Short Answer Assessment: Establishing Links Between Research Strands

Ramon Ziai Niels Ott Detmar Meurers
SFB 833 / Seminar für Sprachwissenschaft
Universität Tübingen
{rziai,nott,dm}@sfs.uni-tuebingen.de

Abstract

A number of different research subfields are concerned with the automatic assessment of student answers to comprehension questions, from language learning contexts to computer science exams. They share the need to evaluate free-text answers but differ in task setting and grading/evaluation criteria, among others.

This paper has the intention of fostering synergy between the different research strands. It discusses the different research strands, details the crucial differences, and explores under which circumstances systems can be compared given publicly available data. To that end, we present results with the CoMiC-EN Content Assessment system (Meurers et al., 2011a) on the dataset published by Mohler et al. (2011) and outline what was necessary to perform this comparison. We conclude with a general discussion on comparability and evaluation of short answer assessment systems.

1 Introduction

Short answer assessment systems compare students' responses to questions with manually defined target responses or answer keys in order to judge the appropriateness of the responses, or in order to automatically assign a grade. A number of approaches have emerged in recent years, each of them with different aims and different backgrounds. In this paper, we will draw a map of the short answer assessment landscape, highlighting the similarities and differences between approaches and the data used for evaluation. We will provide an overview of 12 systems and sketch their attributes. Subsequently, we will zoom into the comparison of two of them, namely CoMiC-EN (Meurers et al., 2011a) and the one which we call the Texas system (Mohler et al., 2011) and discuss the issues that arise with this endeavor. Returning to the bigger picture, we will explore how such systems could be compared in general, in the belief that meaningful comparison of approaches across research strands will be an important ingredient in advancing this relatively new research field.

2 The short answer assessment landscape

2.1 General aspects

Researchers from all directions have settled in the landscape of short answer assessment, each of them with different backgrounds and different goals. In this section, we aim at providing an overview of these research villages, also hoping to construct a road network that may connect them.

Most approaches to short answer assessment are situated in an educational context. Some focus on GCSE[1] tests, others aim at university assessment tests in the medical domain. Another strand of approaches focuses on language teaching and learning. All of these approaches share one theme: they assess short texts written by students. These may be answers to questions that ask for knowledge acquired in a course, e.g., in computer science, or to reading comprehension questions in second language

[1] The General Certificate of Secondary Education (GCSE) is an academic qualification in the United Kingdom, usually taken at the age of 14–16.

learning. While thematically related, short answer assessment is different from essay grading. Short answers are formulated by students in a much more controlled setting. Not only are they short, they usually are supposed to contain only a few facts that answer only one question.

Another common theme of these approaches is that they compare the student answers to one or more previously defined correct answers that are either given in natural language as target answers or as a list of concepts in an answer key. The ways of technically conducting these comparisons vary widely, as we discuss below in Section 2.2.

There also are conceptual differences between the approaches. Some systems focus on assessing whether or not the student has properly answered the question. They put the spot on comparing the meaning of target answers and student answers; they aim at being tolerant of form errors such as spelling or grammar errors. Others aim at giving a grade as accurate as possible, therefore not only assessing meaning but also performing grading similar to human teachers. This can also include modules that take into account form errors.

These two views on a similar task are also reflected in the annotation of the data used in experiments: Systems performing meaning comparison usually operate with labels specifying the relations between target answers and student answers. Grading systems naturally aim at producing numerical grades. Since labels are on a nominal scale, and grades are on an ordinal scale (or even treated as being on an interval scale), the difference between meaning comparison and grading results in a whole string of other differences in methodology.

Researchers also enter the short answer landscape from different home countries: Some projects are interested in the strategies and mechanics of meaning comparison, others aim at reducing the load and costs of large-scale assessment tests, and yet others aim at improving intelligent tutoring systems, requiring additional components that provide useful feedback to students using these systems.

2.2 Approaches

Table 1 summarizes the features of the short answer assessment systems discussed hereafter.

One of the earlier systems is WebLAS, presented by Bachman et al. (2002). A human task creator feeds the system with scores for model answers. Regular expressions are then created automatically from these model answers. Since each regular expression is associated with a score, matching the expression against a student answer yields a score for that answer. Bachman et al. (2002) do not provide an evaluation study based on data.

