Annotating Entailment Relations for Shortanswer Questions

Simon Ostermann, Andrea Horbach, Manfred Pinkal
Department of Computational Linguistics, Saarland University, Saarbrücken, Germany
(simono|andrea|pinkal)@coli.uni-saarland.de

Abstract

This paper presents an annotation project that explores the relationship between textual entailment and short answer scoring (SAS). We annotate entailment relations between learner and target answers in the Corpus of Reading Comprehension Exercises for German (CREG) with a fine-grained label inventory and compare them in various ways to correctness scores assigned by teachers. Our main finding is that although both tasks are clearly related, not all of our entailment tags can be directly mapped to SAS scores and that especially the area of partial entailment covers instances that are problematic for automatic scoring and need further investigation.

1 Introduction

Reading comprehension exercises are a standard task in foreign language education: Students read a text in the language they are learning and answer questions about it. With the advent of computer-based language learning courses, the automatic scoring of such shortanswer questions has become an important research topic (for an overview see Burrows et al. (2015); Ziai et al. (2012)), not only for reading and listening comprehension in the context of foreign language learning, but also e.g. in science questions for native speaker students.

It has been often noted that the SAS task is related to the task of recognizing textual entailment (RTE, e.g. Mohler et al. (2011), Sukkarieh and Blackmore (2009), Dzikovska et al. (2013b)). RTE is the task to decide whether there is an inference relation between two texts; in the case of SAS, these texts are the learner answer (LA), given by a student, and a teacher-specified target answer (TA, i.e. a sample solution). An entailment relation between two texts *A* and *B* is given if people reading *A* and *B* would infer that whenever *A* is true, *B* is most likely true as well (Dagan et al., 2013).

Consider the following example:[1]

(1) **Q:** Why did Julchen come to the kitchen?
TA: She came to the kitchen because of the noise her parents made.
LA: She came to the kitchen because Mr. and Mrs. Muschler became out of breath from laughing.

In this example, the LA textually (but not logically) entails the TA. In a strictly logical sense of entailment, laughing until you are out of breath does not entail making noise. However, it seems plausible to many people that laughing in that way makes a lot of noise. Such a learner answer that is more specific than the target answer – and thus entails the target answer – is likely to be scored as correct by a teacher.

In some aspects, SAS for reading comprehension in a language learning scenario differs from a standard textual entailment scenario: Whereas in standard RTE, two texts are compared, in the SAS scenario the additional context of the question has to be accounted for in terms of information structure and resolution of anaphora and ellipses. Additionally, when processing learner language one often has to deal with ungrammatical sentences and orthographical variance that are challenging for many NLP tools, up to the extent that it is sometimes difficult to understand what the learner wanted to express with an answer (the so-called *target hypothesis*).

In this study, we want to explicitly assess the relation between RTE labels and correctness scores assigned by teachers. We assume that they are related, but we expect that the relation is not a direct

[1] All examples are taken from the CREG corpus and translated by the authors preserving linguistic errors whenever possible.

Beijing, China, July 31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

mapping. We expect, for example, that, if a LA entails a TA and vice versa at the same time, i.e. if they are paraphrases, then the LA will probably be scored as correct by a teacher. On the other hand, the fact that there is only some partial conceptual overlap between a LA and a TA does not constitute entailment, but is in some instances enough for an answer to be scored as correct by a teacher.

We present in this paper the first part of an annotation project that aims at investigating the relationship between SAS and RTE and that compares existing binary correctness scores annotated by teachers to RTE annotations that have been conducted without the correctness or quality of the learner answer in mind. (In future work, we will also look at the relation between reading texts and learner answers.)

Understanding these relations better will potentially help us to leverage techniques from RTE for the task of SAS in a more efficient way and to shed light on the the way teachers score shortanswer questions.

