Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Determining the Quality of a Student Reflective Response

Wencan Luo
Computer Science Department
University of Pittsburgh
Pittsburgh, PA 15260, USA
wencan@cs.pitt.edu

Diane Litman
Computer Science Department and LRDC
University of Pittsburgh
Pittsburgh, PA 15260, USA
litman@cs.pitt.edu

Abstract

The quality of student reflective responses has been shown to positively correlate with student learning gains. However, providing feedback on reflection quality to students is typically expensive and delayed. In this work, we automatically predict the quality of student reflective responses using natural language processing. With the long-term goal of producing informative feedback for students, we derive a new set of predictive features from a human quality-coding rubric. An off-line intrinsic evaluation demonstrates the effectiveness of the proposed features in predicting reflection quality, particularly when training and testing on different lectures, topics, and courses. An extrinsic evaluation shows that both expert-coded quality ratings and quality predictions based on the new features positively correlate with student learning gain.

Introduction

It is widely recognized that while feedback plays an important role in promoting learning and motivating students (Case 2007), students often do not receive such feedback (Ferguson 2011). This is becoming more severe in large courses (e.g., introductory STEM, MOOCs).

In this work, we propose applying natural language processing (NLP) techniques to predict the quality of student responses to reflection prompts, with the long term goal of providing automatic feedback on reflective response quality. *Reflection prompts* (Boud et al. 2013) have been demonstrated to be effective in improving instructors' teaching quality and students' learning outcomes (Van den Boom et al. 2004; Menekse et al. 2011). Furthermore, the quality of student reflective responses has been shown to positively correlate with student learning gains (Menekse et al. 2011).

We have already built a mobile application[1], to support reflective response collection (Fan et al. 2015) and summarization (Luo et al. 2015; Luo and Litman 2015). The motivation of this work is to provide feedback on the collected responses in the application. The concept of automatic feedback is illustrated by the partial mockup in Fig. 1. On the left side of the figure (the current implementation), a student has typed "everything was confusing" in response to the reflection prompt "Describe what was confusing or needed more detail" about today's lecture. On the right side (the mockup), after using the work reported in this paper to determine that the student's response is of low-quality, the system makes a suggestion for improvement by popping-up, "Could you provide more details?"

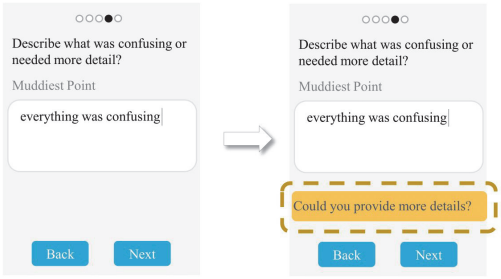

Figure 1: Proposed feedback ("Could you provide more details?") to a low-quality student reflective response.

With this long-term goal of producing informative feedback for students, in this paper we formalize quality prediction as a classification problem and design a new set of features derived from a quality coding rubric provided by domain experts. The contributions of our work are as follows. 1) We propose a new application of NLP techniques motivated by providing automatic feedback on student reflective responses. 2) We develop a new set of features for predicting reflective response quality. Because the proposed features are derived from a human-coding rubric, they are meaningful and interpretable, which should make them useful for producing informative feedback. 3) An intrinsic evaluation on off-line data shows that models which predict reflective response quality with the new features outperform baseline models, especially when training and testing on data from different lectures, topics, and courses. 4) An extrinsic evaluation demonstrates the utility of our features, by showing that correlations between reflective response quality and student learning gain are similar whether using expert quality annotations or automatic predictions.

