The Impact of Spelling Correction and Task Context on Short Answer Assessment for Intelligent Tutoring Systems

Ramon Ziai Florian Nuxoll
Kordula De Kuthy Björn Rudzewitz Detmar Meurers

Collaborative Research Center 833
Department of Linguistics, ICALL Research Group*
LEAD Graduate School & Research Network
University of Tübingen

Abstract

This paper explores Short Answer Assessment (SAA) for the purpose of giving automatic meaning-oriented feedback in the context of a language tutoring system. In order to investigate the performance of standard SAA approaches on student responses arising in real-life foreign language teaching, we experimented with two different factors: 1) the incorporation of spelling normalization in the form of a task-dependent noisy channel model spell checker (Brill and Moore, 2000) and 2) training schemes, where we explored task- and item-based splits in addition to standard tenfold cross-validation.

For evaluation purposes, we compiled a data set of 3,829 student answers across different comprehension task types collected in a German school setting with the English tutoring system FeedBook (Rudzewitz et al., 2017; Ziai et al., 2018) and had an expert score the answers with respect to appropriateness (correct vs. incorrect). Overall, results place the normalization-enhanced SAA system ahead of the standard version and a strong baseline derived from standard text similarity measures. Additionally, we analyze task-specific SAA performance and outline where further research could make progress.

1 Introduction

Short Answer Assessment (SAA) is the task of determining whether an answer to a question is correct or not with respect to meaning. The task is also often called Automatic Short Answer Grading (ASAG) in cases where the outcome to determine is on an ordered scale (e.g., a numeric score). After a surge of attention (cf., e.g., Burrows et al., 2015) including shared tasks at SemEval (Dzikovska et al., 2013) and Kaggle[1], the field has quietened down somewhat, with a couple of recent exceptions (Riordan et al., 2017; Gomaa and Fahmy, 2019).

However, SAA cannot be considered a solved problem. In particular, it is still unclear how well standard SAA approaches work in real-life educational contexts, for example when integrating language tutoring systems into a regular school setting. In such systems, the goal is to give immediate feedback on the language produced by the learner, e.g., to help students complete homework exercises in the system step by step. For meaning-oriented exercises, such as reading and listening comprehension, this is especially challenging, since the system needs to evaluate the meaning provided by the student response and possibly give helpful feedback on how to improve it in the direction of an acceptable answer. SAA can help with the evaluation part: if an answer is deemed correct, the feedback is positive, if not, further diagnosis can be carried out. The purpose of SAA in this context is thus to help the tutoring system decide whether the feedback to be given needs to be positive or negative.

In this paper, we therefore report on SAA work in progress on authentic data from a language tutoring system for 7th grade English currently in use in German schools. We employ an alignment-based SAA system (CoMiC, Meurers et al., 2011a) shown to work well for several data sets where target answers are available (Meurers et al., 2011b; Ott et al., 2013), and use it to train a classifier mimicking a trained language teacher's

* http://icall-research.de

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

Ramon Ziai, Florian Nuxoll, Kordula De Kuthy, Björn Rudzewitz and Detmar Meurers 2019. The impact of spelling correction and task context on short answer assessment for intelligent tutoring systems. *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019)*. Linköping Electronic Conference Proceedings 164: 93–99.

judgments on whether a student response is acceptable or not.

We investigate two main factors for SAA performance: 1) the impact of automatic spelling normalization on SAA using a noisy channel approach (Brill and Moore, 2000), and 2) the influence of using different training/test splits, namely 'unseen answers', 'unseen items' (questions), and 'unseen tasks', following Dzikovska et al. (2013).

Overall, results show that using spelling normalization yields superior performance for the SAA system we use, and that the performance gap widens when only using out-of-domain training data ('unseen tasks'). We also conduct a by-task analysis of spelling and non-spelling variants of the SAA system, revealing that normalization effects are not uniform across tasks.

The paper is organized as follows: Section 2 introduces the data source we use for our experiments before section 3 outlines the spelling correction approach. Section 4 then delves into the setup and results of our experiments before section 5 concludes the paper.

