Using the text to evaluate short answers for reading comprehension exercises

Andrea Horbach, Alexis Palmer and Manfred Pinkal
Department of Computational Linguistics, Saarland University, Saarbrücken, Germany
(andrea|apalmer|pinkal)@coli.uni-saarland.de

Abstract

Short answer questions for reading comprehension are a common task in foreign language learning. Automatic short answer scoring is the task of automatically assessing the semantic content of a student's answer, marking it e.g. as correct or incorrect. While previous approaches mainly focused on comparing a learner answer to some reference answer provided by the teacher, we explore the use of the underlying reading texts as additional evidence for the classification. First, we conduct a corpus study targeting the links between sentences in reading texts for learners of German and answers to reading comprehension questions based on those texts. Second, we use the reading text directly for classification, considering three different models: an answer-based classifier extended with textual features, a simple text-based classifier, and a model that combines the two according to confidence of the text-based classification. The most promising approach is the first one, results for which show that textual features improve classification accuracy. While the other two models do not improve classification accuracy, they do investigate the role of the text and suggest possibilities for developing automatic answer scoring systems with less supervision needed from instructors.

1 Introduction

Reading comprehension exercises are a common means of assessment for language teaching: students read a text in the language they are learning and are then asked to answer questions about the text. The types of questions asked of the learner may vary in their scope and in the type of answers they are designed to elicit; in this work we focus on "short answer" responses, which are generally in the range of 1–3 sentences.

The nature of the reading comprehension task is that the student is asked to show that he or she has *understood* the text at hand. Questions focus on one or more pieces of information from the text, and correct responses should contain the relevant semantic content. In the language learning context, responses classified as correct might still contain grammatical or spelling errors; the focus lies on the content rather than the form of the learner answer.

Automatic scoring of short answer responses to reading comprehension questions is in essence a textual entailment task, with the additional complication that, in order to answer a question correctly, the learner must have identified the right portion of the text. It isn't enough that a student answer is entailed by *some* part of the reading text; it must be entailed by the part of the text which is responsive to the question under discussion.

Previous approaches to automatic short answer scoring have seldom considered the reading text itself, instead comparing student answers to target answers supplied by instructors; we will refer to these as *answer-based models*. In this paper we explore the role of the text for short answer scoring, evaluating several models for considering the text in automatic scoring, and presenting results of an annotation study regarding the semantic links between reading texts and answers to reading comprehension questions.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 286–295, Atlanta, Georgia, June 13-14, 2013.

TEXT: SCHLOSS PILLNITZ

This palace, which lies in the east of Dresden, is to me the most beautiful palace in the Dresden area. (...) One special attraction in the park is the camellia tree. In 1992, the camellia, which is more than 230 years old and 8.90 meters tall, got a new, moveable home, in which temperature, ventilation, humidity, and shade are controlled by a climate regulation computer. In the warm seasons, the house is rolled away from the tree. During the Blossom Time, from the middle of February until April, the camellia has tens of thousands of crimson red blossoms. Every year, a limited number of shoots from the Pillnitz camellia are sold during the Blossom Time, making it an especially worthwhile time to visit.

QUESTION:
A friend of yours would like to see the historic camellia tree. When should he go to Pillnitz, and why exactly at this time?

TARGET ANSWERS:

- From the middle of February until April is the Blossom Time.
- In spring the camellia has tens of thousands of crimson red blossoms.

LEARNER ANSWERS:

- [correct] He should go from the middle of February until April, because then the historic camellia has tens of thousands of crimson red blossoms.
- [incorrect] Every year, a limited number of Pillnitz camellia are sold during the Blossom Time.
- [incorrect] All year round against temperature and humidity are controlled by a climate regulation computer.

Figure 1: Example of reading text with question and answers (translation by authors)

These investigations are done for German language texts, questions, and answers. Figure 1 shows a (translated) sample reading text, question, set of target answers, and set of learner answers.

We show that the use of text-based features improves classification performance over purely answer-based models. We also show that a very simple text-based classifier, while it does not achieve the same performance as the answer-based classifier, does reach an accuracy of 76% for binary classification (correct/incorrect) of student answers. The implication of this for automatic scoring is that reasonable results may be achievable with much less effort on the part of instructors; namely, a classifier trained on the supervision provided by marking the region of a text relevant to a given question performs reasonably well, though not as well as one trained on full target answers.

