference of the European Chapter of the Association for Computational Linguistics*, pages 567–575, Athens, Greece. Association for Computational Linguistics.

Mohler, M., Bunescu, R., and Mihalcea, R. (2011). Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In D. Lin, editor, *Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1 of *HLT '11*, pages 752–762, Portland, Oregon. Association for Computational Linguistics.

Moser, J. R. (2009). *The Electronic Assessor: Design and Prototype of an Automated Free-Text Assessment System*. Master's thesis, Institute for Information Systems and Computer Media, Technical University of Graz, Graz, Austria.

Nielsen, R. D., Ward, W., Martin, J. H., and Palmer, M. (2008a). Annotating Students' Understanding of Science Concepts. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1–8, Marrakech, Morocco. European Language Resources Association.

Nielsen, R. D., Ward, W., and Martin, J. H. (2008b). Learning to Assess Low-Level Conceptual Understanding. In D. Wilson and H. C. Lane, editors, *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference*, pages 427–432, Coconut Grove, Florida. AAAI Press.

Nielsen, R. D., Ward, W., and Martin, J. H. (2009). Recognizing Entailment in Intelligent Tutoring Systems. *Natural Language Engineering*, **15**(4), 479–501.

Ott, N., Ziai, R., and Meurers, D. (2012). Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context. In T. Schmidt and K. Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, volume 14 of *Hamburg Studies on Multilingualism*, pages 47–69. John Benjamins Publishing, Amsterdam, Netherlands.

Ott, N., Ziai, R., Hahn, M., and Meurers, D. (2013). CoMeT: Integrating Different Levels of Linguistic Modeling for Meaning Assessment. In S. Manandhar and D. Yuret, editors, *Proceedings of the Seventh International Workshop on Semantic Evaluation*, pages 608–616, Atlanta, Georgia. Association for Computational Linguistics.

Page, E. B. (1966). The Imminence of Grading Essays by Computer. *Phi Delta Kappan*, **47**(5), 238–243.

Papadimitriou, C. H., Tamaki, H., Raghavan, P., and Vempala, S. (1998). Latent Semantic Indexing: A Probabilistic Analysis. In A. Mendelson and J. Paredaens, editors, *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '98, pages 159–168, Seattle, Washington. ACM.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, editor, *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Pascual-Nieto, I., Perez-Marin, D., O'Donnell, M., and Rodriguez, P. (2008). Enhancing a Free-Text Adaptive Computer Assisted Assessment System with Self-Assessment Features. In P. Díaz, Kinshuk, I. Aedo, and E. Mora, editors, *Proceedings of the Eighth IEEE International Conference on Advanced Learning Technologies*, pages 399–401, Santander, Spain. IEEE.

Pascual-Nieto, I., Santos, O. C., Perez-Marin, D., and Boticario, J. G. (2011). Extending Computer Assisted Assessment Systems with Natural Language Processing, User Modeling and Recommendations Based on Human Computer Interaction and Data Mining. In T. Walsh, editor, *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, volume 3 of *IJCAI '11*, pages 2519–2524, Barcelona, Spain. AAAI Press.

Pearson Education (2010). Intelligent Essay Assessor (IEA) Fact Sheet. Online Brochure. `http://kt.pearsonassessments.com/download/IEA-FactSheet-20100401.pdf`.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::Similarity: Measuring the Relatedness of Concepts. In D. Palmer, J. Polifroni, and D. Roy, editors, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (Demonstration Papers)*, pages 38–41, Boston, Massachusetts. Association for Computational Linguistics.

Pérez, D. and Alfonseca, E. (2005). Adapting the Automatic Assessment of Free-Text Answers to the Students. In *Proceedings of the Ninth Computer Assisted Assessment Conference*, pages 1–12, Loughborough, United Kingdom.

Pérez, D., Alfonseca, E., and Rodríguez, P. (2004a). Application of the BLEU Method for Evaluating Free-Text Answers in an E-Learning Environment. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 1351–1354, Lisbon, Portugal.