Task	input	type	# answers	∅ tokens
2B1	gap-filling	reading	1,511	7.04
3A3a	sentence(s)	reading	463	9.77
1CYP2b	sentence(s)	listening	411	7.83
1ET5	sentence(s)	reading	360	4.68
2CYP3	sentence(s)	reading	255	7.71
1B7b	gap-filling	listening	220	1.79
2C5b	sentence(s)	reading	177	9.24
1AP37	sentence(s)	reading	126	8.90
1AP38	sentence(s)	reading	85	14.15
2ET3	gap-filling	reading	61	2.59
3AP19a	gap-filling	listening	35	1.54
3AP20a	sentence(s)	listening	23	4.13
3AP16a	sentence(s)	listening	17	4.47
2AP34	sentence(s)	listening	15	5.00
3AP32	gap-filling	reading	15	2.27
4AP16	gap-filling	listening	13	8.15
3CYP2b	sentence(s)	listening	9	3.89
2AP33	gap-filling	listening	8	1.25
4AP15b	sentence(s)	listening	8	9.50
4C2	sentence(s)	reading	6	7.83
4B6	gap-filling	listening	5	2.00
3AP33	sentence(s)	reading	2	14.50
4AP17	sentence(s)	listening	2	14.00
4AP31	sentence(s)	listening	1	7.00
6A4	gap-filling	reading	1	1.00
overall			3,829	7.11

Table 1: Data set properties by task

2 Data

Our data comes from the FeedBook (Rudzewitz et al., 2017, 2018; Ziai et al., 2018), an English tutoring system for 7th grade used in German secondary schools as part of a full-year randomized controlled field study (Meurers et al., 2019). The system includes interactive feedback on form for all grammar topics on the curriculum, but also a first version of meaning feedback for meaning-oriented tasks, such as reading and listening comprehension activities.

For our purposes in this paper, we extracted all student responses that were entered in reading or listening tasks where the task objective is meaning-oriented, i.e., comprehension. We excluded duplicate answers. After filtering out answers to tasks that were erroneously classified as meaning-oriented or that require knowledge external to the task material (for example, asking about aspects of the individual student's life), we obtained 3,829 answers entered into 123 answer fields of 25 tasks.

Table 1 lists the tasks in the data set together with the required student input (full sentence(s) vs. gap-filling), comprehension type (reading vs. listening), number of answers, and mean answer token length. The distribution of answers is rather uneven across tasks, with almost 40% of the answers coming from one task. This may be a result of this task being favored by teachers, but reflects real-life usage of the system. On the whole, answers consist of 7.11 tokens on average, with gap-filling tasks typically triggering shorter responses than full sentence tasks.

Figure 1 shows an example gap-filling task for listening comprehension. For the purposes of this paper, we use 'item' to refer to a field that a student can type an answer into, and 'task' refers to the whole exercise that is made up of items and the surrounding context.

In order to obtain a gold standard for our classification approaches to train on, an experienced English teacher rated every response with respect to whether it is an acceptable answer or not. The majority class is 'correct' with a percentage of 62.05% among the 3,829 responses.

3 Task-dependent Spelling Correction

The spelling correction approach we employ is based on the noisy channel model described by Brill and Moore (2000) as implemented by Adri-