Related Work

Automatic assessment of student responses is categorized into summative (for providing grades) and formative assess-

[1]Freely available at www.coursemirror.com/download/app

ment (for providing feedback) (Sadler 1989). Our research is motivated by the latter as our long-term goal is to provide interactive feedback to students so as to allow self-regulated learning. Prior research using NLP to trigger interactive feedback after quality assessment in other educational contexts has inspired us. For example, Rahimi et al. (2014) proposed a set of interpretable essay-scoring features derived from a grading rubric for response to text assessment, while Nguyen, Xiong, and Litman (2014) developed a system to provide instant feedback during web-based peer review whenever a student's review was predicted to be of low quality. Similar to our work, one of the functions of SEEDING (Paiva et al. 2014) is to collect student responses and provide a real-time analysis. However, SEEDING was designed to help instructors quickly browse and review responses, while our work aims to deliver immediate feedback to individual students automatically.

Automatic prediction of reflective response quality is somewhat similar to automatic essay scoring (Burstein 2003; Attali et al. 2006; Lee, Gentile, and Kantor 2008; Balfour 2013). Both are used to assign numeric scores to textual responses. However, reflective responses are slightly different from essays. First, reflective response length and granularity range from single words to several sentences. Second, reflective responses highly refer to content included in external information sources such as textbooks, slides, and other course materials. Finally, student reflective responses are scored from a different perspective than most essays: e.g. specificity instead of traits such as grammar, coherence, organization etc.

Automatic short answer grading has also received a lot of attention. It is the task of assessing short responses to objective questions (Burrows, Gurevych, and Stein 2015) and typically aims to search for alternatives of provided "right" answers (Hirschman et al. 2000). In contrast, student responses are subjective and there is no correct answer.

Task Description and Data

To collect the reflective response corpus described in detail below, students were asked to respond to a carefully designed prompt called "muddiest point," "describe what was confusing or needed more detail," at the end of each lecture. The prompt was the same across lectures, but the student responses were different because they were based on individual course content.

After collection, human experts evaluated the reflective responses based on relevance to lecture content and specificity of explanation. The operational scoring rubric (Menekse et al. 2011) is illustrated in Fig. 2.

The coding rubric followed an ordinal scale of 0-3 to indicate the degree of quality of reflections. Examples of coded reflections from the dataset described below are shown in Table 2. 1) A score of "0" was given if the student did not write anything as a muddiest point, or if the student's reflection was completely irrelevant to any class topic, discussion, and/or assignment. For example, "Elephant stampede in a rainstorm" and "Who will ever tell my random thoughts to once in graduate your class" are not related to any class topic. 2) A score of "1" was given to vague reflections, in

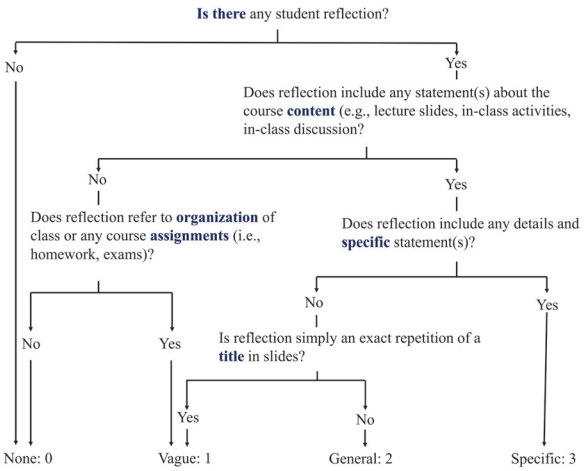

Figure 2: Quality coding rubric

which the student addressed course content, but simply restated one of the broad concepts or titles from a slide. This category also included statements referring to class organization or any course assignment. For example, "size of print and colors are hard to read" is talking about the handout and "Elastic module" is one of the slide titles. 3) A score of "2" was given to general reflections which were neither deep/detailed statements nor simple repetitions of slide titles. 4) A score of "3" in our coding schema referred to deep/specific reflection statements. An example of a reflection that was scored in this category was "computing length, edges and atomic packing factor for FCC." This reflection was very specific about the student's area of concern.

