Summarization Evaluation meets Short-Answer Grading

Margot Mieskes
Hochschule Darmstadt
Germany
margot.mieskes@h-da.de

Ulrike Padó
Hochschule für Technik Stuttgart
Germany
ulrike.pado@hft-stuttgart.de

Abstract

Summarization Evaluation and Short-Answer Grading share the challenge of automatically evaluating content quality. Therefore, we explore the use of ROUGE, a well-known Summarization Evaluation method, for Short-Answer Grading. We find a reliable ROUGE parametrization that is robust across corpora and languages and produces scores that are significantly correlated with human short-answer grades. ROUGE adds no information to Short-Answer Grading NLP-based machine learning features in a by-corpus evaluation. However, on a question-by-question basis, we find that the ROUGE Recall score may outperform standard NLP features. We therefore suggest to use ROUGE within a framework for per-question feature selection or as a reliable and reproducible baseline for SAG.

1 Introduction

Teachers use short free-text questions both in second-language teaching (to evaluate reading comprehension and writing skills) and in content instruction (to probe content understanding and the ability to apply knowledge). Reducing the time needed for grading the answers greatly lightens teacher workloads and allows flexible self-study. Short-Answer Grading (SAG) is the corresponding NLP task of predicting grades for student answers containing up to three sentences.

The most difficult formulation of the SAG problem, which occurs frequently in real-world teaching, is the processing of completely unseen questions and their answers. The prevailing strategy in this situation is to compare student and reference answers and base the grade prediction on any similarities. While very shallow baselines like bag-of-word models are strong for SAG (Dzikovska et al., 2013), they fail to cover deeper levels of meaning. Therefore, features on different levels of language processing have been proposed to solve the central problem of comparing the meaning of two different texts (see Burrows et al. (2015)).[1]

Other NLP tasks facing a similar challenge are Machine Translation evaluation, Natural Language Generation evaluation and Summarization evaluation. Of the three, Summarization evaluation is most closely related to SAG: When determining the quality of an automatic summary, the standard evaluation method ROUGE (derived from Translation evaluation's BLEU) compares candidate summaries against manually created references (Lin, 2004), with the goal of comparing the meaning of the two texts with string-level evaluation tools. Graham (2015) points out that the parameter space of ROUGE is not trivial and that for individual tasks and/or data sets different parameter combinations might give the best results.

In this paper, we exploit the similarities of the tasks by applying ROUGE to SAG. We evaluate on four corpora from the content assessment domain, in English and German. We begin by determining an appropriate, robust set of parameters for ROUGE and by analyzing how well the metric is correlated with the gold grades in the different corpora.[2] We then go on to compare ROUGE with standard SAG features for machine learning. We find that ROUGE is a robust predictor on its own (and could therefore serve as a standardized baseline) and on the question level can outperform the

[1] Recently, neural network approaches have also been explored for educational scoring in general, e.g. Alikaniotis et al. (2016), and SAG in particular (Riordan et al., 2017).

[2] ROUGE results for the parameter sweeping and the ROUGE predictions for our corpora are available at https://bwsyncandshare.kit.edu/dl/fiL6mnSswKKhZttY687GtQgi/MieskesPadoROUGE.zip.

Margot Mieskes and Ulrike Padó 2019. Summarization evaluation meets short-answer grading. *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019)*. Linköping Electronic Conference Proceedings 164: 79–85.

standard SAG features (and is therefore useful for per-question feature selection approaches).

2 Related Work

Within SAG, we follow the research tradition that explores the use of informative features and helpful strategies from other areas of NLP, machine learning and educational research. Examples are the use of features from Information Retrieval, such as text similarity and textual inference (Zesch et al., 2013), the use of the machine learning strategy Active Learning (Horbach and Palmer, 2016) or empirically estimated question difficulty information (Padó, 2017).

ROUGE was presented by Lin (2004) and has since established itself as the de-facto standard evaluation metric in Summarization evaluation used in various summarization related shared tasks[3]. Other metrics have been presented in the past, but none have received a wide-spread usage similar to ROUGE. For an overview of various other methods and their comparison to ROUGE, but also manual evaluations see Louis and Nenkova (2013). ROUGE is based on counting the number of n-grams overlapping in one or several reference text(s) and a comparison text. While n-gram overlap has long been known to be a strong predictor in SAG (see, e.g., Dzikovska et al. (2013)), ROUGE offers a range of other parameters, including skip n-grams, which allow intervening words between the matching words and thus help to cover paraphrases.