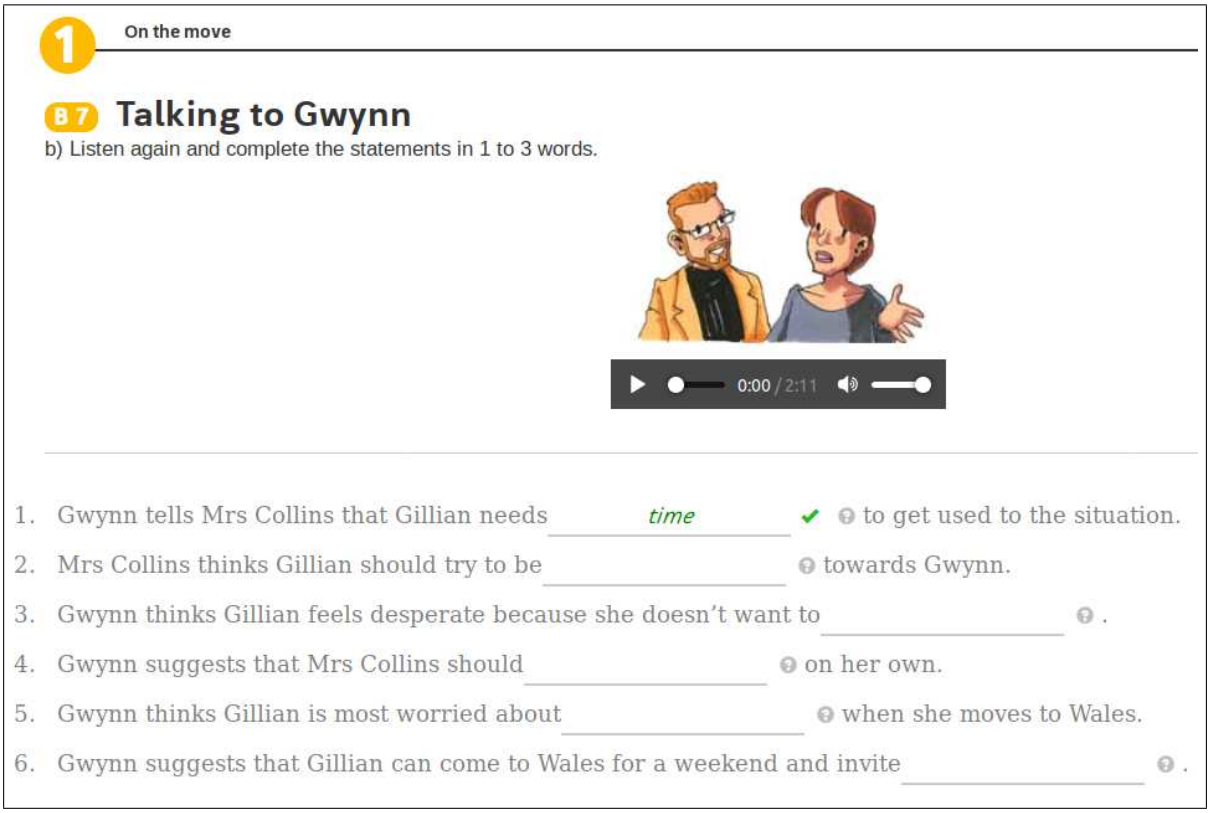

Figure 1: Example listening task

ane Boyd[2]. The approach requires a list of misspellings (non-word/correction pairs) to derive its model from, as well as a dictionary of valid words to draw its suggestions from. Given a non-word, i.e., one that is not found in its dictionary, it returns an n-best list of valid candidate words.

We trained the approach on a list of approximately 10,000 misspellings made by German learners of English, which we extracted from the EFCamDat corpus (Geertzen et al., 2013). The dictionary we used is compiled from the vocabulary list of English school books used in German schools up to 7th grade, approximating the vocabulary that German 7th graders learning English in a foreign language learning setting were exposed to and may use.[3]

In order to make the spelling correction approach somewhat context-aware, we used the weighting of dictionary entries offered by the Brill and Moore approach, giving a weight of 1 to standard entries, and increasing the weight of forms found in the specific task's reading or listening text by their term frequency in that text. As a result of this weighting, task-specific spelling corrections are more likely to happen, given a sufficiently close learner production.

4 Experiments

In this section, we describe the experiments we carried out, and the results obtained.

4.1 Setup

For Short Answer Assessment (SAA), we employed a variant of the CoMiC system (Meurers et al., 2011a). CoMiC is a so-called alignment-based system. It aligns different linguistic units (tokens, chunks, dependencies) of the learner and the target answers to one another and then extracts numeric features based on the number and type of alignments found. The features are then used to train a classifier for new unseen answer pairs.

For the experiments in this paper, we used a Support Vector Machine (SVM) with a polynomial kernel as the classification approach, based on the *kernlab* package (Karatzoglou et al., 2004) in *R* (R Core Team, 2015) via the *caret* machine learning toolkit (Kuhn, 2008). We used default hy-

[2] https://github.com/adrianeboyd/BrillMooreSpellChecker

[3] Naturally, English movies, video games such as Minecraft, and English Let's Play videos are quite popular in the targeted age group and will impact their vocabulary knowledge in a way not captured here.

perparameters for the SVM approach.

Complementing to CoMiC approach, we created a baseline system using nine standard string similarity measures from the *stringdist* package (van der Loo, 2014) in *R*, calculated between student and target response. These similarity scores were used in the same classification setup we used for the CoMiC features.

To incorporate the spelling correction approach described in section 3, we ran it on all student responses as a preprocessing step to obtain a second version of CoMiC enhanced with spelling correction. Apart from this preprocessing, the two CoMiC versions are exactly the same.

