SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge

Myroslava O. Dzikovska
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom
m.dzikovska@ed.ac.uk

Rodney D. Nielsen
University of North Texas
Denton, TX, USA
Rodney.Nielsen@UNT.edu

Chris Brew
Nuance Communications
USA
cbrew@acm.org

Claudia Leacock
CTB McGraw-Hill
USA
claudia_leacock@mheducation.com

Danilo Giampiccolo
CELCT
Italy
giampiccolo@celct.it

Luisa Bentivogli
CELCT and FBK
Italy
bentivo@fbk.eu

Peter Clark
Vulcan Inc.
USA
peterc@vulcan.com

Ido Dagan
Bar-Ilan University
Israel
dagan@cs.biu.ac.il

Hoa Trang Dang
NIST
hoa.dang@nist.gov

Abstract

We present the results of the Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge, aiming to bring together researchers in educational NLP technology and textual entailment. The task of giving feedback on student answers requires semantic inference and therefore is related to recognizing textual entailment. Thus, we offered to the community a 5-way student response labeling task, as well as 3-way and 2-way RTE-style tasks on educational data. In addition, a partial entailment task was piloted. We present and compare results from 9 participating teams, and discuss future directions.

1 Introduction

One of the tasks in educational NLP systems is providing feedback to students in the context of exam questions, homework or intelligent tutoring. Much previous work has been devoted to the automated scoring of essays (Attali and Burstein, 2006; Shermis and Burstein, 2013), error detection and correction (Leacock et al., 2010), and classification of texts by grade level (Petersen and Ostendorf, 2009; Sheehan et al., 2010; Nelson et al., 2012). In these applications, NLP methods based on shallow features and supervised learning are often highly effective. However, for the assessment of responses to short-answer questions (Leacock and Chodorow, 2003; Pulman and Sukkarieh, 2005; Nielsen et al., 2008a; Mohler et al., 2011) and in tutorial dialog systems (Graesser et al., 1999; Glass, 2000; Pon-Barry et al., 2004; Jordan et al., 2006; VanLehn et al., 2007; Dzikovska et al., 2010) deeper semantic processing is likely to be appropriate.

Since the task of making and testing a full educational dialog system is daunting, Dzikovska et al. (2012) identified a key subtask and proposed it as a new shared task for the NLP community. Student response analysis (henceforth SRA) is the task of labeling student answers with categories that could

Example 1	QUESTION	You used several methods to separate and identify the substances in mock rocks. How did you separate the salt from the water?
	REF. ANS.	The water was evaporated, leaving the salt.
	STUD. ANS.	The water dried up and left the salt.
Example 2	QUESTION	Georgia found one brown mineral and one black mineral. How will she know which one is harder?
	REF. ANS.	The harder mineral will leave a scratch on the less hard mineral. If the black mineral is harder, the brown mineral will have a scratch.
	STUD. ANS.	The harder will leave a scratch on the other.

Figure 1: Example questions and answers

help a full dialog system to generate appropriate and effective feedback on errors. System designers typically create a repertoire of questions that the system can ask a student, together with reference answers (see Figure 1 for an example). For each student answer, the system needs to decide on the appropriate tutorial feedback, either confirming that the answer was correct, or providing additional help to indicate how the answer is flawed and help the student improve. This task requires semantic inference, for example, to detect when the student answers are explaining the same content but in different words, or when they are contradicting the reference answers.

Recognizing Textual Entailment (RTE) is a series of highly successful challenges used to evaluate tasks related to semantic inference, held annually since 2005. Initial challenges used examples from information retrieval, question answering, machine translation and information extraction tasks (Dagan et al., 2006; Giampiccolo et al., 2008). Later challenges started to explore the applicability and impact of RTE technology on specific application settings such as Summarization and Knowledge Base Population (Bentivogli et al., 2009; Bentivogli et al., 2010; Bentivogli et al., 2011). The SRA Task offers a similar opportunity.

We therefore organized a joint challenge at SemEval-2013, aiming to bring together the educational NLP and the semantic inference communities. The goal of the challenge is to compare approaches for student answer assessment and to evaluate the methods typically used in RTE on data from educational applications.

We present the corpus used in the task (Section 2) and describe the Main task, including educational NLP and textual entailment perspectives and data set creation (Section 3). We discuss evaluation metrics and results in Section 4. Section 5 describes the Pilot task, including data set creation and evaluation results. Section 6 presents conclusions and future directions.

2 Student Response Analysis Corpus

We used the Student Response Analysis corpus (henceforth SRA corpus) (Dzikovska et al., 2012) as the basis for our data set creation. The corpus contains manually labeled student responses to explanation and definition questions typically seen in practice exercises, tests, or tutorial dialogue.

Specifically, given a question, a known correct 'reference answer' and a 1- or 2-sentence 'student answer', each student answer in the corpus is labelled with one of the following judgments:

- 'Correct', if the student answer is a complete and correct paraphrase of the reference answer;
- 'Partially_correct_incomplete', if it is a partially correct answer containing some but not all information from the reference answer;
- 'Contradictory', if the student answer explicitly contradicts the reference answer;
- 'Irrelevant' if the student answer is talking about domain content but not providing the necessary information;
- 'Non_domain' if the student utterance does not include domain content, e.g., "I don't know", "what the book says", "you are stupid".