The paper proceeds as follows: in Section 2 we discuss the task and related approaches. In Section 3, we describe our baseline model and the data set we use. In Section 4 and Section 5 we discuss our text-based models and present experiments and results.

2 Approaches to short answer scoring

In short answer scoring (SAS) the task is to automatically assign labels to individual learner answers. Those labels can either be binary, a value on some scale of points or grades, or a more fine-grained diagnosis. For example, one fine-grained set of labels (Bailey, 2008) classifies answers as (among others) correct, as missing a necessary concept or concepts, containing extra content, or as failing to answer the question. Our present study is restricted to binary classification.

Previous work on SAS, including early systems like (Leacock and Chodorow, 2003; Pulman and Sukkarieh, 2005; Sukkarieh and Pulman, 2005) is of course not only in the domain of foreign language learning. For example, Mohler et al. (2011) and Mohler and Mihalcea (2009) use semantic graph alignments and semantic similarity measures to assess student answers to computer science questions, comparing them to sample solutions provided by a teacher. Accordingly, not all SAS settings include reading or other reference texts; many involve only questions, target answers, and learner answers. Our approach is relevant for scenarios in which some sort

of reference text is available.

The work we present here is strongly based on approaches towards SAS by Meurers and colleagues (Bailey and Meurers, 2008; Meurers et al., 2011a; Meurers et al., 2011b; Ziai et al., 2012). Specifically, the sentence alignment model described in Section 3 (and again discussed in Section 4) is modeled after the one used by Meurers et al. to align target answers and student answers.

Rather than using answers provided by instructors, Nielsen et al. (2008) represent target answers to science questions as a set of hand-annotated *facets*, i.e. important aspects of the answer, typically represented by a pair of words and the relation that connects them. Student answers, and consequently students' understanding of target science concepts, are then assessed by determining whether the relevant facets are addressed by the learner answers.

Evaluating short answers on the basis of associated reading texts, as we do here, is a task related to textual entailment. In the context of tutoring systems, Bethard et al. (2012) identify students' misconceptions of science concepts in essay writing using textual entailment techniques. They align students' writings to extracted science concepts in order to identify misconceptions, using a similar approach to identify the correct underlying concept.

An excellent and more detailed overview of related work can be found in Ziai et al. (2012).

To our knowledge, there is no previous work that uses reading texts as evidence for short answer scoring in the context of foreign language learning.

3 Answer-based models

In order to compare to previous work, we first implement an alignment-based model following that proposed in (Meurers et al., 2011b). We refer to this class of models as *answer-based* because they function by aligning learner answers to instructor-supplied target answers along several different dimensions, discussed below. Answers are then classified as correct or incorrect on the basis of features derived from these alignments.

Wherever possible/practical, we directly re-implement the Meurers model for German data. In this section we describe relevant aspects of the Meurers model, along with modifications and extensions in our implementation of that model.[1]

Preprocessing

We preprocess all material (learner answers, target answers, questions and reading texts) using standard NLP tools for sentence splitting and tokenization (both OpenNLP[2]), POS tagging and stemming (both Treetagger (Schmid, 1994)), NP chunking (OpenNLP), and dependency parsing (Zurich Parser (Sennrich et al., 2009)). We use an NE Tagger (Faruqui and Padó, 2010) to annotate named entities. Synonyms and semantic types are extracted from GermaNet (Hamp and Feldweg, 1997).

For keywords, which serve to give more emphasis to content words in the target answer, we extract all nouns from the target answer.

Given that we are dealing with learner language, but do not want to penalize answers for typical learner errors, spellchecking (and subsequent correction of spelling errors) is especially important for this task. Our approach is as follows: we first identify all words from the learner answers that are not accepted by a German spellchecker (aspell[3]). We then check for each word whether the word nevertheless occurs in the target answer, question or reading text. If so, we accept it as correct. Otherwise, we try to identify (using Levenshtein distance) which word from the target answer, question, or reading text is most likely to be the form intended by the student.