This paper makes the following contributions:

- We provide a fine-grained annotation of the Corpus of Reading Comprehension Exercises in German (CREG) (Ott et al., 2012) with 7 textual entailment labels that specify the entailment relations between learner answers and target answers.
- We provide an evaluation of our annotations that compares how our label distribution corresponds to the distribution of binary teacher scores in CREG.
- State of the art binary scoring approaches label only about 86% of the corpus correctly (Hahn and Meurers, 2012). In order to understand the challenges of automatic scoring better, we evaluate which instances in terms of our entailment annotation labels are most problematic for automatic scoring with a binary label.
- We will further explore the relation between textual entailment and SAS by comparing, how well features from shortanswer scoring tasks can be used to learn our classification.

2 Related Work

Recognizing textual entailment and automatic shortanswer scoring are two related tasks in which text pairs are labeled with the relation between them:

The RTE task in its original formulation (Dagan and Glickman, 2004) is a binary classification task deciding whether a *text t* entails a *hypothesis h*. The two-way task has been extended to a 3-way task involving the labels *Entailed*, *Contradicted* and *Unknown* (Giampiccolo et al., 2007). Annual RTE shared tasks led to a growing community with a large number of approaches, cf. (Dagan et al., 2013). MacCartney and Manning (2009a) proposed an extension of the classification schema to a much more fine-grained inventory of 7 semantic relations that expresses additional concepts such as *equivalence* and *reverse entailment* and also inspired our label set.

In SAS, the task is to assign a student answer a score that specifies the correctness of the answer. Many approaches to SAS compare learner answers given by a student to target answers specified by a teacher and rely on some measure of surface or semantic overlap between them (e.g. Bailey and Meurers (2008); Meurers et al. (2011); Mohler et al. (2011)) or measure whether teacher-specified aspects of a correct answer (so-called facets) are addressed in the learner answer (Nielsen et al., 2008).

In SAS corpora, the label for an answer is a binary score, stating whether the LA is correct or incorrect. Some data sets also provide annotations with points from an integer scale (e.g. Mohler and Mihalcea (2009) or the kaggle SAS competition [2]). Other data use more meaningful diagnostic labels such as Ott et al. (2012) and Bailey and Meurers (2008) that provide feedback to the learner.

In our study, we primarily rely on binary correctness scores for our comparisons. For the RTE task, we see LA and TA as text and hypothesis and expect that entailment will correlate with correctness: While a LA paraphrasing the TA should definitely count as correct, making the LA more specific should not make it incorrect either. However, omitting crucial information from the TA will potentially make the LA incorrect.

SemEval-2013 task 7 (Dzikovska et al., 2013b) took a first step in bringing together the RTE and the SAS community in a task to label student answers to explanation and definition questions with 5 RTE-labels. The data set used there (Dzikovska et al., 2012) focuses on science questions (Nielsen

[2] https://www.kaggle.com/c/asap-sas/data

et al., 2008) and physics questions from tutorial dialogues (Dzikovska et al., 2010), i.e. in contrast to our scenario they deal with native speakers – thus avoiding problems in processing learner language – and the questions do not refer to a specific reading text. Most importantly, our perspective on the relation between SAS and RTE also differs from the SemEval definition: The SemEval task uses RTE labels that are constructed from labels assigned by teachers as meaningful feedback to students. They assume that there is a direct mapping from RTE labels to binary teacher scores and construct their binary data set from collapsing those labels. Their approach is backed up by a small feasibility study that shows the correspondence of the RTE and SAS label sets in their setting. In our study, we consider RTE and SAS as different tasks and want to explore their relation. We therefore compare labeling from a RTE perspective and scoring from a teacher's point of view.

Both within the context of the SemEval task and already before, RTE approaches have been used for SAS. Levy et al. (2013) try to recognize *partial entailment* based on the facet approach by Nielsen et al. (2008) and aim at exploring its possible impact on recognizing full entailment relations on learner data as part of the SemEval-2013 task 7. Consequently, they also see the tasks of RTE and SAS as equivalent. In contrast to this, Mohler et al. (2011) present a SAS approach that uses techniques from RTE (e.g. a dependency graph matching approach, cf. Haghighi et al. (2005)), but clearly point out that although their system uses those methods, it cannot be seen as RTE system.