The SRA corpus consists of two distinct subsets: BEETLE data, based on transcripts of students interacting with BEETLE II tutorial dialogue system (Dzikovska et al., 2010), and SCIENTSBANK data,

based on the corpus of student answers to assessment questions collected by Nielsen et al. (2008b).

The BEETLE corpus consists of 56 questions in the basic electricity and electronics domain requiring 1- or 2- sentence answers, and approximately 3000 student answers to those questions. The SCIENTSBANK corpus contains approximately 10,000 answers to 197 assessment questions in 15 different science domains (after filtering, see Section 3.3)

Student answers in the BEETLE corpus were manually labeled by trained human annotators using a scheme that straightforwardly mapped into SRA annotations. The annotations in the SCIENTSBANK corpus were converted into SRA labels from a substantially more fine-grained scheme by first automatically labeling them using a set of question-specific heuristics and then manually revising them according to the class definitions (Dzikovska et al., 2012). We further filtered and transformed the corpus to produce training and test data sets as discussed in the next section.

3 Main Task

3.1 Educational NLP perspective

The 5-way SRA task focuses on associating student answers with categorical labels that can be used in providing tutoring feedback. Most NLP research on short answer scoring reports agreement with a numeric score (Leacock and Chodorow, 2003; Pulman and Sukkarieh, 2005; Mohler et al., 2011), which is a potential contrast with our task. However, the majority of the NLP work makes use of underlying representations in terms of concepts, so the 5-way task is still likely to mesh well with the available technology. Research on tutorial dialog has emphasized generic methods that use latent semantic analysis or other machine learning methods to determine when text strings express similar concepts (Hu et al., 2003; Jordan et al., 2004; VanLehn et al., 2007; McCarthy et al., 2008). Most of these methods, like the NLP methods, (with the notable exception of (Nielsen et al., 2008a)), are however strongly dependent on domain expertise for the definitions of the concepts. In educational applications, there would be great value in a system that could operate more or less unchanged across a range of domains and question-types, requiring only a question text and a reference answer supplied by the instructional designers. Thus, the 5-way classification task at SemEval was set up to evaluate the feasibility of such answer assessment, either by adapting the existing educational NLP methods to the categorical labeling task or by employing the RTE approaches.

3.2 RTE perspective and 2- and 3-way Tasks

According to the standard definition of Textual Entailment, given two text fragments called *Text* (T) and *Hypothesis* (H), it is said that T entails H if, typically, a human reading T would infer that H is most likely true (Dagan et al., 2006).

In a typical answer assessment scenario, we expect that a correct student answer would entail the reference answer, while an incorrect answer would not. However, students often skip details that are mentioned in the question or may be inferred from it, while reference answers often repeat or make explicit information that appears in or is implied from the question, as in Example 2 in Figure 1. Hence, a more precise formulation of the task in this context considers the entailing text T as consisting of both the original question and the student answer, while H is the reference answer.

We carried out a feasibility study to check how well the entailment judgments in this formulation align with the annotated response assessment, by annotating a sample of the data used in the SRA task with entailment judgments. We found that some answers labeled as "correct" implied inferred or assumed pieces of information not present in the text. These reflected the teachers' assessment of student understanding but would not be considered entailed from the traditional RTE perspective. However, we observed that in most such cases, a substantial part of the hypothesis was still implied by the text. Moreover, answers assigned labels other than "correct" were always judged as "not entailed".

Overall, we concluded that the correlation between assessment judgments of the two types was sufficiently high to consider an RTE approach. The challenge for the textual entailment community was to address the answer assessment task at varying levels of granularity, using textual entailment techniques, and explore how well these techniques can help in this real-world educational setting.

In order to make the setup more similar to pre-

vious RTE tasks, we introduced 3-way and 2-way versions of the task. The data for those tasks were obtained by automatically collapsing the 5-way labels. In the 3-way task, the systems were required to classify the student answer as either (i) *correct*; (ii) *contradictory*; or (iii) *incorrect* (combining the categories partially correct but incomplete, irrelevant and not in the domain from the 5-way classification).

In the two-way task, the systems were required to classify the student answer as either correct or incorrect (combining the categories contradictory and incorrect from the 3-way classification)

3.3 Data Preparation and Training Data

In preparation of the task four of the organizers examined all questions in the SRA corpus, and decided that to remove some of the questions to make the dataset more uniform.

We observed two main issues. First, a number of questions relied on external material, e.g., charts and graphs. In some cases, the information in the reference answer was sufficient to make a reasonable assessment of student answer correctness, but in other cases the information contained in the questions was deemed insufficient and the questions were removed.

Second, some questions in the SCIENTSBANK dataset could have multiple possible correct answers, e.g., a question asking for any example out of two or more unrelated possibilities. Such questions were also removed as they do not align well with the RTE perspective.

Finally, parts of the data were re-checked for reliability. In BEETLE data, a second manual annotation pass was carried out on a subset of questions to check for consistency. In SCIENTSBANK, we manually re-checked the test data. The automatic conversion from the original SCIENTSBANK annotations into SRA labels was not perfectly accurate (Dzikovska et al., 2012). We did not have the resources to check the entire data set. However, four of the organizers jointly hand-checked approximately 100 examples to establish consensus, and then one organizer hand-checked all of the test data set.