Another earlier system is CarmelTC by Rosé et al. (2003). It has been designed as a component in the Why2 tutorial dialogue system (VanLehn et al., 2002). Even though Rosé et al. (2003) position CarmelTC in the context of essay grading, it may be considered to deal with short answers: in their data, the average length of a student response is approx. 48 words. Their system is designed to perform text classification on single sentences in the student responses, where each class of text represents one possible model response, plus an additional class for 'no match'. They combine decision trees operating on an automatic syntactic analysis, a Naive Bayes text classifier, and a bag-of-words approach. In a 50-fold cross validation experiment with one physics question, six classes and 126 student responses, hand-tagged by two annotators, CarmelTC reaches an F-measure value of 0.85. They do not report on a baseline. Concerning the quality of the gold standard, they report that conflicts in the annotation have been resolved.

C-Rater (Leacock and Chodorow, 2003) is based on a paraphrase recognition approach. It employs correct answer models consisting of essential points formulated in natural language. C-Rater aims at automatic scoring and focuses on meaning, thus tolerating form errors. Leacock and Chodorow (2003) present two pilot studies, one of them dealing with reading comprehension. From 16,625 student answers with an average length of 43 words, they drew a random sample of 100 answers to each of the seven questions. This sample was scored by one human judge using a three-way scoring system (full credit, partial credit, no credit). Their system achieved 84% agreement with the gold standard. Information about the distribution of the scoring categories is given indirectly: A baseline system that assigns scores randomly would have achieved 47% accuracy.

System	Goal	Technique	Domain	Lang.
WebLAS (Bachman et al., 2002)	Assessment of language ability	Auto-generated regular expressions	Foreign language teaching	EN
CarmelTC (Rosé et al., 2003)	Automatic grading	Text classification	Physics	EN
C-Rater (Leacock and Chodorow, 2003)	Assessment test	Paraphrase recognition	Mathematics, Reading comp.	EN
IAT (Mitchell et al., 2003)	Assessment, Automatic grading	Information extraction w/ handwritten patterns	Medical	EN
Oxford (Pulman and Sukkarieh, 2005)	Assessment, automatic grading	Information extraction w/ handwritten patterns	GCSE exams	EN
Atenea (Pérez et al., 2005)	Automatic grading	N-gram overlap, Latent Semantic Analysis	Computer science	ES
Logic-based System (Makatchev and VanLehn, 2007)	Meaning comparison	First-order logic, machine learning	Physics	EN
CAM (Bailey and Meurers, 2008), CoMiC-EN (Meurers et al., 2011a)	Meaning comparison	Alignment, machine learning	Reading comp. in foreign language	EN
Facets System (Nielsen et al., 2009)	Meaning comparison & tutoring systems	Alignment of facets, machine learning	Elementary school science classes	EN
Texas (Mohler et al., 2011)	Automatic grading	Graph alignment, semantic similarity	Computer science	EN
CoMiC-DE (Meurers et al., 2011b)	Meaning comparison	Alignment, machine learning	Reading comp. in foreign language	DE
CoSeC-DE (Hahn and Meurers, 2012)	Meaning comparison	Alignment via Lexical-Resource Semantics	Reading comp. in foreign language	DE

Table 1: Short Answer Assessment systems and their Features

Information extraction templates form the core of the Intelligent Assessment Technologies system (IAT, Mitchell et al. 2003). These templates are created manually in a special-purpose authoring tool by exploring sample responses. They allow for syntactic variation, e.g., filling the subject slot in a sentence with different equivalent concepts. The templates corresponding to a question are then matched against the student answer. Unlike other systems, IAT additionally features templates for explicitly invalid answers. They tested their approach with a progress test that has to be taken by medicine students. Approximately 800 students each plowed through 270 test items. The automatically graded responses then were moderated: Human judges streamlined the answers to achieve a more consistent grading. This step already had been done before with tests graded by humans. Mitchell et al. (2003) state that their system reaches 99.4% accuracy on the full dataset after the manual adjustment of the templates via the moderation process. Summarizing, they report an error of "between 5 and 5.5%" in inter-grader agreement and an error of 5.8% in automatic grading without the moderation step, though it is not entirely clear which data these statistics correspond to. No information on the distribution of grades or a random baseline is provided.