The quality of students' reflections is considered to be indicative of the degree to which the reflections were relatively active or constructive according to Chi's framework (Chi 2009), which in turn correlates to learning gain (Menekse et al. 2011) in the data described below.

Course	Student #	Lecture #	Response #
Spring 2010	27	4	108
Spring 2011	53	23	1149

Table 1: Statistics for the data sets.

The data for our study are student responses collected from two undergraduate courses. Both are introductory materials science and engineering classes, but taught in two different semesters: Spring 2010 and Spring 2011. Additional external resources are also available, including the lecture slides and textbook (Callister and Rethwisch 2010). Data are first preprocessed to convert noisy responses like 'Blank','__', 'N/A', '?' to blank responses based on a manual examination of Spring 2011 data. [2] The statistics of the data are shown in Table 1. The length of the responses varies a lot, as shown in Table 3. Based on a sample of 100 responses from Spring 2011, Kappa for the rubric (between

[2]This affected 72 and 12 responses for Spring 2011 and Spring 2010 respectively.

score=0	‘’ ‘Elephant stampede in a rainstorm’ ‘Not sure if I understand’ ‘Made some kind of sense’ ‘Who will ever tell my random thoughts to once in graduate your class’
score=1	‘size of print and colors are hard to read’ ‘Elastic module’ ‘I tried to follow along but I couldn’t grasp the concepts. Plus it’s hard to see what’s written on the white board when the projector shines on it’
score=2	‘I found a little confusing properties related to bond strength’ ‘The repulsive/ attraction charts’ ‘I didn’t understand the attractive and repulsive force graphs from the third slide’
score=3	‘Part III on worksheet in class, comparing metals. I was confused about why each metal was selected’ ‘computing length, edges and atomic packing factor for FCC’ ‘The working definition of elasticity is not very clear. I think I’m imagining resilience instead’

Table 2: Examples of student responses to the reflection prompt *“Describe what was confusing or needed more detail.”* The left column shows the quality scores given by domain experts. The right column shows student responses belonging to each quality category.

single annotator we use as gold standard and two additional independent annotators) is 0.73.

The Spring 2010 data were analyzed in the study (Menekse et al. 2011). For this study, students took a pre-test before and a post-test after the 4 lectures. After each lecture, students submitted their responses. There are 27 students who regularly submitted their responses and also finished the pre-test and post-test. The quality scores were annotated by domain experts after all the responses were collected according to the rubric (Fig. 2). Thus, each student received 4 quality scores (each for a lecture), ranging from 0 to 3. A positive correlation was found between the total sum of the 4 scores and the learning gain calculated from the pre-test and post-test, $r = 0.473$, $p = 0.013$.

Features

As discussed above, the long-term goal of our research in predicting the quality of student responses is to provide interactive feedback that suggests to students how to write better quality responses. Therefore, we want to design a set of interpretable features that capture the domain knowledge encoded in the coding rubric. To operationalize key decision points in Fig. 2, we extract the following five types of features from student responses, corresponding to each of the key decision points.

Existence (E) is a binary indicator of whether a statement has any words.

Content (C) is defined to capture whether a statement is about course content. We use an exhaustive list of words that appear in the rubric, including *lecture*, *slide*, *activity*, and *discussion*. If any of these keywords (stemmed (Porter 1997)) is present in a statement, the feature is set to True.

Course	score	N	min	max	mean	std
Spring 2010	0	19	0	9	0.6	2.0
	1	20	1	23	8.0	6.6
	2	33	1	11	3.5	2.2
	3	36	2	23	7.1	4.4
Spring 2011	0	447	0	24	1.1	3.4
	1	225	1	36	6.5	6.1
	2	250	2	30	7.9	4.9
	3	227	3	44	13.1	7.1

Table 3: Quality score distribution. ‘N’ is the number of responses; ‘min’, ‘max’, ‘mean’, and ‘std’ show the minimal, maximal, mean, and standard deviation of the number of words in student responses for each score.