Each of the systems just described was given the classification task of determining whether a given response is correct or not, given a prompt and the one or more target answers from the task specification. We used the following test scenarios, roughly following Dzikovska et al. (2013):

- 'unseen answers': tenfold cross-validation across all answers, randomly sampled.
- 'unseen items': for each item, all answers for that item (gap/field) are held out; training is done on all other answers.
- 'unseen tasks': for each task, all answers for that task are held out; training is done on all other answers.

Whereas 'unseen answers' is the most desirable scenario from a computational perspective (training answers for all items are available), 'unseen tasks' is much closer to a real-life situation where educators or material designers add new exercises to the tutoring system for which no pre-scored answers exist. This setting is thus of special importance to a real-life approach.

4.2 Results

We first report and discuss overall results, before diving into a task-specific analysis.

4.2.1 Overall Results

The overall results are shown in Table 2. In addition to the systems described in the previous section, we list the majority baseline ('Majority'). 'CoMiC' is the standard CoMiC system, whereas '+SC' is the variant enhanced by spelling correction preprocessing. We report both accuracy and Cohen's κ (Cohen, 1960).

SAA System	Unseen answers		Unseen items		Unseen tasks	
	%	κ	%	κ	%	κ
Majority	62.05%, κ = 0.00					
stringsim	78.35	0.52	76.97	0.48	75.61	0.45
CoMiC	81.25	0.59	81.20	0.59	80.80	0.58
+SC	**82.63**	**0.62**	**82.63**	**0.61**	**82.45**	**0.61**

Table 2: Overall accuracy (%) and Cohen's κ

All models clearly outperform the majority baseline. The string similarity model is surprisingly strong, showing that many real-life cases can actually be scored with such surface-based methods if one has access to reference answers. However, the majority baseline and the string similarity model are clearly outperformed by CoMiC. This is particularly evident when looking at the κ-values, which include chance correction based on the distribution of labels. Note that CoMiC generalizes much better to 'unseen items' and 'unseen tasks' than the string similarity model, indicating that the higher level of linguistic abstraction bears fruit especially in these settings.

CoMiC is in turn systematically outperformed by its spelling-enhanced counterpart. Interestingly, the performance gap is about the same for 'unseen items' and 'unseen answers', but greater for 'unseen tasks'. This suggests that the effect of spelling correction is more pronounced for out-of-domain training scenarios, which may be due to the fact that the training basis for the spelling correction approach is disjunct from that of the SAA system, and thus does not suffer from generalization problems on this data set.

Since these are the first results on this data set, we cannot directly compare them to any previous ones. Looking at recent related work on similar data, we can see that, e.g., the results of Ziai and Meurers (2018) on reading comprehension data in German are in the same ballpark, though slightly higher. We suspect this is the case because that data was more uniform, both with respect to task diversity and the resulting nature of the answers.

4.2.2 Results by Task

In order to find out more about the effects of adding spelling correction to the CoMiC model, we analyzed the 'unseen tasks' results of 'CoMiC' and 'CoMiC+SC' on a per-task level. These results are listed in Table 3. The tasks are listed in the same order as in Table 1, namely by descend-

Task	CoMiC		CoMiC+SC	
	%	κ	%	κ
2B1	80.15	0.53	**82.46**	**0.57**
3A3a	79.70	0.53	**82.51**	**0.58**
1CYP2b	**88.32**	0.71	88.08	0.71
1ET5	93.33	0.86	**93.61**	**0.87**
2CYP3	72.94	0.45	**75.29**	**0.49**
1B7b	64.09	0.29	**70.45**	**0.42**
2C5b	84.75	0.69	**85.88**	**0.72**
1AP37	**73.81**	**0.44**	70.63	0.38
1AP38	87.06	0.74	87.06	0.74
2ET3	**62.30**	**0.25**	54.10	0.10
3AP19a	88.57	0.60	**91.43**	**0.68**
3AP20a	91.30	0.75	91.30	0.75
3AP16a	82.35	-0.09	82.35	-0.09
2AP34	86.67	0.00	86.67	0.00
3AP32	73.33	0.00	73.33	0.00
4AP16	84.62	0.70	84.62	0.70
3CYP2b	55.56	0.10	55.56	0.10
2AP33	62.50	0.33	62.50	0.33
4AP15b	87.50	0.75	**100.00**	**1.00**
4C2	100.00	1.00	100.00	1.00
4B6	80.00	0.55	80.00	0.55
3AP33	100.00	n/a	100.00	n/a
4AP17	50.00	0.00	50.00	0.00
4AP31	100.00	n/a	100.00	n/a
6A4	100.00	n/a	100.00	n/a

Table 3: Unseen tasks accuracy (%) and κ for CoMiC with and without spelling correction

ing number of answers. For every task, superior results of either model in comparison to the other are marked in **bold**.