The Oxford system (Pulman and Sukkarieh, 2005) is another one to employ an information extraction approach. Again, templates are constructed manually. Motivated by the necessary robustness to process language with grammar mistakes and spelling errors, they use shallow analyses in their pre-processing. In order to overcome the hassle of manually constructing templates, they also investigated machine learning techniques. However, the automatically generated templates were outperformed by the manually created ones. Furthermore, they state that manually created templates can be equipped with messages provided to the student as feedback in a tutoring system. For evaluating their system, they used factual science questions and the corresponding

student answers from GCSE tests. 200 graded answers for each of nine questions served as a training set, while another 60 answers served as a test set. They report that their system achieves an accuracy of 84%. With inconsistencies in the human grading removed, it achieves 93%. However, they do not report on the level of inter-grader agreement or on a random baseline.

Pérez et al. (2005) present the Atenea system, a combined approach that makes use of Latent Semantic Analysis (LSA, Landauer et al. 1998) and n-gram overlap. While n-gram overlap supports comparing target responses and student responses with differing word order, it does not deal with synonyms and related terms. Hence, they use LSA to add a component that deals with semantic relatedness in the comparison step. As a test corpus, they collected nine different questions from computer science exams. A tenth question "[consists] of a set of definitions of 'Operating System' obtained from the Internet." Altogether, they gathered 924 student responses and 44 target responses written by teachers. Since their LSA module had been trained on English but their data were in Spanish, they chose to use Altavista Babelfish to translate the data into English. They do not provide information about the distribution of scores and about inter-grader agreement. Atenea achieves a Pearson's correlation of $r = 0.554$ with the scores in the gold standard.

The approach by Makatchev and VanLehn (2007), which we refer to as the Logic-based System, enters the landscape from the direction of artificial intelligence. It is related to CarmelTC and its dataset, but follows a different route: target responses are manually encoded in first-order predicate language. Similar logic representations are constructed automatically for student answers. They explore various strategies for matching these two logic representation on the basis of 16 semantic classes. In an evaluation experiment, they tested the system on 293 "natural language utterances" with ten-fold cross validation. The test data are skewed towards the 'empty' label that indicates that none of the 16 semantic labels could be attached. They do not report on other properties of the dataset such as number of annotators or number of questions to which the student answers were given. Their winning configuration yields a F-measure value of 0.4974. While Makatchev and VanLehn (2007) position their approach in the context of the Why2 tutorial dialogue system, their use of semantic classes seems to make them more related to meaning comparison than to grading.

The Content Assessment Module (CAM) presented in Bailey (2008) and Bailey and Meurers (2008) utilizes an approach that is different from the systems discussed so far: Following a three-step strategy, the system first automatically generates linguistic annotations for questions, target responses and student responses. In an alignment phase, these annotations are then used to map from elements (words, lemmas, chunks, dependency triples) in the student responses to elements in the target responses. Finally, a machine learning classifier judges on the basis of this alignment, whether or not the student has answered the question correctly. The data used for evaluation was made available as the Corpus of Reading Comprehension Exercises in English (CREE, Meurers et al. 2011a). This corpus consists of 566 responses produced by intermediate ESL learners at The Ohio State University as part of their regular assignments. Students had access to their textbooks and typically answered questions in one to three sentences. All responses were labelled as either appropriate or inappropriate by two independent annotators, along with a detailed diagnosis code specifying the nature of the inappropriateness (missing concept, extra concept, blend, non-answer). In leave-one-out evaluation on the development set containing 311 responses to 47 different questions, CAM achieved 87% accuracy on the binary judgment (response correct/incorrect). For the test set containing 255 responses to 28 questions, the approach achieved 88%. However, the distribution of categories in the data is heavily skewed with 71% of the responses marked as correct in the development set and 84% in the test set. The best result obtained on a balanced set with leave-one-out-testing is 78%. Meurers et al. (2011a) present a re-implementation of CAM called CoMiC-EN (Comparing Meaning in Context in English), achieving an accuracy of 87.6% on the CREE development set and 88.4% on the test set.