Pérez, D., Alfonseca, E., and Rodríguez, P. (2004b). Upper Bounds of the BLEU Algorithm Applied to Assessing Student Essays. In *Proceedings of the Thirtieth International Association for Educational Assessment Conference*, Philadelphia, Pennsylvania.

Pérez, D., Alfonseca, E., Rodríguez, P., Gliozzo, A., Strapparava, C., and Magnini, B. (2005a). About the Effects of Combining Latent Semantic Analysis with Natural Language Processing Techniques for Free-Text Assessment. *Revista Signos: Estudios de Lingüística*, **38**(59), 325–343.

Pérez, D., Postolache, O., Alfonseca, E., Cristea, D., and Rodríguez, P. (2005b). About the Effects of using Anaphora Resolution in Assessing Free-Text Student Answers. In R. Mitkov, editor, *Proceedings of the Eleventh International Conference on Recent Advances in Natural Language Processing*, pages 380–386, Borovets, Bulgaria.

Pérez, D., Gliozzo, A. M., Strapparava, C., Alfonseca, E., Rodríguez, P., and Magnini, B. (2005c). Automatic Assessment of Students' Free-Text Answers Underpinned by the Combination of a BLEU-Inspired Algorithm and Latent Semantic Analysis. In D. Cook, L. Holder, I. Russell, and Z. Markov, editors, *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference*, pages 358–363, Clearwater Beach, Florida. AAAI Press.

Pérez-Marín, D. (2004). *Automatic Evaluation of User's Short Essays by Using Statistical and Shallow Natural Language Processing Techniques*. Diploma thesis, Computer Science Department, Universidad Autónoma of Madrid, Madrid, Spain.

Pérez-Marín, D. (2007). *Adaptive Computer Assisted Assessment of Free-Text Students' Answers: An Approach to Automatically Generate Students' Conceptual Models*. Ph.D. thesis, Computer Science Department, Universidad Autónoma of Madrid, Madrid, Spain.

Pérez-Marín, D. and Pascual-Nieto, I. (2011). Willow: A System to Automatically Assess Students' Free-Text Answers by using a Combination of Shallow NLP Techniques. *International Journal of Continuing Engineering Education and Life Long Learning*, **21**(2), 155–169.

Pérez-Marín, D., Alfonseca, E., and Rodríguez, P. (2006a). A Free-Text Scoring System that Generates Conceptual Models of the Students' Knowledge with the Aid of Clarifying Questions. In L. Aroyo and D. Dicheva, editors, *Proceedings of the Fourth International Workshop on Applications of Semantic Web Technologies for E-Learning*, pages 1–2, Dublin, Ireland.

Pérez-Marín, D., Alfonseca, E., Freire, M., Rodríguez, P., Guirao, J. M., and Moreno-Sandoval, A. (2006b). Automatic Generation of Students' Conceptual Models underpinned by Free-Text Adaptive Computer Assisted Assessment. In Kinshuk, R. Koper, P. Kommers, P. Kirschner, D. Sampson, and W. Didderen, editors, *Proceedings of the Sixth International Conference on Advanced Learning Technologies*, pages 280–284, Kerkrade, Netherlands. IEEE.

Pérez-Marín, D., Alfonseca, E., and Rodríguez, P. (2006c). On the Dynamic Adaptation of Computer Assisted Assessment of Free-Text Answers. In V. P. Wade, H. Ashman, and B. Smyth, editors, *Proceedings of the Fourth International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, volume 4018 of *Lecture Notes in Computer Science*, pages 374–377, Dublin, Ireland.

Pérez-Marín, D., Alfonseca, E., Rodríguez, P., and Pascual-Nieto, I. (2006d). Willow: Automatic and Adaptive Assessment of Students' Free-Text Answers. In F. Pla, editor, *Proceedings of the Twenty-Second International Conference of the Spanish Society for Natural Language Processing*, pages 367–368, Zaragoza, Spain.