3.4 Test Data

We followed the evaluation methodology of Nielsen et al. (2008a) for creating the test data. Since our goal is to support systems that generalize across problems and domains (see Section 3.1), we created three distinct test sets:

1. **Unseen answers (UA)**: a held-out set to assess system performance on the answers to questions contained in the training set (for which the system has seen example student answers). It was created by setting aside a subset if randomly selected learner answers to each question included in the training data set.

2. **Unseen questions (UQ)**: a test set to assess system performance on responses to previously unseen questions but which still fall within the application domains represented in the training data. It was created by holding back all student answers to a subset of randomly selected questions in each dataset.

3. **Unseen domains (UD)**: a domain-independent test set of responses to topics not seen in the training data, available only in the SCIENTSBANK dataset. It was created by setting aside the complete set of questions and answers from three science modules from the fifteen modules in the SCIENTSBANK data.

The final label distribution for train and test data is shown in Table 1.

4 Main Task Results

4.1 Participants

The participants were invited to submit up to three runs in any combination of the tasks. Nine teams participated in the main task, most choosing to attempt all subtasks (5-way, 3-way and 2-way), with 1 team entering only the 5-way and 1 team entering only the 2-way task.

At least 6 (CNGL, CoMeT, CU, BIU, EHUALM, LIMSI) of the 9 systems used some form of syntactic processing, in most cases going beyond parts of speech to dependencies or constituency structure. CNGL emphasized this as an important aspect of the system. At least 5 (CoMeT, CU, EHUALM, ETS UKP) of the 9 systems used a system combination approach, with several components feeding into a final decision made by some form of stacked classifier. The majority of the systems used some kind

label	BEETLE				SCIENTSBANK				
	train (%)	UA	UQ	Test-Total (%)	train (%)	UA	UQ	UD	Test-Total (%)
correct	1665 (0.42)	176	344	520 (0.41)	2008 (0.40)	233	301	1917	2451 (0.42)
pc_inc	919 (0.23)	112	172	284 (0.23)	1324 (0.27)	113	175	986	1274 (0.22)
contra	1049 (0.27)	111	244	355 (0.28)	499 (0.10)	58	64	417	539 (0.09)
irrlvnt	113 (0.03)	17	19	36 (0.03)	1115 (0.22)	133	193	1222	1548 (0.27)
non_dom	195 (0.05)	23	40	63 (0.05)	23 (0.005)	3	0	20	23 (0.004)
incorr-3way	1227 (0.31)	152	231	383 (0.30)	2462 (0.495)	249	368	2228	2845 (0.49)
incorr-2way	2276 (0.58)	263	475	538 (0.59)	2961 (0.596)	307	432	2645	3384 (0.58)

Table 1: Label distribution. Percentages in parentheses. UA, UQ, UD correspond to individual test sets.

of measure of text-to-text similarity, whether the inspiration was LSA, MT measures such as BLEU or in-house methods. These methods were emphasized as especially important by Celi, ETS and SOFTCARDINALITY. These impressions are based on short summaries sent to us by the participants prior to the availability of the full system descriptions. Check the individual system papers for detail.

4.2 Evaluation Metrics

For each evaluation data set (test set), we computed the per-class precision, recall and F_1 score. We also computed three main summary metrics: accuracy, macro-average F_1 and weighted average F_1.

Accuracy is the overall percentage of correctly classified examples.

Macroaverage is the average value of each metric (precision, recall, F_1) across classes, without taking class size into account. It is defined as $1/N_c \sum_c metric(c)$, where N_c is the number of classes (2, 3, or 5 depending on the task). Note that in the 5-way SCIENTSBANK dataset the 'non-domain' class is severely underrepresented, with only 23 examples out of 4335 total (see Table 1). Therefore, we calculated macro-averaged P/R/F_1 over only 4 classes (i.e. excluding the 'non-domain' class) for SCIENTSBANK 5-way data.

Weighted Average (or simply *weighted*) is the average value for each metric weighted by class size, defined as $1/N \sum_c |c| * metric(c)$ where N is the total number of test items and $|c|$ is the number of items labeled as c in gold-standard data.[1]

In general, macro-averaging favors systems that perform well across all classes regardless of class size. Accuracy and weighted average prefer systems that perform best on the largest number of examples, favoring higher performance on the most frequent classes. In practice, only a small number of the systems were ranked differently by the different metrics. We discuss this further in Section 4.7. Results for all metrics are available online, and this paper focuses on two metrics for brevity: weighted and macro-average F_1 scores.

4.3 Results

The evaluation results for all metrics and all participant runs are provided online.[2] The tables in this paper present the F_1 scores for the best system runs. Results are shown separately for each test set (TS), with the simple mean over the five TSs reported in the final column.

We used two baselines: the majority (most frequent) class baseline and a lexical overlap baseline described in detail in (Dzikovska et al., 2012). The performance of the baselines is presented jointly with system scores in the results tables.

For each participant, we report the single run with the best average TS performance, identified by the subscript in the run title, with the exception of ETS. With all other participants, there was almost always one run that performed best for a given metric on *all* the TSs. In the small number of cases where another run performed best on a given TS, we instead report that value and indicate its run with a subscript (these changes never resulted in meaningful changes in the performance rankings). ETS, on the other hand, sub-

[1] This metric is called *microaverage* in (Dzikovska et al., 2012). However, *microaverage* is used to define a different metric in tasks where more than one label can be associated with each data item (Tsoumakas et al., 2010). therefore, we use *weighted* average to match the terminology used by the `Weka` toolkit. The micro-average precision, recall and F_1 computed using the multi-label metric are all equal and mathematically equivalent to accuracy.