With their Facets System, Nielsen et al. (2009) establish a connection to the field of Recognizing Textual Entailment (RTE, Dagan et al. 2009). In

a number of friendly challenges, RTE research has spawned numerous systems that try to automatically answer the following question: Given a text and a hypothesis, is the hypothesis entailed by the text? Short answers assessment can be seen as a RTE task in which the target response corresponds to the text and the student response to the hypothesis. Nielsen et al. (2009) base their system on what they call facets. These facets are meaning representations of parts of sentences. They are constructed automatically from dependency and semantic parses of the target responses. Each facet in the target response is then looked up in the corresponding student response and equipped with one of five labels[2] ranging from unaddressed (the student did not mention the fact in this facet) to expressed (the student named the fact). This step is taken via machine learning. From a tutoring system in real-life operation, they gathered responses from third- to sixth-grade students answering questions for science classes. Two annotators worked on these data, producing 142,151 facets. Furthermore, all facets were looked up in the corresponding student responses and annotated accordingly, using the mentioned set of labels. The best result of the Facets System is 75.5% accuracy on one of the held-out test sets. With ten-fold cross validation on the training set, it achieves 77.1% accuracy. The majority label baselines are 51.1% and 54.6% respectively. Providing this more fine-grained analysis of facets that are searched for in student responses, Nielsen et al. (2009) claim to "enable more intelligent dialogue control" in tutoring systems. From the point of view of grading vs. meaning comparison, their approach can be counted towards the latter, since their labels can be conflated to produce a single yes/no decision.

Another recent approach is described by Mohler et al. (2011), hereafter referred to as the Texas system. Student responses and target responses are annotated using a dependency parser. Thereupon, subgraphs of the dependency structures are constructed in order to map one response to the other. These alignments are generated using machine learning. Dealing with subgraphs allows for variation in word order between the two responses that are to be compared. In order to account for meaning, they combine lexical semantic similarity with the aforementioned alignment. They make use of several WordNet-based measures and two corpus-based measures, namely Latent Semantic Analysis and Explicit Semantic Analysis (ESA, Gabrilovich and Markovitch 2007). For evaluating their system, Mohler et al. (2011) collected student responses from an online learning environment. 80 questions from ten introductory computer science assignments spread across two exams were gathered together with 2,273 student responses. These responses were graded by two human judges on a scale from zero to five. The judges fully agreed in 57% of all cases, their Pearson correlation computes to $r = 0.586$. The gold standard has been created by computing the arithmetic mean of the two judgments for each response. The Texas system achieves $r = 0.518$ and a Root Mean Square Error of 0.978 as its best result. Mohler et al. (2011) mention that "[t]he dataset is biased towards correct answers". Data are publicly available. We used these in an evaluation experiment with the CoMiC-EN system, discussed in Section 3.

While almost all short answer assessment research has targeted answers written in English, there are two recent approaches dealing with German answers. The CoMiC-EN reimplementation of CAM discussed above was motivated by the need for a modular architecture supporting a transfer of the system to German, resulting in its counterpart named CoMiC-DE (Meurers et al., 2011b). The German system utilizes the same strategies as the English one, but with language-dependent processing modules being replaced. Meurers et al. (2011b) evaluated CoMiC-DE on a subset of the Corpus of Reading Comprehension Questions in German (CREG, Ott et al. 2012), collected in collaboration with the German programs at The Ohio State University and the University of Kansas. Like in CREE, all responses are rated by two annotators with both binary and detailed diagnosis codes.[3] The aforementioned subset contains 1,032 learner responses and 223 target responses to 177 questions. Furthermore, it features an even distribution of correct and incorrect answers according to the judgement of two human

[2] In human annotation, they use eight labels, which are grouped into five broader categories as used by their system.

[3] In CREG, correct answers as well as incorrect ones can be labelled with missing concept, extra concept, or blend.

annotators. On that subset, CoMiC-DE achieved an accuracy of 84.6% in the binary classification task. CREG is freely available for research purposes under a Creative Commons by-nc-sa license.