3 Annotations

3.1 Data Set

We use the Corpus of Reading Comprehension Exercises in German (CREG) (Ott et al., 2012), a prominent resource for shortanswer scoring data for German as a Foreign Language, as basis for our annotations. It contains 1032 learner answers (half of which have been scored as correct, the other half as incorrect by teachers), answering 177 different questions about a total of 32 texts together with teacher-specified target answers. Sometimes the corpus contains more than one target answer for a question. In such cases the corpus provides annotations that link every learner answer to exactly one best-fitting target answer. We use these annotations in creating LA-TA pairs for our annotations.

3.2 LA-TA Annotation Scheme

The aim of this first part of our annotation project is to investigate the textual entailment relations between TAs and LAs.

We use an extended and slightly modified version of the entailment classes proposed by MacCartney and Manning (2009b) that we adapted to our scenario of answer pairs, instead of self-contained text pairs (or even sentence pairs as in the early RTE tasks). Our labels are as follows:

paraphrase: TA and LA are paraphrases, i.e. express the same semantic content.

entailment: The LA textually entails the TA, i.e. it is more specific than the TA.

reverse entailment: The TA textually entails the LA.

partial entailment: There is a semantic overlap between TA and LA but there is no clear entailment relation in any direction[3].

contradiction: LA and TA are mutually exclusive, i.e. they cannot both be true at the same point in time.

topical non-entailment: The LA is in principle a valid answer to the question (it is *on-topic*) but there is no semantic overlap to the TA that would qualify it for one of the other entailment categories.

off-topic: While answers with any of the previous labels addressed the right question, i.e. were *on-topic*, for this label, the LA is *off-topic*[4] , i.e. it either answers a different question or is a non-answer and therefore cannot be compared to the TA.

Table 1 gives examples for all entailment types.

Note that our label set is a refinement of the classical 3-way entailment definition: While our *entailment* and *paraphrase* labels (if considering the LA to be the *text* and the TA to be the *hypothesis* in the classical RTE problem) correspond to *entailment* in the 3-way task, and our *contradiction* label directly corresponds to *contradictions* in classical RTE, all our other labels refine the *unknown* class.

[3] The *partial entailment* relation is discussed in more detail in Nielsen et al. (2009) and Levy et al. (2013)

[4] Note that this label is similar to the notion of incongruence introduced by von Stechow (1990).

Label	Question	Target Answer	Learner Answer
paraphrase	How much does BA earn monthly?	BA earns less than 300 Euro in a month.	less than 300 Euro monthly
entailment	What can you do in Dresden apart from sightseeing?	You can take a walk by the waterfront.	You can enjoy a relaxing walk by the waterfront.
reverse entailment	Where did the halo originate from?	The halo originated from the light out of the oven.	It originated from the oven.
partial entailment	List two places where one can sit outside!	there are two large terraces and a sunny garden.	In the garden or forest area.
contradiction	Is the apartment located in a new or an old building?	The apartment is in a new building.	The apartment is in an old building.
topical-non-entailment	What was the topic of the survey?	The survey was about things you can't do without.	The topic was usage of the Internet.
off-topic	Who made lawn gnomes famous?	Philipp Griebel made lawn gnomes famous.	It was famous in the Thuringian.

Table 1: Examples for the 7 entailment annotation labels

Due to the difference between classical RTE settings and the task and data we use, our annotation manual contains some guidelines that differ from those for a standard textual entailment task:

Learner Language Issues: One feature of the data that makes the annotation in general difficult is the fact that the LAs in CREG often come in an ungrammatical form or use lexically inappropriate material since they are formulated by language learners. Similarly to teachers in a short answer grading task, our annotators were instructed to ignore such errors. That means they had to implicitly build a so-called *target hypothesis* for each learner answer, i.e. an error-free version of what the learner presumably wanted to express (cf. Ellis (1994)), a task which is known to be problematic even for experienced teachers (Lüdeling, 2008).