Prior to alignment, we remove from the answer all punctuation, stopwords (restricted to determiners and auxiliaries), and material present in the question.

Alignment

The alignment process in short answer scoring approximates determination of semantic equivalence between target answer and learner answer. During alignment, we identify matches between answer pairs on a number of linguistic levels: tokens, chunks, and dependency triples.

On the token level, we consider a number of different metrics for identity between tokens, with each

[1] Some extensions were made in order to bring performance of our re-implementation closer to the figures reported in previous work.

[2] http://opennlp.apache.org/index.html

[3] http://aspell.net/

metric associated with a certain alignment weight. After weights have been determined for all possible token pairs, the best applicable weight is used as input for a traditional marriage alignment algorithm (Gale and Shapley, 1962).

We use the following types of identity (id), weighted in the following order:

```
token id > lemma id >
spelling id > synonym & NE id >
similarity id>
NE type, semantic type & POS id
```

For synonym identity, we take a broad notion of synonymy, extracting (from GermaNet) as potential synonyms all words which are at most two levels (in either direction) away from the target word. Similarity identity is defined as two words having a GermaNet path relatedness above some threshhold. In order to have semantic type identity, two words must have a common GermaNet hypernym (from a predetermined set of relevant hypernyms). Only some closed-class words are eligible for POS identity. We treat e.g. all types of determiners as POS identical.

Unlike, for example, alignment in machine translation, in which every token pair is considered a candidate for alignment, under the Meurers model only candidates with at least one type of token identity are available for alignment. This aims to prevent completely unrelated word pairs from being considered for alignment.

In order to favor alignment of content words over alignment of function words, and in departure from the Meurers model, we use a content word multiplier for alignment weights.

Chunks can only be aligned if at least one pair of tokens within the respective chunks has been aligned, and the percentage of aligned tokens between learner and target answer chunks is used as input for the alignment process. Dependency triple pairs are aligned when they share dependency relation, head lemma, and dependent lemma.

Features and classifier

After answers have been aligned, the following features are extracted as input for the classifier: keyword overlap (percentage of aligned keywords), target token overlap (percentage of aligned target tokens), learner token overlap (percentage of aligned learner tokens), token match (percentage of token alignments that are token identical), lemma match, synonym match, type match, target triple overlap, learner triple overlap, target chunk overlap, learner chunk overlap, target bigram overlap, learner bigram overlap, target trigram overlap, learner trigram overlap, and variety of alignment (number of different token alignment types).

The n-gram features are the only new features in our re-implementation of the Meurers model, hoping to capture the influence of linear ordering of aligned tokens. These features did not in the end improve the model's performance.

For classification, we use the timbl toolkit (Daelemans et al., 2009) for k-nearest neighbors classification. We treat all features as numeric values and evaluate performance via leave-one-out cross-validation. Further details appear in Section 5.

Data

For all work reported in this paper, we use the German CREG corpus (Ott et al., 2012) of short answers to questions for reading comprehension tasks. More specifically, we use a balanced subset of the CREG corpus containing a total of 1032 learner answers. This corpus consists of 30 reading texts with an average of 5.9 questions per text. Each question is associated with one or more target answers, specified by a teacher. For each question in turn there are an average of 5.8 learner answers, each manually annotated according to both binary and fine-grained labeling schemes. When there are several target answers for a question, the best target answer for each learner answer is indicated.

4 Text-based approach

Previous approaches to this task take the instructor-supplied target answer(s) as a sort of supervision; the target answer is meant to indicate the semantic content necessary for a correct student answer. Alignment between student answer and target answer is then taken as a way of approximating semantic equivalence. The key innovation of the current study is to incorporate the reading text into the evaluation of student answers. In this section we describe and evaluate three approaches to incorporating the text. The aim is to consider the semantic

relationships between target answer, learner answer, and the text itself.[4]

A target answer is in fact just one way of expressing the requisite semantic content. Teachers who create such exercises are obviously looking at the text while creating target answers, and target answers are often paraphrases of one or more sentences of the reading text. Some learner answers which are scored as incorrect by the answer-based system may in fact be variant expressions of the same semantic content as the target answer. Due to the nature of the reading comprehension task, in which students are able to view the text while answering questions, we might expect students to express things in a manner similar to the text. This is especially true for language learners, as they are likely to have a limited range of options both for lexical expression and grammatical constructions.