ROUGE has been applied in the context of spoken (Loukina et al., 2014) and written (Madnani et al., 2013) learner summaries, thus providing a first bridge from Summarization evaluation to the educational domain. Gütl (2008) proposed the use of ROUGE for SAG in the e-Examiner system, but there is no formal evaluation. ROUGE is demonstrably suited to texts of similar length as short answers: In the DUC-2004 challenge, Task 1 resulted in texts which are at most 75 bytes long and Task 5 aimed at summaries of lengths up to 665 bytes – short answers in our largest data set (ASAP) range between 52 and 500 bytes.

3 Method and Data

We use four SAG corpora (see Table 1) in our experiments. The three English corpora (ASAP, SEB and Beetle) are large enough to have separate test sets for result verification. We use the development sets of SEB and Beetle and the training set of ASAP[4] for finding optimal parameter settings for ROUGE. We evaluate the final parameters on the unseen test sets and on the full data of CSSAG, the smallest corpus. This German corpus allows us to determine how well ROUGE performs across languages.

English Corpora	**Dev #Q/#A**	**Test #Q/#A**
ASAP (`www.kaggle.com/c/asap-sas`)	5/8182	5/2218
SEB (Dzikovska et al., 2013)	15/1070	15/733
Beetle (Dzikovska et al., 2013)	9/1236	9/819
German Corpus		
CSSAG (Padó and Kiefer, 2015)	–	31/1926

Table 1: Corpus sizes and characteristics (source, number of questions and answers in development (ASAP: training) and test sections)

We evaluate the ROUGE predictions by correlating the gold human grades to ROUGE scores using Kendall's τ (Kendall, 1955). The standard Pearson's r is not applicable here, since our data are not normally distributed. We therefore choose a non-parametric correlation method. Specifically, Kendall's τ is less sensitive to ties than Spearman's ρ. Given the small number of grade levels in the human annotations, this property is key for a correlation-based approach. Note that τ as a non-parametric method is more conservative than Pearson's r and will produce smaller coefficient values than r would for the same data sets.

3.1 Experiment 1: Optimal ROUGE Parameters

We experimentally determine the set of ROUGE parameters that yields the best correlation of ROUGE scores against human SAG grades across corpora. As detailed in Graham (2015) there is a wide range of possible combinations. Therefore, our first step is a parameter sweeping experiment to determine the best settings for the following parameters[5]:

Stemming yes/no
Stopwords yes/no
ROUGE variant unigrams to 4-grams, longest common subsequence (LCS) and skip n-grams (S*)

[3] `https://duc.nist.gov/`

[4] We only use the five questions that have explicit reference answers.

[5] We did not experiment with the sampling size (`-r`), as the parameter space was large to begin with.

	ASAP	Beetle	SEB
Stemming	**y**	n	**y**
Stopwords	**n**	**n**	y
ROUGE	**S***	**S***	LCS
Eval Basis	**s**	**s**/t	**s**/t
Model	**best**	all	all
Measure	**R**	$F_{0.5}$	$F_{0.5}$
Conf Level	**95**	**95**/99	**95**/99
optimal τ	0.581	0.469	0.313
final τ	0.581	0.449	0.286

Table 2: Optimal ROUGE parametrizations with corresponding τs and τs for the final parametrization. Final parameter values in bold face.

Evaluation Basis sentences (s), tokens (t) or raw counts (r)
Model average or best
Measure Recall, $F_{0.5}$ or $F_{1.0}$
Confidence Interval 99% or 95%

Stemming and stopwords are options for text pre-processing, intended as rough measures to normalize the input and focus on content words.

The ROUGE measure itself can be calculated in different variants: Four are based on plain n-grams (uni- up to 4-grams), and there are the longest common subsequence (LCS) and skip bigrams model (S*, initially with a skip interval of 4), giving a total of 6 scores. We do not consider ROUGE-W* as it rarely produced stable results.[6]

The evaluation basis can be either ROUGE for all the tokens in the document or the average over sentence ROUGE scores; raw counts can also be output independent of ROUGE.