Therefore, depending on the interpretation of the annotator, the chosen label can differ, as is illustrated by the following example:

(2) **Q:** Where and when could most garden gnomes be found?
TA: Most garden gnomes could be found in the postwar period in West Germany.
LA: Die Gartenzwerge setzte aus den Wald.
a) The garden gnomes released in the woods.
b) The garden gnomes sets out of the woods.

The LA in this example is ungrammatical and could either be interpreted as "The garden gnomes [were] released into the woods" or "The garden gnomes put [something] out of the woods", leading to *topical non-entailment* as the most plausible label for the first (a) and *off-topic* for the second (b) interpretation.

Note, that the label *contradiction* is not an option for this answer: Although the question presupposes that there is only one correct answer and the topical reading of the learner answer gives a different location than the TA, the two locations "western Germany" and "in the forest" are not mutually exclusive, but the learner answer rather addresses a different type of location than the TA. A clear case of a contradictory answer is instead the following LA: "Most garden gnomes could be found between 1948 and 1952 in the GDR", because GDR refers to a different location than western Germany.

Annotating Answers in Relation to the Question: In contrast to other RTE data sets that compare two texts, our data has the form of answer pairs with both answers referring to the same question. The question is made available to the annotators to resolve anaphoric expressions such as pronouns occurring in the answers and to expand answers in the form of ellipses to *full answers*: Semantic material introduced by the question is explicitly addressed in a *full answer* and omitted in a *term answer* (cf. the example for *paraphrase* in table 1) in the terminology of e.g. Krifka (2001), following von Stechow and Zimmermann (1984). Otherwise, the annotators were instructed to treat short and full answers in the same way. Specifically, only semantic content which has not been introduced by the question should be taken into consideration when deciding between partial entailment and topical-non-entailment. In doing so, we want to avoid that a learner answer is already partially entailed by the TA as soon as it is on-topic and repeats material from the question.

3.3 Annotation Process

All material has been double-annotated by two German native speakers with a background in linguistics using the *WebAnno* annotation tool (Yimam et al., 2013). The annotators were shown the question together with each LA-TA pair, but could not see the corresponding text and did not know whether a LA has been graded as correct or incorrect. We did so to avoid that they would explicitly or implicitly base their labelling decision on the knowledge of whether an answer is correct or supported by the text. Cases of disagreement have been additionally annotated by a third annotator and then be resolved through majority voting. Instances where all three annotators gave a different label have been resolved by one of the authors.

4 Evaluation

This section presents an analysis of our RTE annotations and comparisons to SAS scores.

4.1 Agreement

Our annotators reached a Cohen's Kappa of *0.69* which – according to Landis and Koch (1977) – indicates *substantial agreement*. The confusion matrix is given in table 2. Our results show that the labels *paraphrase*, *entailment* and *reverse entailment* can be reliably identified by the annotators. However, the confusion matrix highlights 2 problems: First, the identification of *partial entailment* is not trivial, as can be seen from a relatively high rate of misclassifications between *partial entailment* and almost any other label. Second, it is challenging to tell apart the three entailment classes *contradiction*, *off-topic* and *topical non-entailment*. As these labels – as we will later see – primarily belong to answers scored as incorrect, we will refer to them as *negative entailment* labels. When collapsing the three labels , our Kappa score improves to *0.78*.

4.2 Comparison of Teacher Scores and Entailment Labels

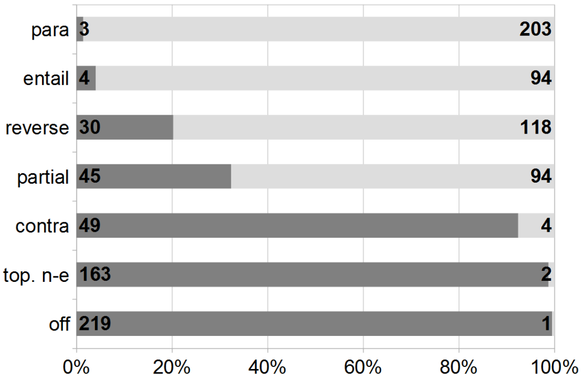

Figure 1: Distribution of entailment labels over binary labels, relative and absolute values (correct: light grey, incorrect: dark grey).