Hahn and Meurers (2012) present the CoSeC-DE approach based on Lexical Resource Semantics (LRS, Richter and Sailer 2003). In a first step, they create LRS representations from POS-tagged and dependency-parsed data. These underspecified LRS representations of student responses and target responses are then aligned. Using A* as heuristic search algorithm, a best alignment is computed and equipped with a numeric score representing the quality of the alignment of the formulae. If this best alignment scores higher than a threshold, the system judges student response and target response to convey the same meaning. The alignment and comparison mechanism does not utilize any linguistic representations other than the LRS semantic formulae. These semantic representations abstract away from surface features, e.g., by treating active and passive voice equally. Hahn and Meurers (2012) claim that that "[semantic representations] more clearly expose those distinction which do make a difference in meaning." They evaluate the approach on the above-mentioned subset of CREG containing 1,032 learner responses and report an accuracy of 86.3%.

3 A concrete system comparison

After discussing the broad landscape of Short Answer Evaluation systems, the main characteristics and differences, we now turn to a comparison of two concrete systems, namely CoMiC-EN (Meurers et al., 2011a) and the Texas system Mohler et al. (2011), to explore what is involved in such a concrete comparison of two systems from different contexts. While CoMiC-EN was developed with meaning comparison in mind, the purpose of the Texas system is answer grading. We pick these two systems because they constitute recent and interesting instances of their respective fields and the corresponding data are freely available.

3.1 Data

In evaluating the Texas system, Mohler et al. (2011) used a corpus of ten assignments and two exams from an introductory computer science class. In total, the Texas corpus consists of 2,442 responses, which were collected using an online learning platform. Each response is rated by two annotators with a numerical grade on a 0–5 scale. Annotators were not given any specific instructions besides the scale itself, which resulted in an exact agreement of 57.7%. In order to arrive at a gold standard rating, the numerical average of the two ratings was computed. The data exist in raw, sentence-segmented and parsed versions and are freely available for research use. Table 2 presents a breakdown of the score counts and distribution statistics of the Texas corpus. A bias towards correct answers can be observed, which is also mentioned by Mohler et al. (2011).

Score	#	**Score**	#
0.000	24	3.250	1
0.500	3	3.500	187
1.000	23	3.625	1
1.500	46	3.750	1
1.750	1	4.000	220
2.000	93	4.125	2
2.250	2	4.500	310
2.500	125	4.750	1
3.000	164	5.000	1238

$\overline{x} = 4.19, s = 1.11$

Table 2: Details on the gold standard scores in the Texas corpus. Non-integer scores result from averaging between raters and normalization onto the 0–5 scale.

3.2 Approaches

CoMiC-EN uses a three-step approach to meaning comparison. *Annotation* uses NLP to enrich the student and target answers, as well as the question text, with linguistic information on different levels (words, chunks, dependency triples) and types of abstraction (tokens, lemmas, distributional vectors, etc.). *Alignment* maps elements of the learner answer to elements of the target response using annotation. The global alignment solution is computed using the Traditional Marriage Algorithm (Gale and Shapley, 1962). Finally, *Classification* analyzes the possible alignments and labels the learner response with a binary or detailed diagnosis code. The features used in the classification step are shown in Table 3.

For the Texas system, Mohler et al. (2011) used a combination of bag-of-words (BOW) features and

Features	Description
1. Keyword Overlap	Percent of keywords aligned (relative to target)
2./3. Token Overlap	Percent of aligned target/learner tokens
4./5. Chunk Overlap	Percent of aligned target/learner chunks
6./7. Triple Overlap	Percent of aligned target/learner triples
8. Token Match	Percent of token alignments that were token-identical
9. Similarity Match	Percent of token alignments that were similarity-resolved
10. Type Match	Percent of token alignments that were type-resolved
11. Lemma Match	Percent of token alignments that were lemma-resolved
12. Synonym Match	Percent of token alignments that were synonym-resolved
13. Variety of Match (0-5)	Number of kinds of token-level alignments

Table 3: Features used in the CoMiC-EN system

dependency graph alignment in connection with two different machine learning approaches. Among the BOW features are WordNet-based similarity measures such as the one by Lesk (1986) and vector space measures such as $tf * idf$ (Salton and McGill, 1983) and the more advanced LSA (Landauer et al., 1998). The dependency graph alignment approach builds on a node-to-node matching stage which computes a score for each possible match between nodes of the student and target response. In the next stage, the optimal graph alignment is computed based on the node-to-node scores using the Hungarian algorithm.