ROUGE usually evaluates against a number of samples – in a SAG context, this corresponds to having multiple reference answers. The evaluation can then be reported using the average results across all the reference samples for Precision, Recall and F-Score, or just for the best sample. We follow Summarization evaluation practice and experiment with Recall and F-Score, with different weightings of Precision and Recall. Finally, we varied the required confidence interval between 0.99 and 0.95.

ROUGE proved quite robust to many parameter instantiations. There were results for 75% (864) of parametrizations on the Beetle data, and for all parametrizations on SEB. In contrast, though, only 168 (14.5%) out of 1152 possible parameter combinations yielded results for ASAP. Beetle and ASAP evaluations both failed for all runs which use raw counts as the basis of evaluation. This result is unproblematic in practice, since the raw scores are not a standard evaluation tool and are not in focus here. ASAP evaluations additionally failed for all runs that evaluated across all models, and yielded no results in the 0.99 confidence interval. The reason for the difficulties on ASAP may be that the model answers are quite long. The questions ask students to give multiple aspects or key points, and the model answers aim to list many possible correct aspects. However, any given student will answer with just the required number of aspects, so there is usually a relatively large difference between the student answers and the models they are compared to. Despite this drawback, we find throughout that the ROUGE output performs similarly for ASAP as for the other corpora, so it appears justified to use the ASAP data.

Table 2 shows the optimal parameters for the three corpora. While the ROUGE tool was brittle on ASAP, this corpus shows the largest correlation of ROUGE results and human ratings. Inversely, the correlation is lowest for SEB, the corpus without any failed ROUGE runs.

During parameter sweeping, the largest drops in τ compared to the optima are observed (in order) by changing the ROUGE variant, the F weighting and the combination of stemming and stop words (for all three corpora). Worst case, changing to ROUGE-4 on ASAP costs $\Delta\tau = 0.39$, and $\Delta\tau = 0.27$ on Beetle. This is in line with observations from the Summarization community, where the numerically highest scores are usually achieved using ROUGE-1 and the lowest using ROUGE-4. This pattern is ultimately due to sparse data caused by linguistic variation, which greatly reduces the chance of finding exactly matching 4-grams in two different documents compared to unigrams. The changes in F weighting and stemming/stop words cause much smaller drops in the range between $\Delta\tau = 0.1$ and 0.01, underscoring again the robustness of ROUGE performance to variations in parameter settings.

We found several more, stable patterns across parametrizations that helped inform our choice of final parametrization. For each pattern, we also discuss its plausibility in a SAG context.

To begin with the pre-processing steps, **stopwords** alone are detrimental for all three corpora. In combination with **stemming**, they work well for

[6] We also experimented with various weight settings for ROUGE-W.

SEB, but not at all for Beetle and not optimally for ASAP. This possibly points to a domain dependence of stopword lists. Stemming without stopwords is the best setting for ASAP and the second best by a small margin for Beetle and SEB. Since stemming is a step away from pure string comparison, this result is plausible for SAG.

ROUGE-S* using the standard skip of 4 tokens between the elements of a bigram works best for ASAP and Beetle, while LCS outperforms it slightly for SEB. In addition to the standard skip of 4 tokens, we also experimented with 2 and 6 tokens, but found the performance using a skip of size 4 to achieve the best numeric results. As mentioned above, ROUGE-4 is consistently the worst choice across corpora while ROUGE-S* proved to be quite robust. In a SAG context, this result is plausible, as ROUGE-S* flexibly allows paraphrases. In contrast, ROUGE-4 looks for a specific, fairly long sequence. With short answers of 2 to 3 sentences, the probability to find matching 4-grams drops considerably due to linguistic variation. ROUGE-1 fails to show optimal performance, but yields robust slightly lower results across the remaining parameters, in line with observations in Summarization evaluation.