Figure 1 shows the distribution of our entailment labels compared to the binary CREG labels that indicate whether an answer is correct or incorrect. We can see that some of our labels clearly correspond to correct (paraphrase, entailment) or incorrect answers (contradiction, off-topic, topical-non-entailment). From the definition of these labels, this is an expected result: Whenever a LA is a paraphrase of a TA or more specific than a TA it should be correct and whenever a LA contradicts the TA, does not answer the question or answers the question without overlap with the target answer, it is most likely incorrect. However, the labels *partial entailment* and *reverse entailment* cannot be as easily mapped to binary scores, providing evidence for the existence of some substantial differences between the two tasks of RTE and SAS. These two labels have in common, that only some information from the TA is entailed by the LA (while in *partial entailment* the LA additionally entails information not present in the TA). One possible explanation why such answers sometimes are still scored as correct is that often TAs are formulated in an exhaustive way and more elaborate than the teacher would expect the learner to answer. It is not clear however from the TA which facts are necessary to make the LA correct and which facts are not. Example 3 shows one such answer pair, where the binary label is *correct*, although the entailment type clearly is *reverse entailment*. [5]

(3) **Q:** What is needed for paper production?
TA: You need wood, water and energy to produce paper.

[5] An answer just stating *water is needed* does not occur in our corpus, but we would consider it plausible that teachers label such an answer as incorrect, due to the the more prominent role of wood in the paper production process.

	para	entail	reverse	partial	contra	top. n-e	off
para	180	4	12	9	0	2	0
entail	6	78	0	15	0	2	0
reverse	7	5	112	28	2	1	3
partial	5	8	15	75	8	3	10
contra	0	0	0	2	47	1	1
top. n-e	1	0	2	10	35	100	30
off	0	1	3	3	5	31	169

Table 2: Confusion matrix between the two annotators for our labels. Abbreviations: **para**phrase, **entail**ment, **reverse** entailment, **partial** entailment, **contra**diction, **top**ical **n**on-**e**ntailment, **off**-topic.

LA: Wood is needed for paper production.

There are a few curious cases of label score combinations that seem implausible, such as answers with a negative entailment label that are scored as correct answer. The following example (4) illustrates this. While our schema clearly labels the LA as *off-topic*, since question material is paraphrased in a wrong way, the teacher decided to accept the answer by implicitly substituting the location of *Erfurt* with *Frankfurt*.

(4) **Q:** For how long does the company hold a branch at Frankfurt?
TA: The company holds a branch at Frankfurt for 15 years.
LA: It holds a branch at Erfurt for 15 years.

Similarly, there are rare examples of *entailment* or *paraphrase* items that are labeled as *incorrect*. Example 5 shows one such pair, where, both for the entailment label and the correctness score, different options are plausible depending on the interpretation of *warm light* (temperature vs. colour):

(5) **Q:** Why did the man put the wood into the plate oven?
TA: He put the wood into the oven to make the room warmer.
LA: For a warm light through the room.