Pérez-Marín, D., Pascual-Nieto, I., Alfonseca, E., and Anguiano, E. (2007). A Study on the Impact of the Use of an Automatic and Adaptive Free-Text Assessment System during a University Course. In J. Fong and F. L. Wang, editors, *Proceedings of the Workshop on Blended Learning*, ICWL '07, pages 186–195, Edinburgh, United Kingdom. Pearson.

Pérez-Marín, D., Pascual-Nieto, I., and Rodríguez, P. (2009). Computer-Assisted Assessment of Free-Text Answers. *The Knowledge Engineering Review*, **24**(4), 353–374.

Potthast, M. (2011). *Technologies for Reusing Text from the Web*. Ph.D. thesis, Bauhaus-Universität Weimar, Weimar, Germany.

Prettenhofer, P. and Stein, B. (2011). Cross-Lingual Adaptation using Structural Correspondence Learning. *Transactions on Intelligent Systems and Technology*, **3**, 13:1–13:22.

Pulman, S. G. and Sukkarieh, J. Z. (2005). Automatic Short Answer Marking. In J. Burstein and C. Leacock, editors, *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 9–16, Ann Arbor, Michigan. Association for Computational Linguistics.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In C. R. Perrault and C. S. Mellish, editors, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 1 of *IJCAI '95*, pages 448–453, Montreal, Canada. Morgan Kaufmann Publishers.

Richter, F. and Sailer, M. (2003). Basic Concepts of Lexical Resource Semantics. In A. Beckmann and N. Preining, editors, *Proceedings of the Fifteenth European Summer School in Logic Language and Information*, volume 5 of *Collegium Logicum*, pages 87–143, Vienna, Austria. Kurt Gödel Society.

Rodgers, J. L. and Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, **42**(1), 59–66.

Sargeant, J., McGee Wood, M., and Anderson, S. M. (2004). A Human-Computer Collaborative Approach to the Marking of Free Text Answers. In *Proceedings of the Eighth Computer Assisted Assessment Conference*, pages 361–370, Loughborough, United Kingdom. Loughborough University.

Shermis, M. D. and Burstein, J. (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, Mahwah, New Jersey, first edition.

Shermis, M. D. and Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge, New York City, New York, first edition.

Shermis, M. D., Burstein, J., and Leacock, C. (2008). Applications of Computers in Assessment and Analysis of Writing. In C. A. MacArthur, S. Graham, and J. Fitzgerald, editors, *Handbook of Writing Research*, chapter 27, pages 403–416. Guilford Press, New York City, New York, first edition.

Siddiqi, R. and Harrison, C. J. (2008a). A Systematic Approach to the Automated Marking of Short-Answer Questions. In M. K. Anis, M. K. Khan, and S. J. H. Zaidi, editors, *Proceedings of the Twelfth International Multitopic Conference*, pages 329–332, Karachi, Pakistan. IEEE.

Siddiqi, R. and Harrison, C. J. (2008b). On the Automated Assessment of Short Free-Text Responses. In *Proceedings of the Thirty-Fourth International Association for Educational Assessment Annual Conference*, pages 1–11, Cambridge, United Kingdom.

Siddiqi, R., Harrison, C. J., and Siddiqi, R. (2010). Improving Teaching and Learning through Automated Short-Answer Marking. *IEEE Transactions on Learning Technologies*, **3**(3), 237–249.

Sim, J. and Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, **85**(3), 257–268.

Sima, D., Schmuck, B., Szöll, S., and Miklós, A. (2007). Intelligent Short Text Assessment in eMax. In *Proceedings of the Eighth Africon Conference*, pages 1–6, Windhoek, South Africa. IEEE.

Sima, D., Schmuck, B., Szöll, S., and Miklós, A. (2009). Intelligent Short Text Assessment in eMax. In I. J. Rudas, J. Fodor, and J. Kacprzyk, editors, *Towards Intelligent Engineering and Information Technology*, volume 243 of *Studies in Computational Intelligence*, pages 435–445. Springer.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, **15**(1), 72–101.