Mohler et al. (2011) also employ a technique they call "question demoting", which refers to the exclusion of words from the alignment process if they already appeared in the question string. Incidentally, the technique is also used in the earlier CAM system (Bailey and Meurers, 2008), but called "Givenness filter" there, following the long tradition of research on givenness (Schwarzschild, 1999) as a notion of information structure investigated in formal pragmatics.

To produce the final system score, the Texas system uses two machine learning techniques based on Support Vector Machines (SVMs), SVMRank and Support Vector Regression (SVR). Both techniques are trained with several combinations of the dependency alignment and BOW features. While with SVR one trains a function to produce a score on the 0–5 scale itself, SVMRank produces a ranking of student answers which does not produce a 0–5 grade. Therefore, Mohler et al. (2011) employ isotonic regression to map the ranking to the 0–5 scale.

In terms of performance, Mohler et al. (2011) report that the SVMRank system produces a better correlation measure ($r = 0.518$) while the SVR system yields a better RMSE (0.978).

3.3 Evaluation

We now turn to the evaluation of CoMiC-EN on the Texas corpus as it is a publicly available dataset. As mentioned before, CoMiC-EN performs meaning comparison based on a system of categories while the Texas system is a scoring approach, trying to predict a grade. While the former is a *classification* task, the latter is better characterized as a *regression* problem because of the desired numerical outcome. Of course, one could simply pretend that individual grades are classes and treat scoring as a classification task. However, a classification approach has no knowledge of numerical relationships, i.e., it does not 'know' that 4 is a higher grade than 3 and a much higher grade than 1 (assuming a 0–5 scale). As a result, if an evaluation metric such as Pearson correlation is used, classification systems are at a disadvantage because some misclassifications are punished more than others. We discuss this point further in Section 4.

For these reasons, to obtain a more interesting comparison, we modified CoMiC-EN to perform scoring instead of meaning comparison. This means that the memory-based learning approach CoMiC-EN had employed so far was no longer applicable and had to be replaced with a regression-capable learning strategy. We chose Support Vector Regression (SVR) using libSVM[4] since that is one of the methods employed by Mohler et al. (2011). However, all other parts of CoMiC-EN such as the processing pipeline and the alignment approach and the extracted features remained the same.

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm

The evaluation procedure was carried out as a 12-fold cross-validation due to the 12 assignments in the Texas corpus. For each fold, one complete assignment was held out as test set. Parameters for the SVR were determined using a grid search using the tools provided with libSVM. As kernel function, we used a linear kernel as it was also used in the evaluation of the Texas system and thus constitutes a vital part of the evaluation setup. In general, we designed to evaluation procedure to be as close as possible to the Texas one.

Table 4 presents detailed results on the 12 folds as well as the overall results and a baseline which always predicts the median value 5.

Assignment	# responses	*r*	RMSE
1	203	0.416	0.958
2	210	0.349	1.221
3	217	0.335	0.969
4	210	0.338	1.212
5	112	0.010	1.030
6	182	0.646	0.702
7	182	0.265	0.991
8	189	0.521	0.942
9	189	0.220	0.942
10	168	0.699	0.990
11 (exam)	300	0.436	1.076
12 (exam)	280	0.619	1.165
Median Baseline	2442	–	1.375
Overall	2442	0.405	1.016

Table 4: Detailed results of CoMiC-EN on Texas corpus

The CoMiC-EN system on the Texas data set does not quite reach the level achieved by the Texas system on their data set. We obtained a Pearson correlation of $r = 0.405$ and an RMSE of 1.016 over all 12 folds. However, let us keep in mind the objective of this experiment as exemplifying the process needed to directly compare two systems from different research strands on the same dataset.