The findings from this evaluation show that, in our labeling scenario, SAS and RTE are two separate tasks – in contrast to findings by Dzikovska et al. (2013a), who assume that the two tasks do not differ essentially from each other. Thus, their label set contains labels for scoring the LA and exploring its entailment relation simultaneously: They distinguish the label *correct* for complete paraphrases of the TA – which they expect to be the only correct type of answers – and *Partially_correct_incomplete* for LAs that lack information; furthermore *Contradictory* and *Irrelevant* for answers that are on-topic, but either contradictory to the TA or containing the wrong information; and finally *Non_domain* for answers that do not address the question. Our labels are slightly more fine grained: *Partial entailment* has no correspondence in their 5-way label set, but forms for our data the most interesting case for further investigation because of its coverage of both correct and incorrect answers. There is also no correspondence for our *entailment* label. From a SAS perspective, the difference between *paraphrase* and *entailment* seems not to be crucial, as both labels almost exclusively cover answers that are scored binary as correct in our data.

	correct	missing concept	extra concept	blend	non-answer
para	194	7	3	2	0
entail	73	3	16	6	0
reverse	74	53	1	20	0
partial	50	37	8	46	0
contra	1	10	0	42	0
top. n-e	1	15	1	148	0
off	1	40	0	175	4

Table 3: Confusion matrix for teacher assessments and entailment labels.

In addition to binary scores, the CREG corpus also contains a 5-way set of teacher scores (Ott et al. (2012), following Bailey and Meurers (2008)): In these annotations, *missing concept* and *extra concept* were used if the answer missed important information or contained additional, not necessary information, respectively. Therefore we would expect them to match our *reverse entailment* and *entailment* labels, while their *correct* label should correspond to our *paraphrase*. The label *blend* is a combination of *missing* and *extra concept*, seemingly similar to our *partial entailment*. The label

non-answer was used for LAs that did not address the question as with our *off-topic*.
From the label descriptions, we would have expected to see a good fit between the two label sets. Instead, we find that a clear mapping between our labels and the 5-way scores is not possible, as can be seen in the confusion matrix in table 3. Similar to the comparison to the SemEval7 labels, this is mainly the case because the 5-way scores mix aspects of SAS and RTE in an unsuitable way.

5 Machine Learning Experiments

We explore the relation between RTE and SAS through a series of machine learning evaluations. In the first part, we evaluate a SAS classifier asking which LAs in terms of entailment type are most difficult for automatic labeling. We then present a modeling experiment that explores the impact of using our entailment labels as features for a SAS system and finally a series of experiments that aim at testing how well entailment information is modeled by alignment-based machine learning features.

For all experiments, we used the *Logistic* classifier in the *Weka* package, that is based on a logistic regression algorithm (Hall et al., 2009). We use alignment-based features in a re-implementation of Meurers et al. (2011) that reaches an accuracy of 86% on CREG. All experiments were evaluated via leave-one-out cross validation.

Task Setting	**Accuracy**	**Kappa**
teacher-alignment	*0.861*	*0.723*
teacher-entailment	*0.922*	*0.843*
entailment-7	*0.473*	*0.36*
entailment-5	*0.641*	*0.489*
entailment-3	*0.749*	*0.562*
entailment-2	*0.837*	*0.668*

Table 5: Overview of the classifier performances Abbreviations: **teacher** scores as class with **alignment** features and alignment+**entailment** features. **7**-way **entailment** type as class and collapsed entailment class sets by combining entailment types into **5**, **3** or **2** classes, all with alignment features

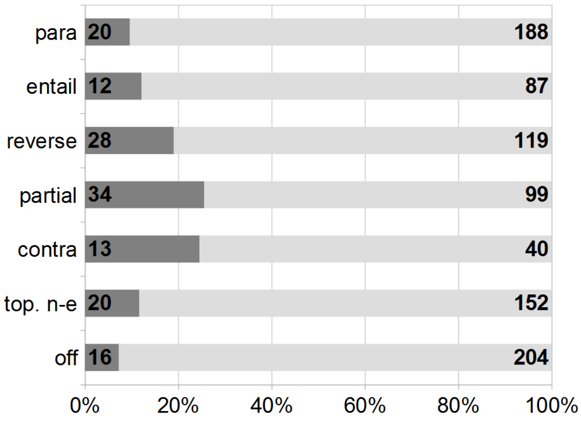

Figure 2: Correctly (light grey) and incorrectly (dark grey) classified instances per entailment class, relative and absolute values.