Stern, A. and Dagan, I. (2011). A Confidence Model for Syntactically-Motivated Entailment Proofs. In R. Mitkov and G. Angelova, editors, *Proceedings of the Fourteenth International Conference on Recent Advances in Natural Language Processing*, pages 455–462, Hissar, Bulgaria. Association for Computational Linguistics.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, **103**(2684), 667–680.

Sukkarieh, J. Z. (2010). Using a MaxEnt Classifier for the Automatic Content Scoring of Free-Text Responses. In A. Mohammad-Djafari, J.-F. Bercher, and P. Bessiére, editors, *Proceedings of the International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 1305 of *AIP Conference Proceedings*, pages 41–18, Chamonix, France. American Institute of Physics.

Sukkarieh, J. Z. and Blackmore, J. (2009). c-rater: Automatic Content Scoring for Short Constructed Responses. In H. C. Lane and H. W. Guesgen, editors, *Proceedings of the Twenty-Second International Conference of the Florida Artificial Intelligence Research Society*, pages 290–295, Sanibel Island, Florida. AAAI Press.

Sukkarieh, J. Z. and Bolge, E. (2008). Leveraging c-rater's Automated Scoring Capability for Providing Instructional Feedback for Short Constructed Responses. In B. P. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, editors, *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems*, ITS '08, pages 779–783, Montreal, Canada. Springer.

Sukkarieh, J. Z. and Bolge, E. (2010). Building a Textual Entailment Suite for Evaluating Content Scoring Technologies. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC '10, pages 1–8, Valletta, Malta. European Language Resources Association.

Sukkarieh, J. Z. and Kamal, J. (2009). Towards Agile and Test-Driven Development in NLP Applications. In K. B. Cohen and M. Light, editors, *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '09, pages 42–44, Boulder, Colorado. Association for Computational Linguistics.

Sukkarieh, J. Z. and Pulman, S. G. (2005). Information Extraction and Machine Learning: Auto-Marking Short Free Text Responses to Science Questions. In C.-K. Looi, G. I. McCalla, B. Bredeweg, and J. Breuker, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence in Education*, Frontiers in Artificial Intelligence and Applications, pages 629–637, Amsterdam, Netherlands. IOS Press.

Sukkarieh, J. Z. and Stoyanchev, S. (2009). Automating Model Building in c-rater. In C. Callison-Burch and F. M. Zanzotto, editors, *Proceedings of the First ACL/IJCNLP Workshop on Applied Textual Inference*, TextInfer '09, pages 61–69, Suntec, Singapore. Association for Computational Linguistics.

Sukkarieh, J. Z., Pulman, S. G., and Raikes, N. (2003). Auto-marking: Using Computational Linguistics to Score Short, Free Text Responses. In *Proceedings of the Twenthy-Ninth Annual Conference of the International Association for Educational Assessment*, pages 1–15, Manchester, United Kingdom.

Sukkarieh, J. Z., Pulman, S. G., and Raikes, N. (2004). Auto-marking 2: An update on the UCLES-Oxford University Research into using Computational Linguistics to Score Short, Free Text Responses. In *Proceedings of the Thirtieth Annual Conference of the International Association for Educational Assessment*, Philadelphia, Pennsylvania.

Swithenby, S. and Jordan, S. (2008). Supporting Open Learners by Computer Based Assessment with Short Free-Text Responses and Tailored Feedback. In F. Welsch, F. Malpica, A. Tremante, J. V. Carrasquero, and A. Oropeza, editors, *Proceedings of the Second International Multi-Conference on Society, Cybernetics and Informatics*, IMSCI '08, Orlando, Florida. International Institute of Informatics and Systemics.