There is a small preference for sentences as **evaluation unit**, while tokens perform just as well for SEB and Beetle. Raw scores, tested for the sake of completeness, lower the correlation for SEB and evaluation on raw scores breaks down for Beetle and ASAP. The standard SAG setting of using the **Best Model**, i.e., using the highest score produced by comparison to any reference is consistent with Beetle and SEB and optimal for ASAP.

While correlations of $F_{0.5}$-Scores with the human grades often are numerically slightly higher than correlations of **Recall** and human grades, the Recall predictions are much more robust across different combinations of parameters. This is plausible for the SAG task, since the recall of n-gram overlap between the student answer and reference answer shows how much of the reference answer content the student replicated. Precision would correspond to predicting a high human grade if the student only produced correct answer portions (but maybe missed important parts of the answer).

The chosen **confidence interval** did not make a difference to the results for SEB and Beetle, but there were no results in the 0.99 interval for ASAP (probably due to the form of the model answers).

Corpus	τ **dev**	τ **test**	**Language**
ASAP	0.581	0.356	
Beetle	0.449	0.306	EN
SEB	0.286	0.223	
CSSAG	–	0.385	DE

Table 3: Correlations between ROUGE predictions and manual grades for seen (dev) and unseen (test) corpus portions. All correlations significant at $p < 0.001$.

Given the optimal parametrizations and our general observations for the English data, we chose the parameters that work for the majority of corpora. The only departure from this rule is our use of Recall, which yields slightly lower figures, but seems overall more robust than F. We use stemming without stopwords, S* with gaps of up to four intervening words, and evaluate on the sentence level using the best model.[7] Incidentally, this is the optimal parametrization for ASAP, and causes only a small drop in τ for Beetle and SEB (see the bottom line in Table 2).

These parameters hardly differ from the most commonly used settings in Summarization evaluation (i.e. as used in DUC 2004). The only deviation from that standard is that we do not include unigrams in the skip-bigram (ROUGE-S*) calculation. This underlines the similarities between the summarization and SAG tasks. From a SAG perspective, the resulting parameters are also plausible given previous work, as discussed above.

3.2 Experiment 2: Robustness of Parameters

We next test the generalizability of these parameters for new data sets and a new language. We first try the test sets of the three English corpora and then the German corpus. For the German data, instead of the stemming step we externally performed lemmatization (using the TreeTagger, Schmid (1995)) to do more justice to German morphology.

Table 3 presents results for the optimal parameter setting determined in Exp.1. The top three rows of the table repeat the development set results for the final parametrization for the three English corpora and show performance on the unseen test

[7] The full parameter set for the ROUGE package is `-n 4 -m -s -2 4 -c 95 -r 1000 -p 0.5 -t 0`. Please note that we performed the lemmatization for the German data offline and removed this parameter when calling ROUGE for the German data.

sets. For all three data sets, performance drops, as must be expected. The most affected data set is ASAP. This was the most brittle corpus in parameter sweeping, so the optimal parameters possibly overfit the training data used for parameter setting. Least affected is SEB, which showed the highest drop between optimal and final parameters. All correlations remain highly significant (and recall that τ is a conservative measure).

For the German corpus, which was not used in parameter sweeping, the correlation is numerically the strongest of all. This allows us to conclude that the parameter set can be ported to another language with a similar outcome as porting to the unseen test portion of the development data. The method is clearly robust when using language-specific preprocessing tools.

In sum, we find that the ROUGE parameters we have determined on the training sets of three English SAG corpora are stable across corpora and languages. However, we find signs of brittleness and overfitting for our largest English corpus, ASAP, which are probably due to the nature of the available model answers. We therefore expect the identified parameters to be portable to new corpora, especially if model and students answers are comparable (as for SEB, Beetle and CSSAG).

3.3 Experiment 3: ROUGE and Standard SAG Features

Our final experiments evaluate the usefulness of the ROUGE predictions in combination with existing features for grade prediction by machine learning. We use the system from Padó (2016), which extracts features on the basis of a range of levels of linguistic analysis, such as n-grams, textual similarity, dependency parses, semantic representations and textual entailment.