4 Comparability of approaches & datasets

It seems clear that for systems to be comparable and results to be reproducible, datasets must be publicly available, as is the case with the Texas corpus. However, data availability alone does not ensure meaningful comparison. Depending on the context the corpus was drawn from, datasets will differ just like the corresponding systems:

- Data source: Reading comprehension task in language learning setting, language tutoring context, automated grading of short answer exams
- Language properties: Native vs. learner language, domain-specific language (e.g., computer science)
- Assessment scheme: nominal vs. interval scale

Especially the last point deserves some further discussion. Depending on the kind of assessment scheme, which in turn is motivated by the task, different evaluation methods may be chosen. Scoring systems are often evaluated using a correlation metric in order to capture the systems' tendency to assign similar but not necessary equal grades as the human raters. Conversely, with category-based schemes one usually reports accuracy, which expresses how many items were classified correctly.

The question that arises is how a system coming from one paradigm can be compared to one from the other paradigm in a meaningful way. One might argue that the tasks are simply too different: scoring might take form errors into account while meaning comparison by definition does not. Moreover, while classification labels say something explicit and absolute about a piece of data, grades by definition are relative to the scale they come from. It thus seems impossible to somehow unify the two schemes as they express fundamentally different ideas.

However, the strategies systems use to tackle scoring or meaning comparison are undoubtedly similar and should be comparable, as we argue in this paper. So in order for researchers to learn from other approaches and also compare their results to those of other systems which tackle a different task, changes to systems seem necessary and should be preferred over changes to the gold standard data. In the case presented here, a meaning comparison system was turned into a scoring system by changing the machine learning component from classification to regression, which requires a certain level of system modularity.

Having compared the two systems using Pearson correlation and RMSE, it also makes sense to consider the relevance of these evaluation metrics. For example, it is the case that pairwise correlation assumes a normal distribution whereas datasets like

the Texas corpus are heavily skewed towards correct answers (see Table 2). Mohler et al. (2011) also note that in distributions with zero variance, correlation is undefined, which is not a problem as such but limits the use of correlation as evaluation metric. Mohler et al. (2011) propose that RMSE is better suited to the task since it captures the relative error a system makes when trying to predict scores. However, RMSE is scale-dependent and thus RMSE values across different studies cannot be compared. We can only suggest that in order to sufficiently describe a system's performance, several metrics need to be reported.

Finally, an important point concerns the quality of gold standards. Given the relatively low inter-annotator agreement in the Texas corpus ($r = 0.586, RMSE = 0.659$) it seems fair to ask whether answers without perfect agreement should be used in training and testing systems at all. In the CREE and CREG corpora, answers with disagreement among the annotators have either been excluded from experiments or resolved by an additional judge. This approach is also supported by recent literature (cf., e.g., Beigman and Beigman Klebanov 2009; Beigman Klebanov and Beigman 2009). However, for the Texas corpus, Mohler et al. (2011) have opted to use the arithmetic mean of the two graders as gold standard. While mathematically a viable solution, it seems questionable whether the mean is reliable with only two graders, especially if they have not operated on the grounds of explicit guidelines. It would be interesting to see whether in this case, a system trained on more, singly annotated data would perform better than one on less, doubly annotated data, as argued for by Dligach et al. (2010). In any case, if many disagreements occur, one should ask the question whether the annotation task is defined well enough and whether machines should really be expected to perform it consistently if humans have trouble doing so.

5 Conclusion

We discussed several issues in the comparison of short answer evaluation systems. To that end, we gave an overview of the existing systems and picked two for a concrete comparison on the same data, the CoMiC-EN system (Meurers et al., 2011a) and the Texas system (Mohler et al., 2011). In comparing the two, it was necessary to turn CoMiC-EN into a scoring system because the Texas corpus as the chosen gold standard contains numeric scores assigned by humans. Taking a step back from the concrete comparison, we gave a more general description of what is necessary to compare short answer evaluation systems. We observed that more datasets need to be publicly available in order for performance comparisons to have meaning, a point also made earlier by Pulman and Sukkarieh (2005). Moreover, we noted how datasets differ in similar aspects as systems do, such as task context and assessment scheme. We then criticized the use of correlation measures as evaluation metrics for short answer scoring. Finally, we discussed the importance of gold standard quality.