The results show that for the task with by far the most answers, '2B1', spelling correction had a very noticeable positive impact (+2.45%). For other tasks, the effect seems to be less pronounced, though still present, e.g., '1ET5'. For some tasks, the effect is actually negative (e.g., '1AP37' and '2ET3'), suggesting that spelling correction introduced additional noise for these tasks. One hypothesis for this phenomenon would be that for these tasks, spelling correction over-corrected wrong answers or non-answers into more acceptable versions, which then got scored better than they should have been. After inspecting concrete normalization cases, we indeed found examples such as the following one for '1AP37':

(1) Prompt: 'Robin ran away because of trouble with his father.'

A_{orig}: 'Robin ran away because of trouble with his stepfather.'

A_{corr}: 'Robin ran away because of trouble with his stepmother.'

Here, the task is to correct the statement in the prompt with the help of a reading text (not shown here). 'stepfather' apparently neither occurred in the general dictionary nor anywhere in the reading text and was thus corrected to 'stepmother', which is wrong in this context and is not aligned to 'stepfather' by CoMiC.

We also suspected that the task properties we showed in Table 1, such as the task type (reading vs. listening), the input (gap-filling vs. sentence(s)), or the mean length of answers would interact in some manner with the addition of spelling correction. For example, very short answers, occurring systematically in gap-filling exercises such as '2ET3', could proportionally be altered more by automatic spelling correction, thus potentially introducing more noise for the SAA classifier. However, this suspicion does not seem to be supported by the results in Tables 1 and 3. For example, both '2B1' and '2ET3' are gap-filling tasks, but while there is a performance gain for the former, there is a drop for the latter.

In search for reasons for the positive impact of spelling correction, we manually inspected some of the student responses given for task '2B1', which is shown in Figure 2, since due to the higher number of answers, the improved result for this task is the most stable. We found that a number of the spelling problems in responses to this task were related to the Welsh proper names introduced by the reading text, such as 'Gruffudd' or 'Llandysul'. These are very hard to spell for 7th grade English learners, but were successfully corrected by our spelling correction approach. Based on this information, we hypothesize that the effect of spelling correction is connected to the lexical material involved in the task rather than its more formal properties. In order to investigate this hypothesis, a systematic analysis of lexical complexity and/or complex word identification (cf., e.g., Yimam et al. 2018) within SAA could be a promising avenue to follow.

5 Conclusion

We presented work in progress on Short Answer Assessment (SAA) on data from the FeedBook, an English language tutoring system we employed in a real-life school setting in Germany. The purpose of SAA in this context is to help the tutoring system decide whether the feedback to be given needs to be positive or negative.