5.1 Distribution of correctly and incorrectly automatically scored instances over the entailment types

We investigate if some of the entailment types are challenging for a SAS system. Figure 2 shows the distribution of incorrectly classified instances over our entailment types.

In general, LAs that are labeled with *partial entailment* or *reverse entailment* are more problematic for the SAS model than the other labels. This observation reminds of the finding that these labels do not clearly correspond to one correctness score: An alignment based SAS model that covers among its features the percentage of TA tokens and chunks covered in the LA can not differentiate whether a unit not covered was crucial or not. The machine learner also struggles in general with the *contradiction* class. This is because many contradicting answer pairs still provide a high overlap but differ in just a small but critical detail.

Our finding again underlines the difficulty which the evaluation of the semantic overlap between two texts, as can be found in the *partial entailment* group, poses to SAS approaches and reinforces the need for more sophisticated semantic features for modeling these entailment phenomena and consequently for a better shortanswer scoring.

5.2 Can entailment classes improve an SAS feature set?

We enhanced the feature set used by the classifier with our annotated entailment label as an ad-

classified real	para	entail	reverse	partial	contra	top. n-e	off	recall
para	136	11	26	18	1	3	11	*0.66*
entail	20	44	2	24	0	1	7	*0.449*
reverse	24	1	82	17	0	1	23	*0.554*
partial	20	15	20	46	0	9	32	*0.324*
contra	5	2	7	7	1	3	28	*0.019*
top. n-e	2	3	10	13	2	16	119	*0.097*
off	6	3	14	7	1	26	163	*0.741*
precision	*0.638*	*0.557*	*0.51*	*0.348*	*0.2*	*0.271*	*0.426*	

Table 4: Confusion matrix for the machine learner on our labels with precision and recall for all classes.

ditional feature in order to explore whether our annotations, could they be determined automatically, would be helpful in a SAS task. This raises the classifier's performance from *86.1% (κ=0.723)* to *92.2% (κ=0.843)*, as can be seen in table 5. Although we showed that the RTE and SAS scenario differ substantially, this outcome emphasizes that they also have a lot in common.

However, the obvious problem here is that the usage of a manually annotated feature is comparable to the use of a human oracle and is therefore not feasible for a fully automatic approach. Thus, further research has to concentrate on how we can automatically model entailment types computationally. To do so, we will for example try to enhance the current TA-LA alignment based SAS approach. This leads to the question in how far the model is already able to predict our entailment types. One first evaluation trial of this question is presented in the next section.

5.3 Are entailment relations learnable with an SAS system?

In this last set of experiments we address the question in how far the automatic prediction of entailment labels is possible with the feature sets of an alignment based SAS approach. Although the focus in an educational application would be the automatic scoring of the correctness of a LA rather than its entailment relation to its TA, this experiment might shed additional light on the relatedness of the two tasks.

We therefore train our classifier on the LA data and use the entailment labels as class, which leads to an accuracy of *47%* (table 5) and a kappa indicating *poor agreement*. The confusion matrix for this classification (table 4) shows that the machine learner especially struggles with labeling the negative classes, because the features it uses are computed based on the alignment between TA and LA, while the question is not taken into account. Therefore the machine learner is unable to decide if an answer addresses the question or not. *Partial entailment* poses a large difficulty again as well, resulting in an F1-Score of *0.336 (P=0.348/R=0.324)* for that class. In contrast, the F1-Score for paraphrase reaches a modest level of *0.649 (P=0.638/R=0.66)*.

To narrow down the difficulties for our machine learner, we stepwise collapsed our entailment labels, by first subsuming the negative entailment classes *topical non-entailment*, *off-topic* and *contradiction* as one class, which leads to only 5 entailment classes and an accuracy of *64.1%*. In the next step, we subsumed *entailment*, *reverse entailment* and *paraphrase* under one "positive" label, but left partial entailment out, which lead to 3 classes (positive, negative, partial) and an accuracy of *74.9%*. Finally, we added *partial entailment* to the positive class and achieved a performance of *83.7%*. Although it is in general not surprising that the performance increases as the number of labels decreases, it is interesting that the inclusion or exclusion of *partial entailment* has a rather high impact on the performance.