Along similar lines, one potential source of incorrect answers is an inability on the part of the student to correctly identify the portion of the text that is relevant to the question at hand. Our hypothesis therefore is that a learner answer which links to the same portion of the reading text as the target answer is likely to be a **correct** answer. Similarly, a learner answer which closely matches some part of the text that is *not* related to the target answer is likely to be **incorrect**.

Our text-based models investigate this hypothesis in several different ways, described in Section 4.2.

4.1 Annotation study

The CREG data includes questions, learner answers, target answers, and reading texts; associations between text and answers are not part of the annotations. We undertook an annotation project in order to have gold-standard **source sentences** for both learner and target answers. This gold-standard is then used to inform the text-based models described below.

After removing a handful of problematic questions and their associated answers, we acquired human annotations for 889 of the 1032 learner answers from the balanced subset of the CREG corpus, in addition to 294 target answers. Each answer was labeled separately by two (of three) annotators, who were given the reading text and the question and asked to identify the single best source sentence from the text. Annotators were not told whether any given instance was a target or learner answer, nor whether learner answers were correct or incorrect.

Although we expected most answers to correspond directly to a single text passage (Meurers et al., 2011b), annotators were asked to look for (and annotate appropriately) two different conditions in which more than one source sentence may be relevant. We refer to these as the repeated content condition and the distributed content condition.

In the *repeated content condition*, the same semantic content may be fully represented in more than one sentence from the original text. In such cases, we would expect the text to contain sentences that are paraphrases or near-paraphrases of one another. The *distributed content condition* occurs when the relevant semantic content spans multiple sentences, and some degree of synthesis or even inference may be required to arrive at the answer. Annotators were instructed to assume that pronouns had been resolved; in other words, a sentence should not be considered necessary semantic content simply because it contains the NP to which a pronoun in another sentence refers. For both of these multiple-sentence conditions, annotators were asked to select one single-best source sentence from among the set and also to mark the alternative source sentences.

For 31.2% of the answers annotated, one or more annotator provided more than one possible source sentence. Upon closer inspection, though, the annotations for these conditions are not very consistent. In the repeated content condition, there is very little agreement between annotators regarding when the text contains more than one full-sentence source for the answer. In the distributed content condition, sometimes annotators disagree on the primary sentence, and in many instances, one annotator identified multiple sentences and the other only one. Due to these inconsistencies, for the purpose of this study we decided to treat the multiple-sentence conditions in an underspecified fashion. When an annotator has identified either of these conditions, we convert the annotations to a single-best sentence and a set of alternatives.

The annotations were processed to automatically

[4] In future work we will also consider semantic relationships between the question and the text.

Answer type	agree	altagree	disagree	nolink
Learner answers (all)	70.3%	9.4%	16.9%	3.4%
Learner answers (correct)	75.1%	11.7%	12.7%	0.5%
Learner answers (incorrect)	65.9%	7.3%	20.7%	6.4%
Target	73.1%	8.1%	17.3%	1.4%

Table 1: Inter-annotator agreement for linking answers to source sentences in text

produce a gold-standard set of source sentence IDs, indicating the single sentence in the reading text to which each answer is most closely linked. We identify four distinct categories with respect to agreement between annotators. Agreement figures appear in Table 1.

** **agree:** In this case, both annotators linked the answer to the same source sentence, and that sentence is identified as the gold-standard link to the answer.

** **altagree:** This category covers two different situations in which the two annotators fail to agree on the single-best sentence. First, there are cases in which the best sentence selected by one annotator is a member of the set of alternatives indicated by the other. Second, in a small number of cases, both annotators agree on one member of the set of alternatives. In other words, the single sentence in the intersection of the sets of sentences identified by the two annotators is taken as the gold-standard annotation. There was no (non-**agree**) case in which that intersection contained more than one sentence.