We conclude that it is interesting and relevant to compare short answer evaluation systems even if the concrete task they tackle, such as grading or meaning comparison, is not the same. However, the availability and quality of the datasets will decide to what extent systems can sensibly be compared. For progress to be made in this area, more publicly available datasets and systems are needed. The upcoming SemEval-2013 task on "Textual entailment and paraphrasing for student input assessment"[5] will hopefully become one important step into this direction (see also Dzikovska et al. 2012).

Acknowledgements

We are grateful to the three anonymous BEA reviewers for their detailed and helpful comments.

References

Lyle Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Michael Pan, Chris Salvador, and Yasuyo Sawaki. 2002. A reliable approach to automatic assessment of short answer free responses. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1–4.

Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In Joel Tetreault, Jill Burstein, and Rachele De Felice, editors, *Proceedings of the*

[5] http://www.cs.york.ac.uk/semeval-2013/task4/

3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08, pages 107–115, Columbus, Ohio.

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

Eyal Beigman and Beata Beigman Klebanov. 2009. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 280–287. Association for Computational Linguistics.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii, 10.

Dmitriy Dligach, Rodney D. Nielsen, and Martha Palmer. 2010. To annotate more accurately or to annotate more. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 64–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12.

David Gale and Lloyd S. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly*, 69:9–15.

Michael Hahn and Detmar Meurers. 2012. Evaluating the meaning of answers to reading comprehension questions: A semantics-based approach. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*, Montreal.

Thomas Landauer, Peter Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37:389–405.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, Toronto, Ontario, Canada.

Maxim Makatchev and Kurt VanLehn. 2007. Combining baysian networks and formal reasoning for semantic classification of student utterances. In *Proceedings of the International Conference on AI in Education (AIED)*, Los Angeles, July.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. 2011a. Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *IJCEELL. Special Issue on Automatic Free-text Evaluation*, 21(4):355–369.

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

Tom Mitchell, Nicola Aldrige, and Peter Broomhead. 2003. Computerized marking of short-answer free-text responses. Paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752–762, Portland, Oregon, USA, June. Association for Computational Linguistics.

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

Niels Ott, Ramon Ziai, and Detmar Meurers. 2012. Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM). Benjamins, Amsterdam. to appear.

Diana Pérez, Enrique Alfonseca, Pilar Rodríguez, Alfio Gliozzo, Carlo Strapparava, and Bernardo Magnini. 2005. About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. *Revista signos*, 38(59):325–343.

Stephen G. Pulman and Jana Z. Sukkarieh. 2005. Automatic short answer marking. In Jill Burstein and Claudia Leacock, editors, *Proceedings of the Second*

Workshop on Building Educational Applications Using NLP, pages 9–16, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Frank Richter and Manfred Sailer. 2003. Basic concepts of lexical resource semantics. In Arnold Beckmann and Norbert Preining, editors, *ESSLLI 2003 – Course Material I*, volume 5 of *Collegium Logicum*, pages 87–143, Wien. Kurt Gödel Society.

Carolyn Penstein Rosé, Antonio Roque, Dumisizwe Bhembe, and Kurt VanLehn. 2003. A hybrid approach to content analysis for automatic essay grading. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL-Short '03, pages 88–90, Edmonton, Canada. Association for Computational Linguistics.

Gerard Salton and Michael J. McGill. 1983. *Introduction to modern information retrieval*. McGraw-Hill, New York.

Roger Schwarzschild. 1999. GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics*, 7(2):141–177.

Kurt VanLehn, Pamela W. Jordan, Carolyn Penstein Rosé, Dumisizwe Bhembe, Michael Boettner, Andy Gaydos, Maxim Makatchev, Umarani Pappuswamy, Micheal Ringenberg, Antonio Roque, Stephanie Siler, and Ramesh Srivastava. 2002. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, volume 2363, pages 158–167, Biarritz, France and San Sebastian, Spain, June 2-7. Springer LNCS.