Szpektor, I. and Dagan, I. (2007). Learning Canonical Forms of Entailment Rules. In R. Mitkov and G. Angelova, editors, *Proceedings of the Sixth International Conference on Recent Advances in Natural Language Processing*, pages 1–6, Borovets, Bulgaria.

Tandalla, L. (2012). Scoring Short Answer Essays. ASAP '12 SAS methodology paper.

Thomas, P. (2003). The Evaluation of Electronic Marking of Examinations. In *Proceedings of the Eighth Annual Conference on Innovation and Technology in Computer Science Education*, ITiCSE '03, pages 50–54, Thessaloniki, Greece. ACM.

Valenti, S., Neri, F., and Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education*, **2**, 319–330.

VanLehn, K., Jordan, P. W., Rosé, C. P., Bhembe, D., Böttner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S., and Srivastava, R. (2002). The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing. In S. A. Cerri, G. Gouarderes, and F. Paraguacu, editors, *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems*, volume 2363 of *Lecture Notes in Computer Science*, pages 158–167, Biarritz, France. Springer.

Wachsmuth, H., Stein, B., and Engels, G. (2011). Constructing Efficient Information Extraction Pipelines. In B. Berendt, A. de Vries, W. Fan, C. MacDonald, I. Ounis, and I. Ruthven, editors, *Proceedings of the Twentieth ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2237–2240, Glasgow, Scotland. ACM.

Wachsmuth, H., Stein, B., and Engels, G. (2013). Information Extraction as a Filtering Task. In Q. He and A. Iyengar, editors, *Proceedings of the Twenty-Second ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 2049–2058, San Francisco, California. ACM.

Wang, H.-C., Chang, C.-Y., and Li, T.-Y. (2008). Assessing Creative Problem-Solving with Automated Text Grading. *Computers & Education*, **51**(4), 1450–1466.

Wang, X., Evanini, K., and Zechner, K. (2013). Coherence Modeling for the Automated Assessment of Spontaneous Spoken Responses. In L. Vanderwende, H. Daumé, and K. Kirchhoff, editors, *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 814–819, Atlanta, Georgia. Association for Computational Linguistics.

Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, **31**(1), 2–13.

Willis, A. (2010). Inductive Logic Programming to Support Automatic Marking of Student Answers in Free Text. Final report, COLMSCT, The Open University, Milton Keynes, United Kingdom.

Wise, M. J. (1993). String Similarity via Greedy String Tiling and Running Karp-Rabin Matching. Technical report, Department of Computer Science, University of Sydney, Sydney, Australia.

Wood, M. M., Jones, C., Sargeant, J., and Reed, P. (2006). Light-Weight Clustering Techniques for Short Text Answers in HCC CAA. In M. Danson, editor, *Proceedings of the Tenth CAA International Computer Assisted Assessment Conference*, pages 291–308, Loughborough, United Kingdom. Loughborough University.

Wu, Z. and Palmer, M. (1994). Verb Semantics and Lexical Selection. In J. Pustejovsky, editor, *Proceedings of the Thirty-Second Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Las Cruces, New Mexico. Association for Computational Linguistics.

Zbontar, J. (2012). Short Answer Scoring by Stacking. ASAP '12 SAS methodology paper.

Zenisky, A. L. and Sireci, S. G. (2002). Technological Innovations in Large-Scale Assessment. *Applied Measurement in Education*, **15**(4), 337–362.

Zesch, T., Levy, O., Gurevych, I., and Dagan, I. (2013). UKP-BIU: Similarity and Entailment Metrics for Student Response Analysis. In S. Manandhar and D. Yuret, editors, *Proceedings of the Seventh International Workshop on Semantic Evaluation*, volume 2, pages 285–289, Atlanta, Georgia. Association for Computational Linguistics.

Ziai, R., Ott, N., and Meurers, D. (2012). Short Answer Assessment: Establishing Links Between Research Strands. In J. Tetreault, J. Burstein, and C. Leacock, editors, *Proceedings of the Seventh Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 190–200, Montreal, Canada. Association for Computational Linguistics.