** **disagree:** This category also includes two different types of cases. In the first, one of the two annotators failed to identify a source sentence to link with the answer. In that case, we consider the annotators to be in disagreement, and for the gold-standard we use the sentence ID provided by the one responding annotator. In the second case, the annotators disagree on the single-best sentence and there is no overlap between indicated alternative sentences. In those cases, for the gold standard we choose from the two source sentences that which appears first in the reading text.[5]

** **nolink:** For a small number of answers (n=34), both annotators found no link to the text. One example of such a case is an answer given entirely in English. For these cases, the gold standard provides no best source sentence.

If we consider both **altagree** and **nolink** to be forms of agreement, interannotator agreement is about 74% for both learner and target answers.

4.2 Text-based models

In this paper we consider two different models for incorporating the reading text into automatic short answer scoring. In the first approach, we employ a purely text-based model. The second combines either text-based features or the text-based model with the answer-based model described in Section 3. Evaluation of all three approaches appears in Section 5.

4.2.1 Simple text-based model

This model classifies student answers by comparing the source sentence most closely associated with the student answer to that associated with the target answer. If the two sentences are identical, the answer is classified as **correct**, and otherwise as **incorrect**.

We consider both the annotated best sentences (**goldlink**) and automatically-identified answer-sentence pairs (**autolink**). For automatic identification, we use the alignment model described in Section 3 to identify the best matching source sentence in the text for both learner and target answers. We use the token alignment process to align a given answer with each sentence from its respective reading text; the best-matching source sentence is that with the highest alignment weight. Chunk alignments are used only for correction of token alignments, and dependency alignments are not considered.

This model takes an extremely simple approach to answer classification, and could certainly be refined and improved. At the same time, its relatively strong

[5]This is a relatively arbitrary decision motivated by the desire to have a source sentence associated with as many answers as possible. Future work may include adjudication of annotations to reduce the noise introduced to the gold standard by this category of responses.

performance (see Table 3) suggests that the minimal level of supervision offered by teachers simply marking the sentence of a text most relevant to a given reading comprehension question may be beneficial for automatic answer scoring.

4.2.2 Combining text-based and answer-based models

In addition to the purely text-based model, we explore two ways of combining text- and answer-based models.

Textual features in the answer-based model In the first, we extract four features from the alignments between answers and source sentences and incorporate these as additional features in the answer-based model.

Features 1, 3, and 4 are each computed in two versions, using source sentences from either the annotated gold standard (**goldlink**), or the alignment model (**autolink**).

1. **SourceAgree** This boolean feature is true if both learner and target answer link to the same source sentence, and false otherwise (also if no source sentence was annotated or automatically found).

2. **SourceEntropy** For this feature we look at the two most-likely source sentences for the learner answer, as determined by automatic alignment scores. We treat the alignment weights as probabilities, normalizing so that they sum up to one. We then take the entropy between these two alignment weights as indicative of the confidence of the automatic alignment for the learner answer.

3. **AgreeEntropy** Here we weight the first feature according to the second, taking the entropy as a confidence score for the binary feature. Specifically, we value **SourceAgree** at 0.5 when the feature is true, -0.5 when false, and multiply this with $(1 - \text{entropy})$.

4. **TextAdjacency** This feature captures the distance (in number of sentences) between the source sentence linked to the learner answer and that linked to the target answer. With this feature we aim to capture the tendency of adjacent passages in a text to exhibit topical coherence (Mirkin et al., 2010).

Classifier combination In the second approach, we combine the output of the answer-based and text-based classifiers to arrive at a final classification system, allowing the text-based classifier to predominate in those cases for which it is most confident and falling back to the answer-based classifier for other cases. Confidence of the text-based classifier is determined based on entropy of the two highest-scoring alignments between learner answer and source sentence. The entropy threshhold was determined empirically to 0.5.

5 Experiments and results

This section discusses experiments on short answer scoring (binary classification) for German, in the context of reading comprehension for language learning. Specifically, we investigate the text-based models described in Section 4.2. In all cases, features and parameter settings were tuned on a development set which was extracted from the larger CREG corpus. In other words, there is no overlap between test and development data. For testing, we perform leave-one-out cross-validation on the slightly-smaller subset of the corpus which was used for annotation.