[2] `http://bit.ly/11a7QpP`

Dataset:	BEETLE		SCIENTSBANK			
Run	UA	UQ	UA	UQ	UD	Mean
$CELI_1$	0.423	0.386	0.372	0.389	0.367	0.387
$CNGL_2$	*0.547*	0.469	0.266	0.297	0.294	0.375
$CoMeT_1$	***0.675***	0.445	***0.598***	0.299	0.252	0.454
$EHUALM_2$	*0.566*	0.416_3	$*0.525*_3$	0.446	*0.437*	0.471
ETS_1	*0.552*	*0.547*	*0.535*	***0.487***	*0.447*	0.514
ETS_2	***0.705***	***0.614***	***0.625***	0.356	*0.434*	0.547
$LIMSIILES_1$	0.505	0.424	0.419	***0.456***	*0.422*	0.445
$SoftCardinality_1$	*0.558*	0.450	*0.537*	***0.492***	***0.471***	0.502
$UKP\text{-}BIU_1$	0.448	0.269	***0.590***	0.397_2	0.407	0.418
Median	0.552	0.445	0.535	0.397	0.422	0.454
Baselines:						
Lexical	0.483	0.463	0.435	0.402	0.396	0.436
Majority	0.229	0.248	0.260	0.239	0.249	0.245

Table 2: Five-way task weighted-average F_1

Dataset:	BEETLE **5**way		SCIENTSBANK **4**way			
Run	UA	UQ	UA	UQ	UD	Mean
$CELI_1$	0.315	0.300	0.278	0.286	0.269	0.290
$CNGL_2$	0.431	0.382	0.252	0.262	0.239	0.313
$CoMeT_1$	***0.569***	0.300	***0.551***	0.201	0.151	0.354
$EHUALM_2$	*0.526*	0.370_3	$*0.447*_3$	**0.353**	*0.340*	0.407
ETS_1	0.444	0.461	*0.467*	***0.372***	*0.334*	0.416
ETS_2	***0.619***	***0.552***	***0.581***	0.274	*0.339*	0.473
$LIMSIILES_1$	0.327	0.280	0.335	**0.361**	*0.337*	0.328
$SoftCardinality_1$	0.455	0.436	*0.474*	***0.384***	***0.375***	0.425
$UKP\text{-}BIU_1$	0.423	0.285	***0.560***	0.325_2	*0.348*	0.388
Median	0.444	0.370	0.467	0.325	0.337	0.388
Baselines:						
Lexical	0.424	0.414	0.375	0.329	0.311	0.371
Majority	0.114	0.118	0.151	0.146	0.148	0.135

Table 3: Five-way task macro-average F_1

mitted results for systems that were substantially different from one another, with performance varying from being the top rank to nearly the lowest. Hence, it seemed more appropriate to report two separate runs.[3] In the rest of the discussion *system* is used to refer to a row in the tables as just described.

Systems with performance that was not statistically different from the best results for a given TS are all shown in **bold** (significance was not calculated for the TS mean). Systems with performance statistically better than the lexical baseline are displayed in *italics*. Statistical significance tests were conducted using approximate randomization test (Yeh, 2000) with 10,000 iterations; $p \leq 0.05$ was considered statistically significant.

4.4 Five-way Task

The results for the five-way task are shown in Tables 2 and 3.

Comparison to baselines All of the systems performed substantially better than the majority class baseline ("correct" for both BEETLE and SCIENTSBANK), on average exceeding it on the TS mean by 0.21 on the weighted F_1 and 0.25 on the macro-average F_1. Six systems outperformed the lexical baseline on the mean TS results for the weighted F_1 and five for the macro-average F_1. Nearly all of the top results on a given TS (shown in bold in the tables) were statistically better than corresponding lexical baselines according to significance tests (indicated by italics in the tables).

Comparing UA and UQ/UD performance The BEETLE UA (BUA) and SCIENTSBANK UA (SUA) test sets represent questions with example answers in training data, while the UQ and UD test sets represent transfer performance to new questions and new domains respectively.

The top performers on UA test sets were $CoMeT_1$ and ETS_2, with the addition of $UKP\text{-}BIU_1$ on SUA. However, there was not a single best performer on UQ and UD sets. ETS_2 performed statistically better than all other systems on BEETLE UQ (BUQ), but it performed *statistically worse* than the lexical baseline on SCIENTSBANK UQ (SUQ), resulting in no overlap in the top performing systems on the two UQ test sets. $SoftCardinality_1$ performed statistically better than all other systems on SUD and was among the three or four top performers on SUQ, but was not a top performer on the other three TSs, generally not performing statistically better than the lexical baseline on the BEETLE TSs.

Group performance The two UA TSs had more systems that performed statistically better than the lexical baseline (generally six systems) than did the UQ TSs where on average only two systems performed statistically better than the lexical baseline. Over twice as many systems outperformed the lexical baseline on UD as on the UQ TSs. The top performing systems according to the macro-average F_1 were nearly identical to the top performing systems according to the weighted F_1.

[3] In a small number of cases, ETS's third run performed marginally better, see full results online.

4.5 Three-way Task

The results for the three-way task are shown in Tables 4 and 5.