We experiment with an SVM and a Random Forest (RF) learner for the correct-incorrect decision. All the corpora we work on provide several target labels representing partial credit. Prediction tasks with many target labels are harder than predicting a small number of labels. Our corpora have nine labels (CSSAG), five labels (ASAP) and two labels (SEB and Beetle, two-task annotation). In order to standardize the difficulty of the annotation task, we normalize the annotation of ASAP and CSSAG to a binary correct-incorrect annotation by labeling as correct all student answers that receive at least the middle label (50% of points).

We report F scores as the standard measure for classification tasks and in accordance with previous work for SEB, Beetle and CSSAG (Dzikovska et al., 2013; Padó, 2016). As mentioned in the Introduction, we consider the hardest instantiation of the label prediction task, the unseen question setting, where any questions in the test set are completely unseen (so no question-specific models can be trained). In order to achieve this, we use leave-one-question-out evaluation on the training portion of ASAP (the provided test data is for seen questions) and on the full (previously unused) CSSAG data. SEB and Beetle have test sets with unseen data.[8]

Table 4 shows evidence of the high unigram baseline for SAG at at least F=59.7 (RF on SEB; F=65.1 SVM) and up to F=86.7 (RF on ASAP). We also report the majority baseline (the performance of a hypothetical classifier that always predicts the more frequent class) as a learning algorithm-independent (low) baseline. The majority baseline is easy to beat for all classifiers and feature sets, but it highlights the strong label imbalance for ASAP, which is mirrored in its high numerical prediction results throughout.

Over all feature sets, the RF classifier deals better with the data than the SVM. The ROUGE scores alone perform robustly, but below the unigram baseline in most cases. They beat it numerically for RF on SEB and CSSAG. This verifies that ROUGE is predictive for the SAG task, and quite strongly in some configurations.

The deeper features in the NLP feature set generally numerically improve performance over the baseline (except for CSSAG and ASAP RF). Using ROUGE scores as features in addition to the NLP-based features yields no significant improvement and mixed trends. Results for CSSAG improve numerically. On the other hand, we see a small drop for both learners on the Beetle data and for ASAP and SEB, we observe a decrease for one learner, but an increase for the other. This indicates that ROUGE incorporates information also found in the standard NLP features. Since we work with ROUGE-S* skip n-grams, we assume that the shared information can be found in the uni-, bi- and trigrams in the standard NLP features.

We further investigate the impact of ROUGE by

[8] Note that our results are therefore not directly comparable to literature results for ASAP, but are comparable to the literature for SEB, Beetle and CSSAG in both evaluation measure and evaluation procedure.

	Majority	Unigram		ROUGE		NLP		NLP+R	
		RF	SVM	RF	SVM	RF	SVM	RF	SVM
ASAP	58.1	86.3	70.1	80.7	64.0	**86.8**	69.4	86.4	69.9
Beetle	42.6	72.8	71.3	60.7	55.1	**73.6**	73.0	71.9	72.6
SEB	43.7	59.7	65.1	61.4	58.1	66.7	65.2	**67.0**	64.7
CSSAG	45.3	66.2	**70.1**	67.7	64.0	67.6	69.4	68.3	69.9

Table 4: Grade prediction F-Scores for the majority and unigram baselines, ROUGE, all NLP features, and NLP+ROUGE. Random Forest (RF) and SVM classifiers. Best result per corpus in bold.

zooming in on performance on the question level for each corpus. We compute prediction F-scores for each question in the test sets (or in the leave-one-out setting) separately. We find that ROUGE alone performs the same or better than the NLP features for 52% of the 31 CSSAG questions (using RF). The standard NLP features always outperform ROUGE for the five ASAP questions, the nine questions from the Beetle test set and the fifteen questions from the SEB test set. However, for Beetle and SEB, we also analysed the questions in the (previously unused) training set by applying leave-one-question-out evaluation (recall that we always use this evaluation strategy for CSSAG and ASAP). ROUGE outperforms the standard NLP features for 16% of the 47 Beetle training questions and 44% of the 135 SEB questions. In sum, ROUGE is a good predictor for a sizeable subset of our data, but for that subset only.