5.1 Answer-based baseline

As a baseline for our text-based models we take our implementation of the answer-based model from (Meurers et al., 2011b). As previously mentioned, our implementation diverges from theirs at some points, and we do not quite reach the performance reported for their model (accuracy of 84.6% on the balanced CREG corpus) and are far from reaching the current state of the art accuracy of 86.3%, as reported in Hahn and Meurers (2012).

Our answer-based model appears as **baseline** in Table 2. During development, the one extension to the baseline which helped most was the use of extended synonyms. This variant of the model appears in the results table with the annotation **+syn**.

model	k=5	k=15	k=30
baseline	0.817	0.820	0.822
baseline+syn	0.822	0.826	0.825
text: goldlink	0.827	0.827	0.829
text+syn:goldlink	0.830	0.835*	0.837*
text:autolink	0.837*	0.836*	0.825
text+syn:autolink	0.844*	0.836*	0.832
combined	0.810	0.819	0.816
combined+syn	0.817	0.822	0.818

Table 2: Classification accuracy for answer-based baseline (**baseline**), answer-based plus textual features (**text**), and classifier combination (**combined**). **+syn** indicates expanded synonymy features, **goldlink** indicates identifying the source sentences via annotated links, **autolink** indicates determining source sentences using the alignment model, k=number of neighbors. Results marked with * are significant compared to the best baseline model. See Section 5.2.1 for details.

	autolink	**goldlink**	**alt-set**
Accuracy	0.762	0.722	0.747
P correct	0.805	0.781	0.753
R correct	0.667	0.585	0.702
F correct	0.729	0.668	0.727
P incorrect	0.735	0.689	0.742
R incorrect	0.851	0.849	0.788
F incorrect	0.789	0.761	0.764

Table 3: Classification accuracy, precision, recall, and f-score for simple text-based classifier, under three different conditions. See Section 5.2.2 for details.

5.2 Text-based models

As described in Section 4.2, we consider three different approaches for incorporating the reading text into answer classification: use of textual features in the answer-based model, combination of separate answer-based and text-based models, and a simple text-based classifier.

5.2.1 Combining text-based and answer-based models

We explore two ways of combining text- and answer-based models.

Adding textual features to the answer-based model

We evaluate the contribution of the four new text-based features, computed in two variations: with source sentences as they are identified in the gold standard (**goldlink**) and as they are computed using the alignment model (**autolink**). We add those additional features to the two answer-based systems: the baseline (**text**) and the baseline with extended synonym set (**text+syn**). Results are presented in Table 2.

We present results for using the 5, 15, and 30 nearest neighbors for classification, as the influence of various features changes with the number of neighbors. We calculate the significance for the difference between the best baseline model (0.826) and each model which uses textual features, using a resampling test (Edgington, 1986). The results marked with a * in the Table 2 are significant at $p \leq 0.01$.

Although the impact of the textual features is clearly not as big with a stronger baseline model, we still see a clear pattern of improved accuracy. We may expect this difference to increase with more data and with additional and/or improved text-based features.

Classifier combination

Combining the two classifiers (answer-based and text-based) according to confidence levels results in decreased performance compared to the baseline. These results appear in Table 2 as **combined**.

5.2.2 Simple text-based classification

We have seen that textual features improve classification accuracy over the answer-driven model, yet this approach still requires the supervision provided by teacher-supplied target answers. In our third model, we investigate how the system performs without this degree of supervision, considering how far we can get by using *only* the text.

The simple text-based classifier, rather than taking a feature-based approach to classification, bases its decision solely on whether or not the learner and target answers link to the same source sentence. We compare three different methods for obtaining these links. The first approach (**autolink**) automatically links each answer to a source sentence from the text, based on alignments as described in Section 3. The second (**goldlink**) uses links as provided by the gold standard; in this case, learner answers without

a linked sentence (e.g. **nolink** cases) are immediately classified as incorrect. The third approach (**alt-set**) exploits that fact that in many cases annotators provided alternate source sentences. Under this approach, an answer is classified as correct provided that there is a non-empty intersection between the set of possible source sentences for the learner answer and that for the target answer. For the second and third approaches, we classify as incorrect those learner answers lacking a gold-standard annotation for the corresponding target answer.