Comparison to baselines All of the systems performed substantially better than the majority baseline ("correct" for BEETLE and "incorrect" for SCIENTSBANK), on average exceeding it on the TS mean by 0.28 on the weighted F_1 and 0.31 on the macro-average F_1. Five of the eight systems outperformed the lexical baseline on the mean TS results for the weighted F_1 and five on the macro-average F_1, and all top systems outperformed the lexical baseline with statistical significance.

Comparing UA and UQ/UD performance The top performers on both BUA and SUA were CoMeT$_1$ and ETS$_2$. As for the 5-way task there was no single best performer for UQ and UD sets, and no overlap in top performing systems on BUQ and SUQ test sets, with ETS$_2$ being the top performer on BUQ, but statistically worse than the baseline on SUQ and SUD. On the weighted F_1, SoftCardinality$_1$ performed statistically better than all other systems on SUD and was among the two statistically best systems on SUQ, but was not a top performer on BUQ or BUA/SUA TSs. On the macro-average F_1, UKP-BIU$_1$ became one of the statistically best performers on all SCIENTSBANK TSs but, along with SoftCardinality$_1$, never performed statistically better than the lexical baseline on the BEETLE TSs.

Group performance With the exception of SUA, only around two systems performed statistically better than the lexical baseline on each TS. The top performing systems were nearly the same according to the weighted F_1 and the macro-average F_1.

4.6 Two-way Task

The results for the two-way task are shown in Table 6. Because the labels are roughly balanced in the two-way task, the results on the weighted and macro-average F_1 are very similar and the top performing systems are identical. Hence this section will focus only on the macro-average F_1.

As in the previous tasks, all of the systems performed substantially better than the majority baseline ("incorrect" for all sets), on average exceeding it on the TS mean by 0.25 on the weighted F_1 and 0.30 on the macro-average F_1. However, just four of the nine systems in the two-way task outperformed the lexical baseline on the mean TS results. In fact, the average performance fell below the lexical baseline. The differences in the macro-average F_1 between the top results on a SCIENTSBANK TS and the corresponding lexical baselines were all statistically significant. Two of the top results on BUA were not statistically better than the lexical baseline, and all systems performed below the baseline on BUQ.

Dataset:	BEETLE		SCIENTSBANK			
Run	UA	UQ	UA	UQ	UD	Mean
CELI$_1$	0.519	0.463	0.500	0.555	0.534	0.514
CNGL$_2$	0.592	0.471	0.383	0.367	0.360	0.435
CoMeT$_1$	***0.728***	0.488	***0.707***	0.522	0.550	0.599
ETS$_1$	0.619	0.542	*0.603*	***0.631***	*0.600*	0.599
ETS$_2$	***0.723***	***0.597***	***0.709***	0.537	0.505	0.614
LIMSIILES$_1$	0.587	0.454	0.532	0.553	0.564	0.538
SoftCardinality$_1$	0.616	0.451	*0.647*	***0.634***	***0.620***	0.594
UKP-BIU$_1$	0.472	0.313	*0.670*	*0.573*	0.577$_2$	0.521
Median	0.604	0.467	0.625	0.554	0.557	0.566
Baselines:						
Lexical	0.578	0.500	0.523	0.520	0.554	0.535
Majority	0.229	0.248	0.260	0.239	0.249	0.245

Table 4: Three-way task weighted-average F_1

Dataset:	BEETLE		SCIENTSBANK			
Run	UA	UQ	UA	UQ	UD	Mean
CELI$_1$	0.494	0.441	0.373	0.412	0.415	0.427
CNGL$_2$	0.567	0.450	0.330	0.308	0.311	0.393
CoMeT$_1$	***0.715***	0.466	***0.640***	0.380	0.404	0.521
ETS$_1$	0.592	*0.521*	*0.477*	***0.459***	*0.439*	0.498
ETS$_2$	***0.710***	***0.585***	***0.643***	0.389	0.367	0.539
LIMSIILES$_1$	0.563	0.431	0.404	0.409	0.429	0.447
SoftCardinality$_1$	0.596	0.439	*0.555*	***0.469***	***0.486***	0.509
UKP-BIU$_1$	0.468	0.333	***0.620***	***0.458***	***0.487***	0.473
Median	0.580	0.446	0.516	0.411	0.422	0.485
Baselines:						
Lexical	0.552	0.477	0.405	0.390	0.416	0.448
Majority	0.191	0.197	0.201	0.194	0.197	0.196

Table 5: Three-way task macro-average F_1

4.7 Discussion

All of the systems consistently outperformed the most frequent class baseline. Beating the lexical overlap baseline proved to be more challenging, being achieved by just over half of the results with about half of those being statistically significant improvements. This underscores the fact that there is still a considerable opportunity to improve student

Dataset:	BEETLE		SCIENTSBANK			
Run	UA	UQ	UA	UQ	UD	Mean
$CELI_1$	0.640	0.656	0.588	0.619	0.615	0.624
$CNGL_2$	**0.800**	0.666	0.591_1	0.561	0.556	0.635
$CoMeT_1$	***0.833***	**0.695**	***0.768***	0.579	*0.670*	0.709
CU_1	0.778	**0.689**	0.603	0.638	*0.673*	0.676
ETS_1	**0.802**	**0.720**	*0.705*	*0.688*	*0.683*	0.720
ETS_2	***0.833***	**0.702**	***0.762***	0.602	0.543	0.688
$LIMSIILES_1$	0.723	0.641	0.583	0.629	0.648	0.645
$SoftCardinality_1$	0.774	0.635	*0.715*	***0.737***	***0.705***	0.713
$UKP\text{-}BIU_1$	0.608	0.481	0.726	0.669	0.666_2	0.630
Median	0.778	0.666	0.705	0.629	0.666	0.676
Baselines:						
Lexical	0.788	**0.725**	0.617	0.630	0.650	0.682
Majority	0.375	0.367	0.362	0.371	0.367	0.368

Table 6: Two-way task macro-average F_1

response assessment systems.