This intriguing picture of a light-weight stand-in for our range of NLP features – but only in some cases – matches up well with findings from Padó (2016), who also found that n-gram (or n-gram and textual similarity) features suffice for reliable grade prediction for 18 out of the 30 CSSAG questions that were considered. Padó (2016) suggested question-specific feature selection to optimize overall system performance and processing effort. In our experiments on CSSAG, ROUGE also outperformed the n-gram features in 11 out of the 16 cases where it beat the NLP features. Taken together, these findings indicate that ROUGE should not be used as an addition to already established feature sets, but that it is a strong candidate for inclusion in a feature selection strategy that could further improve the overall classification result while at the same time simplifying the model. We expect the same to be true for SEB and Beetle.

A second take-away from our results is the possibility of using ROUGE as a well-defined, reproducible baseline for SAG. ROUGE-S* captures much of information present in a bag-of-words baseline while clearly defining implementational detail like the use of stemming and stop words. This increases transparency and reproducibility of results for the community.

4 Conclusions

We presented experiments on the transferability of the ROUGE metric, an established evaluation tool in the Automatic Summarization domain, to the related task of Short Answer Grading. Our first result is a ROUGE parametrization for the SAG task that is stable across corpora and languages and plausible both from the point of view of SAG evaluation and of best practices in the source domain of Summarization.

Our further experiments show that ROUGE robustly predicts human short-answer grades, although it does not add to the performance of existing NLP features. However, on the question level, it can outperform the NLP features and can therefore serve to replace them in a question-specific feature selection strategy to improve overall results at reduced processing effort. We also suggest to use ROUGE as a well-defined and reproducible baseline to be used for future experiments. As the package has been stable for several years and is widely used in the Summarization community, it allows for reproducible experiments – unlike individual baseline implementations which may use a range of undocumented parameters.

4.1 Future Work

There are a range of questions to address in the future. The first would be to extend these experiments to other evaluation metrics from summarization evaluation. In particular, the PYRAMID method, which compares the content, rather than the n-gram overlap, of two texts, might give additional insight by alllowing us to move away from the restrictions of string-level comparison. This could be further extended to include methods from the wider field of Natural Language Generation (NLG).

Another strand of investigation would be to determine the reasons for large variations within some parametrizations. For example, the ASAP data set was overall more brittle to parameter changes. We also found that stopwords helped for some corpora, but harmed performance on others. This could lead to the development of corpus-specific stopword lists.

Additionally, we plan a deeper analysis of which of the questions gave better results using ROUGE and on which questions it performed worse. This could support the development of more differentiated methods for automatic SAG.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of ACL 2016*.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25:60–117.

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 SemEeval-2013,* Atlanta, Georgia.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of EMNLP 2015*.

Christian Gütl. 2008. Moving towards a fully automatic knowledge assessment tool. *International Journal of Emerging Technologies in Learning*.

Andrea Horbach and Alexis Palmer. 2016. Investigating active learning in short-answer scoring. In *Proceedings of BEA-11,* San Diego, California.

Maurice Kendall. 1955. *Rank Correlation Methods*. Hafner Publishing Co., New York.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of ACL Text Summarization Branches Out Workshop*.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Anastassia Loukina, Klaus Zechner, and Lei Chen. 2014. Automatic evaluation of spoken summaries: the case of language assessment. In *Proceedings of BEA-9,* Baltimore, Maryland.

Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha OReilly. 2013. Automated scoring of a summary-writing task designed to measure reading comprehension. In *Proceedings of BEA-8,* Atlanta, Georgia.

Ulrike Padó. 2016. Get semantic with me! The usefulness of different feature types for short-answer grading. In *Proceedings of COLING 2016,* Osaka, Japan.

Ulrike Padó. 2017. Question difficulty – How to estimate without norming, how to use for automated grading. In *Proceedings of BEA-12,* Copenhagen, Denmark.

Ulrike Padó and Cornelia Kiefer. 2015. Short answer grading: When sorting helps and when it doesn't. In *4th NLP4CALL Workshop at Nodalida*, Vilnius, Lithuania.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of BEA-12,* Copenhagen, Denmark.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.*

Torsten Zesch, Omer Levy, Irina Gurevych, and Ido Dagan. 2013. UKP-BIU: Similarity and entailment metrics for student response analysis. In *Proceedings of SemEeval-2013,* Atlanta, Georgia.