6 Conclusions and Future Work

This paper presented a study that labels LA TA pairs from the CREG corpus with a set of fine-grained textual entailment annotations. Our main finding is that there is a clear correspondence between some textual entailment classes and a binary correctness score. But there is also an area that needs further investigation. This concerns the *partial* and *reverse entailment* cases and illustrates that the tasks of RTE and SAS are related, but not equivalent for our scenario.

One next step will be to investigate the structure of answers that are labeled as *partial* or *reverse entailment* as those instances seem to be particularly problematic for automatic SAS. For advances in automatic scoring it is important to determine which parts of a target answer are crucial for a correct LA and which are not.

In the next step of this annotation project, we will focus on the relation between reading texts and answers. We expect that the combination of this variant of the RTE setting with our current annotations helps us to gather further insights into the nature of shortanswer questions.

Acknowledgements

We thank the three anonymous reviewers for their helpful comments. We also thank our annotators Fernando Ardente, Sophie Henning and Maximilian Wolf for their help with this study and alexis Palmer for valuable feedback for our annotation guidelines.

This work was funded by the Cluster of Excellence "Multimodal Computing and Interaction" of the German Excellence Initiative.

References

Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, page 107115, Columbus, Ohio, USA, June.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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 *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 13–18, 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*, pages 200–210. Association for Computational Linguistics.

Myroslava O. Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013a. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. **SEM 2013: The First Joint Conference on Lexical and Computational Semantics*.

Myroslava O Dzikovska, Rodney D Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa T Dang. 2013b. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Technical report, DTIC Document.

Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment* *and Paraphrasing*, RTE '07, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aria Haghighi, Andrew Y. Ng, and Christopher D. Manning. 2005. Robust textual inference via graph matching. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 387–394, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Michael Hahn and Detmar Meurers. 2012. Evaluating the meaning of answers to reading comprehension questions - a semantics-based approach. *The 7th Workshop on the Innovative Use of NLP for Building Educational Applications.*

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Manfred Krifka. 2001. For a structured meaning account of questions and answers. *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, pages 287–319.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):pp. 159–174.

Omer Levy, Torsten Zesch, Ido Dagan, and Iryna Gurevych. 2013. Recognizing partial textual entailment. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, page 451455, Sofia, Bulgaria, August 4-9.

Anke Lüdeling. 2008. Mehrdeutigkeiten und kategorisierung: Probleme bei der annotation von lernerkorpora. *Fortgeschrittene Lernervarietäten*, pages 119–140.

Bill MacCartney and Christopher D Manning. 2009a. An extended model of natural logic. In *Proceedings of the eighth international conference on computational semantics*, pages 140–156. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2009b. An extended model of natural logic. In *Proceedings of the Eighth International Conference on Computational Semantics*, IWCS-8 '09, pages 140–156, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 567–575, Stroudsburg, PA, USA. Association for Computational Linguistics.

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-Volume 1*, pages 752–762. Association for Computational Linguistics.

Rodney D Nielsen, Wayne Ward, James H Martin, and Martha Palmer. 2008. Annotating students' understanding of science concepts. In *LREC*.

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

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.

Jana Zuheir Sukkarieh and John Blackmore. 2009. c-rater: Automatic content scoring for short constructed responses. In *FLAIRS Conference*.

Arnim von Stechow and Thomas Ede Zimmermann. 1984. Term answers and contextual change. *Linguistics*, 22:3–40.

Arnim von Stechow, 1990. *Discourse Particles*, chapter Focusing and Background Operators.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible,web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 1–6, Stroudsburg, PA, USA, August. Association for Computational Linguistics.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, pages 190–200, Montreal, Canada. Association for Computational Linguistics.