The set of top performing systems on the weighted F_1 for a given TS were also always in the top on the macro-average F_1, but a small number of additional systems joined the top performing set on the macro-average F_1. Specifically, one, three, and two results joined the top set in the five-way, three-way, and two-way tasks, respectively. In principle, the metrics could differ substantially, because of the treatment of minority classes, but in practice they rarely did. Only one pair of participants swap adjacent TS mean rankings on the macro-average F_1 relative to the weighted F_1 on the two-way task. On the five-way task, two pairs swap rankings and another participant moved up two positions in the ranking, ending at the median value.

Most (28/34) rank changes were only one position and most (21/34) were in positions at or below the median ranking. In the five-way task, a pair of systems, $UKP\text{-}BIU_1$ and ETS_1, had a meaningful performance rank swap on the macro-average F_1 relative to the weighted F_1 on the UD test set. Specifically, $UKP\text{-}BIU_1$ moved up four positions from rank 6, where it was not statistically better than the lexical baseline, to the second best performance.

Not surprisingly, performance on UA was substantially higher than on UQ and UD, since the UA is the only set which contains questions with example answers in training data. Performance on BUA was usually better than performance on SUA, most likely because BUA contains more similar questions and answers, focusing on a single science area, Electricity and Magnetism, compared to 12 distinct science topics in SUA). In addition, the BEETLE study participants may have used simpler language, since they were aware that they were talking to a computer system instead of writing down answers for human teachers to assess as in SCIENTSBANK.

Performance on BUQ versus SUQ was much more varied, presumably since there was no direct training data for either TS. For the five-way task, the best performance on the weighted F_1 measure for BUQ is 0.09 below the best result for BUA and the analogous decrease from SUA to SUQ is 0.13, with an additional 0.02 drop on SUD. On the two-way task, the best weighted F_1 for BUQ drops 0.11 from the best BUA value, but the decrease from SUA to SUQ is just 0.03, with another 0.03 drop to SUD. While the drop in performance is fairly similar from BUA to BUQ on all tasks and either metric, the decrease from SUA to SUQ seems to potentially be dependent on the task, ranging from 0.13 on the five-way task to 0.08 on the three-way task and 0.03 on the two-way task.

5 Pilot Task on Partial Entailment

The SCIENTSBANK corpus was originally developed to assess student answers at a very fine-grained level and contains additional annotations that break down the answers into "facets", or low-level concepts and relationships connecting them (henceforth, SCIENTSBANK Extra). This annotation aims to support educational systems in recognizing when specific parts of a reference answer are expressed in the student answer, even if the reference answer is not entailed as a whole (Nielsen et al., 2008b). The task of recognizing such partial entailment relationships may also have various uses in applications such as summarization or question answering, but it has not been explored in previous RTE challenges.

Therefore, we proposed a pilot task on partial entailment, in which systems are required to recognize whether the semantic relation between specific parts of the Hypothesis is expressed by the Text, directly or by implication, even though entailment might not be recognized for the Hypothesis as a whole, based on the SCIENTSBANK facet annotation.

Each reference answer in SCIENTSBANK data is broken down into facets, where a facet is a triplet

consisting of two key terms (both single words and multi-words, e.g. *carbon dioxide, each other, burns out*) and a relation linking them, as shown in Figure 2. The student answers were then annotated with regards to each reference answer facet in order to indicate whether the facet was (i) expressed, either explicitly or by assumption or easy inference; (ii) contradicted; or (iii) left unaddressed. Considering the SCIENTSBANK reference answers as Hypotheses, the facets capture their atomic components, and facet annotations may correspond to the judgments on the sub-parts of the H which are entailed by T.

We carried out a feasibility study to explore this idea and to verify how well the facet annotations align with traditional entailment judgments. We focused on the reference answer facets labeled in the gold standard annotation as *Expressed* or *Unaddressed*. The working hypothesis was that *Expressed* labels assigned in SCIENTSBANK annotations corresponded to *Entailed* judgments in traditional textual entailment annotations, while *Unaddressed* labels corresponded to *No-entailment* judgments.

Similarly to the feasibility study reported in Section 3.2, we concluded that the correspondence between educational labels and entailment judgments was not perfect due to the difference in educational and textual entailment perspectives. Nevertheless, the two classes of assessment appeared to be sufficiently well correlated so as to offer a good testbed for partial entailment in a natural setting.

5.1 Task Definition

Given (i) a text T, made up of a Question and a Student Answer; (ii) a hypothesis H, i.e. the Reference Answer for that question and (iii) a facet, i.e. a pair of key terms in H, the task consists of determining whether T expresses, either directly or by implication, the same relationship between the facet words as in H. In other words, for each of H's facets the system assign one of the following judgments: *Expressed*, if the Student Answer expresses the same relationship between the meaning of the facet terms as in H; *Unaddressed*, if it does not.