In Table 3 we present classification accuracy, precision, recall, and f-score for the three different conditions. Precision, recall, and f-score are reported separately for correct and incorrect learner answers. The 76% accuracy reached using the simple text-based classifier suggests that a system which has teachers supply source sentences instead of target answers and then automatically aligns learner answers to the text, while nowhere near comparable to the state-of-the-art supervised system, still achieves a reasonably accurate classification.

6 Conclusion

In this paper we have presented the first use of reading texts for automatic short answer scoring in the context of foreign language learning. We show that, for German, the use of simple text-based features improves classification accuracy over purely answer-based models. We plan in the future to investigate a wider range of text-based features. We have also shown that a simple classification model based only on linking answers to source sentences in the text achieves a reasonable classification accuracy. This finding has the potential to reduce the amount of teacher supervision necessary for authoring short answer exercises within automatic answer scoring systems. In addition to these findings, we have presented the results of an annotation study linking both target and learner answers to source sentences.

In the near-term future we plan to further investigate the role of the reading text for short answer scoring along three lines. First, we will address the question of the best size of text unit for alignment. In many cases, the best answers are linked not to entire sentences but to regions of sentences; in others, answers may correspond to more than one sentence. Our current approach ignores this issue. Second, we are interested in the variety of semantic relationships holding between questions, answers and texts. Along these lines, we will further investigate the sets of alternatives provided by annotators, as well as bringing in notions from work on paraphrasing and recognizing textual entailment. Finally, we are interested in moving from simple binary classification to the fine-grained level of diagnosis.

Acknowledgments

We would like to thank Erik Hahn, David Alejandro Przybilla and Jonas Sunde for carrying out the annotations. We thank the three anonymous reviewers for their helpful comments. This work was funded by the Cluster of Excellence "Multimodal Computing and Interaction" of the German Excellence Initiative and partially funded through the INTERREG IV A programme project ALLEGRO (Project No.: 67 SMLW 11137).

References

Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 107–115, Columbus, Ohio, June.

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.

Steven Bethard, Haojie Hang, Ifeyinwa Okoye, James H. Martin, Md. Arafat Sultan, and Tamara Sumner. 2012. Identifying science concepts and student misconceptions in an interactive essay writing tutor. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 12–21.

Walter Daelemans, Jakub Zavrel, Ko Sloot, and Antal Van Den Bosch. 2009. TiMBL: Tilburg Memory-Based Learner, version 6.2, Reference Guide. ILK Technical Report 09-01.

Eugene S Edgington. 1986. *Randomization tests*. Marcel Dekker, Inc., New York, NY, USA.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

David Gale and Lloyd S. Shapley. 1962. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):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 (BEA7)*, pages 326–336, Montreal, Canada. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. Germanet - a lexical-semantic net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

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

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. 2011a. 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.

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, Scottland, UK.

Shachar Mirkin, Ido Dagan, and Sebastian Padó. 2010. Assessing the role of discourse references in entailment inference. In *ACL*.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In Alex Lascarides, Claire Gardent, and Joakim Nivre, editors, *EACL*, pages 567–575.

Michael Mohler, Razvan C. Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 752–762.

Rodney D. Nielsen, Wayne Ward, and James H. Martin. 2008. Learning to assess low-level conceptual understanding. In David Wilson and H. Chad Lane, editors, *FLAIRS Conference*, pages 427–432.

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), pages 47–69. Benjamins, Amsterdam.

Stephen G. Pulman and Jana Z. Sukkarieh. 2005. Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, EdAppsNLP 05, pages 9–16.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, United Kingdom.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A new hybrid dependency parser for German. In Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten ? From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference 2009*, pages 115–124. Narr, Tübingen.

Jana Z. Sukkarieh and Stephen G. Pulman. 2005. Information extraction and machine learning: Automarking short free text responses to science questions. In Chee-Kit Looi, Gordon I. McCalla, Bert Bredeweg, and Joost Breuker, editors, *AIED*, volume 125 of *Frontiers in Artificial Intelligence and Applications*, pages 629–637.

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)*, Montreal, Canada.