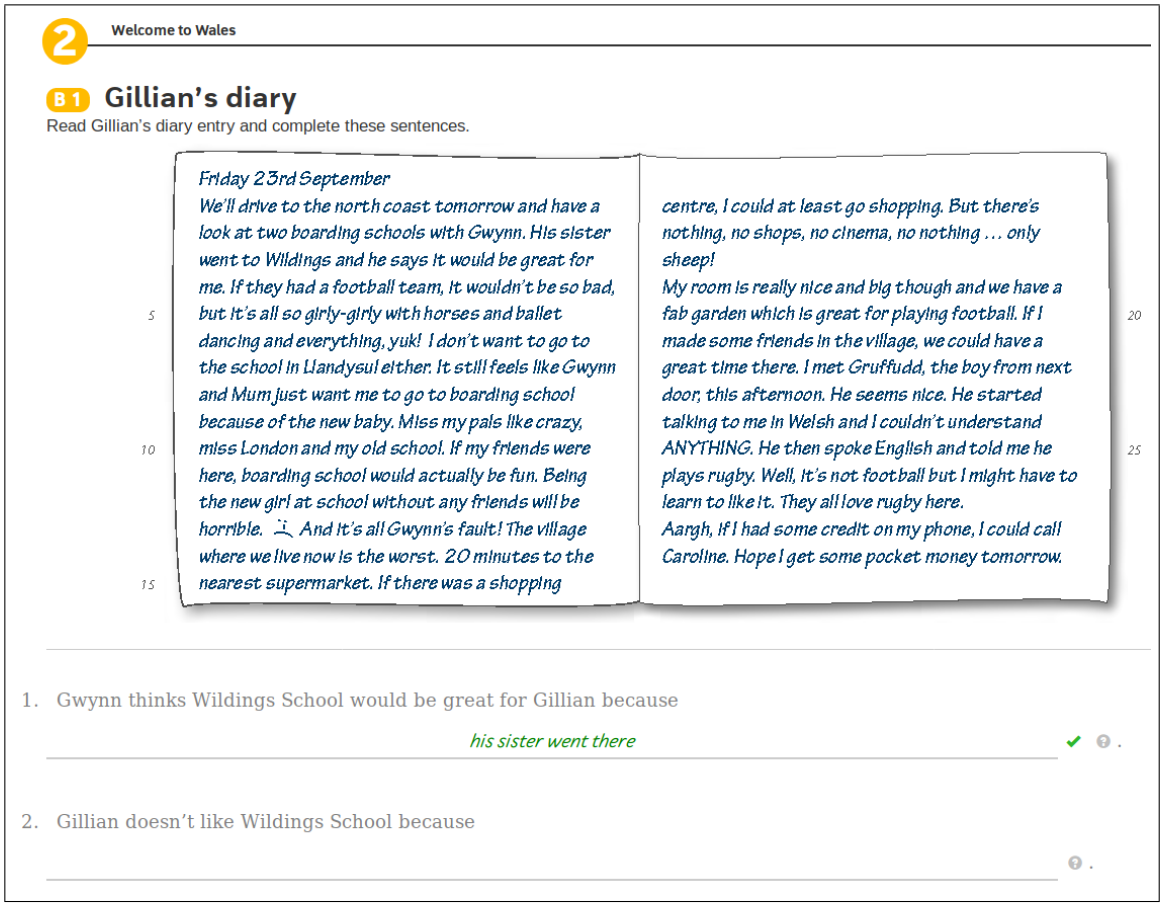

Figure 2: Reading task '2B1' (abbreviated)

To investigate the influence of spelling correction on SAA, we added a noisy channel spelling correction component to a standard SAA approach and found that it generally increases classification performance for the data we collected. In addition, we found that spelling correction helps the SAA system generalize to out-of-domain data.

A task-by-task analysis revealed that the effect of spelling correction is not uniform across tasks. Manual inspection of relevant student responses indicated that this may be related to lexical characteristics of the language employed in the task context. To investigate this hypothesis, it would be interesting to systematically analyze different aspects of lexical complexity, and integrating complex word identification (Yimam et al., 2018) within SAA could be a promising avenue to follow.

Acknowledgments

We would like to thank Louisa Lambrecht for training and tuning the spell checking approach. We also thank the two anonymous reviewers for their helpful comments. Furthermore, we are grateful to the Westermann Gruppe who collaborated with us on the FeedBook project and enabled work such as the one described in this paper, and finally to the Deutsche Forschungsgemeinschaft for funding the project in the first place.

References

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, Hong Kong. ACL.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang

Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum (SLRF)*. Cascadilla Press.

Wael Hassan Gomaa and Aly Aly Fahmy. 2019. Ans2vec: A scoring system for short answers. In *Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)*, pages 586–595, Cham. Springer International Publishing.

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. 2004. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Max Kuhn. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Detmar Meurers, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2019. Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, 39:161–188.

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

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

Niels Ott, Ramon Ziai, Michael Hahn, and Detmar Meurers. 2013. CoMeT: Integrating different levels of linguistic modeling for meaning assessment. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval)*, pages 608–616, Atlanta, GA. ACL.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168, Copenhagen, Denmark. ACL.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, and Detmar Meurers. 2017. Developing a web-based workbook for English supporting the interaction of students and teachers. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition*.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. 2018. Generating feedback for English foreign language exercises. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 127–136. ACL.

M.P.J. van der Loo. 2014. The stringdist package for approximate string matching. *The R Journal*, 6:111–122.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. ACL.

Ramon Ziai and Detmar Meurers. 2018. Automatic focus annotation: Bringing formal pragmatics alive in analyzing the Information Structure of authentic data. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 117–128, New Orleans, LA. ACL.

Ramon Ziai, Björn Rudzewitz, Kordula De Kuthy, Florian Nuxoll, and Detmar Meurers. 2018. Feedback strategies for form and meaning in a real-life language tutoring system. In *Proceedings of the 7th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)*, pages 91–98. ACL.