Consider the example shown in Figure 2. For facet 3, the system must decide whether the same relation between the two terms *'contains'* and *'seeds'* in H (the reference answer) is expressed, explicitly or implicitly, in T (the combination of question and student response). If the student answer is '*The part of a plant you are observing is a fruit if it has seeds.*', the answer to the question is *'yes'* and the correct judgment is *'Expressed'*. But if the student says '*My rule is has to be sweet.*', T does not express the same semantic relationship between *'contains'* and *'seeds'* exhibited in H, thus the correct judgment is *'Unaddressed'*. Note that even though this is an exercise in textual entailment, student response assessment labels were used instead of traditional entailment judgments, due to the partial mismatch between the two assessment classes found in the feasibility study.

5.2 Dataset

We used a subset of the SCIENTSBANK Extra corpus (Nielsen et al., 2008b) with the same problematic questions filtered out as the main task (see Section 3.3). We further filtered out all the student answer facets which were labeled other than *'Expressed'* or *'Unaddressed'* in the gold standard annotation; the facets in which the relationship between the two key terms, as classified in the manual annotation, proved to be problematic to define and judge, namely *Topic, Agent, Root, Cause, Quantifier, Neg*; and inter-propositional facets, i.e. facets that expressed relations between higher-level propositions. Finally, the facet relations were removed from the dataset, leaving the relationship between the two facet terms unspecified so as to allow a more fuzzy approach to the inference problem posed by the exercise.

We used the same training/test split as reported in Section 3.4. The training set created from the Training SCIENTSBANK Extra corpus contains 13,145 reference answer facets, 5,939 of which were labeled as *'Expressed*' in the student answers and 7,206 as *'Unaddressed'*. The Test set was created from the SCIENTSBANK Extra unseen data and is divided into the same subsets as the main task (Unseen Answers, Unseen Questions and Unseen Domains). It contains 16,263 facets total, with 5,945 instances labeled as *'Expressed'*, and 10,318 labeled as '*Unaddressed*'.

5.3 Evaluation Metrics and Baselines

The metrics used in the Pilot task were the same as in the Main task, i.e. Overall Accuracy, Macroaverage

QUESTION:	What is your "rule" for deciding if the part of a plant you are observing is a fruit?
REFERENCE ANSWER:	If a part of the plant contains seeds, that part is the fruit.
FACET 1:	Relation *NMod_of* Term1 *part* Term2 *plant*
FACET 2:	Relation *Theme* Term1 *contains* Term2 *part*
FACET 3:	Relation *Material* Term1 *contains* Term2 *seeds*
FACET 4:	Relation *Be* Term1 *fruit* Term2 *part*

Figure 2: Example of facet annotations supporting the partial entailment task

Run	**UA**	**UQ**	**UD**	**UA**	**UQ**	**UD**
	Weighted Averaged			*Macro Average*		
Run1	0.756	0.71	0.76	0.7370	0.686	0.755
Run 2	0.782	0.765	0.816	0.753	0.73	0.804
Run 3	0.744	0.733	0.77	0.719	0.7050	0.761
Baseline	0.54	0.547	0.478	0.402	0.404	0.384

Table 7: Weighted-average and macro-average F_1 scores (*UA: Unseen Answers; UQ: Unseen Questions; UD Unseen Domains*)

.

and Weighted Average Precision, Recall and F_1, and computed as described in Section 4.2. We used only a majority class baseline, which labeled all facets as *'Unaddressed'*. Its performance is presented in Section 5.4 jointly with the system results.

5.4 Participants and results

Only one participant, UKP-BIU, participated in the Partial Entailment Pilot task. The UKP-BIU system is a hybrid of two semantic relationship approaches, namely (i) computing semantic textual similarity by combining multiple content similarity measures (Bär et al., 2012), and (ii) recognizing textual entailment with BIUTEE (Stern and Dagan, 2011). The two approaches are combined by generating indicative features from each one and then applying standard supervised machine learning techniques to train a classifier. The system used several lexical-semantic resources as part of the BIUTEE entailment system, together with SCIENTSBANK dependency parses and ESA semantic relatedness indexes from Wikipedia.

The team submitted the maximum allowed of 3 runs. Table 7 shows Weighted Average and Macro Average F_1 scores respectively, also for the majority baseline. The system outperformed the majority baseline on both metrics. The best performance was observed on Run 2, with the highest results on the Unseen Domains test set.

6 Conclusions and Future Work

The Joint Student Response Analysis and 8th Recognizing Textual Entailment challenge has proven to be a useful, interdisciplinary task using a realistic dataset from the educational domain. In almost all cases the best systems significantly outperformed the lexical overlap baseline, sometimes by a large margin, showing that computational linguistics approaches can contribute to educational tasks. However, the lexical baseline was not trivial to beat, particularly in the 2-way task. These results are consistent with similar findings in previous RTE exercises. Moreover, there is still significant room for improvement in the absolute scores, reflecting the interesting challenges that both educational data and RTE tasks present to computational linguistics.

The educational setting places new stresses on semantic inference technology because the educational notion of *'Expressed'* and the RTE notion of *'Entailed'* are slightly different. This raises the educational question of whether RTE can work in this setting, and the RTE question of whether this setting is meaningful for evaluating RTE system performance. The experimental results suggests that the answer to both questions is 'yes', a significant finding for both educators and RTE technologists going forward.

The Pilot task, aimed at exploring notions of partial entailment, so far not explored in the series of RTE challenges, has proven to be an interesting, though challenging exercise. The novelty of the task, namely performing textual entailment not on a pair of full texts, but between a text and a hypothesis consisting of a pair of words, may have represented a more complex task than expected for some textual entailment engines. Despite this, the encouraging results obtained by the team which carried out the exercise has shown that this partial entailment task is worthy of further investigation.

Acknowledgments

The research reported here was supported by the US ONR award N000141010085 and by the Institute of Education Sciences, U.S. Department of Education, through GrantR305A120808 to the University of North Texas. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The RTE-related activities were partially supported by the Pascal-2 Network of Excellence, ICT-216886-NOE. We would also like to acknowledge the contribution of Alessandro Marchetti and Giovanni Moretti from CELCT to the organization of the challenge.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning, and Assessment*, 4(3), February.

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the 6th International Workshop on Semantic Evaluation, held in conjunction with the 1st Joint Conference on Lexical and Computational Semantics*, pages 435–440, Montreal, Canada, June.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernando Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of Text Analysis Conference (TAC) 2009*.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2010. Thesixth PASCAL recognizing textual entailment challenge. In *Notebook papers and results, Text Analysis Conference (TAC)*.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2011. The seventh PASCAL recognizing textual entailment challenge. In *Notebook papers and results, Text Analysis Conference (TAC)*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 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*. Springer.

Myroslava O. Dzikovska, Johanna D. Moore, Natalie Steinhauser, Gwendolyn Campbell, Elaine Farrow, and Charles B. Callaway. 2010. Beetle II: a system for tutoring and computational linguistics experimentation. In *Proc. of ACL 2010 System Demonstrations*, pages 13–18.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proc. of 2012 Conference of NAACL: Human Language Technologies*, pages 200–210.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. 2008. The fourth PASCAL recognizing textual entailment challenge. In *Proceedings of Text Analysis Conference (TAC) 2008*, Gaithersburg, MD, November.

Michael Glass. 2000. Processing language input in the CIRCSIM-Tutor intelligent tutoring system. In *Papers from the 2000 AAAI Fall Symposium, Available as AAAI technical report FS-00-01*, pages 74–79.

A. C. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, and R. Kreuz. 1999. Autotutor: A simulation of a human tutor. *Cognitive Systems Research*, 1:35–51.

Xiangen Hu, Zhiqiang Cai, Max Louwerse, Andrew Olney, Phanni Penumatsa, and Art Graesser. 2003. A revised algorithm for latent semantic analysis. In *Proceedings of the 18th International Joint Conference on Artificial intelligence (IJCAI'03)*, pages 1489–1491, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Pamela W. Jordan, Maxim Makatchev, and Kurt VanLehn. 2004. Combining competing language understanding approaches in an intelligent tutoring system. In *Proc. of Intelligent Tutoring Systems Conference*, pages 346–357.

Pamela Jordan, Maxim Makatchev, Umarani Pappuswamy, Kurt VanLehn, and Patricia Albacete. 2006. A natural language tutorial dialogue system for physics. In *Proc. of 19th Intl. FLAIRS conference*, pages 521–527.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel R. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Philip M. McCarthy, Vasile Rus, Scott A. Crossley, Arthur C. Graesser, and Danielle S. McNamara. 2008. Assessing forward-, reverse-, and average-entailment indices on natural language input from the intelligent tutoring system, iSTART. In *Proc. of 21st Intl. FLAIRS conference*, pages 165–170.

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.

Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical report, Student Achievement Partners. `http://www.ccsso.org/Documents/2012/Measures%20ofText%20Difficulty_fina%l.2012.pdf`.

Rodney D. Nielsen, Wayne Ward, and James H. Martin. 2008a. Learning to assess low-level conceptual understanding. In *Proc. of 21st Intl. FLAIRS Conference*, pages 427–432.

Rodney D. Nielsen, Wayne Ward, James H. Martin, and Martha Palmer. 2008b. Annotating students' understanding of science concepts. In *Proceedings of the Sixth International Language Resources and Evaluation Conference, (LREC08)*, Marrakech, Morocco.

Sarah Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer, Speech and Language*, 23(1):89–106.

Heather Pon-Barry, Brady Clark, Karl Schultz, Elizabeth Owen Bratt, and Stanley Peters. 2004. Advantages of spoken language interaction in dialogue-based intelligent tutoring systems. In *Proc. of ITS-2004 Conference*, pages 390–400.

Stephen G Pulman and Jana Z Sukkarieh. 2005. Automatic short answer marking. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 9–16, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Kathryn M. Sheehan, Irene Kostin, Yoko Futagi, and Michael Flor. 2010. Generating automated text complexity classifications that are aligned with targeted text complexity standards. Technical Report RR-10-28, Educational Testing Service.

Mark D. Shermis and Jill Burstein, editors. 2013. *Handbook on Automated Essay Evaluation: Current Applications and New Directions*. Routledge.

Asher Stern and Ido Dagan. 2011. A confidence model for syntactically-motivated entailment proofs. In *Recent Advances in Natural Language Processing (RANLP 2011)*, pages 455–462, Hissar, Bulgaria, September.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2010. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US.

Kurt VanLehn, Pamela Jordan, and Diane Litman. 2007. Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In *Proc. of SLaTE Workshop on Speech and Language Technology in Education*, Farmington, PA, October.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational linguistics (COLING 2000)*, pages 947–953, Stroudsburg, PA, USA. Association for Computational Linguistics.