Organization and Assignment (OA) captures whether a statement refers to the organization of class or any course assignments. Similar to C feature, we include keywords *organization*, *assignment*, *homework*, *HW*, *exam*, and *examination* to compute this feature.

Specific (S) features are extracted to capture the node “does reflection include any detailed and specific statements”. High-quality response includes specific examples, reasons or an explanation of why some points are confusing. Most existing work (Caraballo and Charniak 1999; Ryu and Choi 2006; Reiter and Frank 2010) focuses on specificity at the term level, which cannot meet our needs. Recently, a freely available tool called Speciteller (Li and Nenkova 2015) was released, which judges whether a sentence is general or specific by including neural network word embedding and word cluster features. It computes how much detail is present in a sentence by giving a rating ranging from 0 (most general) to 1 (most detailed). We use this tool to extract two features derived from a response’s specific rating: the raw rating and a binary decision using the threshold 0.5.

Title (T) features are used to capture “is reflection simply an exact repetition of a title in slides?” We first extract all the titles from the lecture slides automatically. Next, we extract three features in this category: 1) whether a statement repeats part of a title or its entirety; 2) number of title words in a statement; 3) ratio of title words. The title words are defined as words in the titles but not stop words.

The features above are inspired by the rubric (Fig. 2), and thus named as **rubric (R)** features. In addition to these features, we also extracted **Word Count (WC)** and **Lexicon (L)** unigram as features because they are widely used and effective in automatic text scoring (Rahimi et al. 2014; Attali et al. 2006; Lee, Gentile, and Kantor 2008), text classification (Joachims 1998; Rousseau, Kiagias, and Vazirgiannis 2015), etc. In our problem, longer statements tend to have higher quality scores (Table 3). These two types of features are named as **baseline (B)** features.

Experiments

To test how our model performs with the proposed features in different situations, we configure a series of experiments to test the validity of the following hypotheses.

- **H1***: Models with rubric features will outperform or at least perform equally well at predicting reflection quality, compared to models using only baseline features.*
- **H2***: Models with rubric features will transfer better (i.e. have better predictive utility when trained and tested on different lectures, topics, and different courses) compared to models using only baseline features.*
- **H3***: Relationships between reflection quality and student learning gain will be the same whether using human or automatically-predicted quality scores.*

To test H1, we use Spring 2011 data and perform a 10-fold cross-validation (CV) since this data set is larger than Spring 2010; for H2 and H3, we use both Spring 2010 and Spring 2011 data. Among the three hypotheses, H1 and H2 are evaluated intrinsically and the performance measures we report are accuracy, Kappa, and Quadratic Weighted Kappa (QWKappa). For the classifier, we used SVM, with default parameters implemented in Weka. We include Kappa for evaluation since the distribution of quality scores is not even and Kappa measures how the classifier agrees with humans after correcting for chance. Since our quality scores are ordered and incorrect predictions have different costs (e.g., predicting '3' as '1' is more severe than predicting '3' as '2'), we also report QWKappa. H3 is evaluated extrinsically by performing an end-to-end evaluation to test whether we can observe the original correlation between response quality and learning gain when using automatically predicted (rather than manually annotated) scores.

H1

We first examine the hypothesis that adding the rubric-based features will outperform or at least perform equally well as the baseline (H1). It is based on the observation that although the baseline features (WC and Lexicon) captured some information (not exactly the same) of the Existence (E), Content (C), Organization and Assignment (OA) features, or even Specific (S) to some extent, it is hard to capture the title information. Therefore, we hypothesize the rubric features provide more information about the student response and thus yield better performance.

In this experiment, we do 10-fold cross validation and the results are shown in Table 4. The rubric features[3] perform significantly worse than the baseline in terms of accuracy and Kappa. However, adding the rubric features together with the baseline features yields significantly better performance on all three metrics.

Feature	Accuracy	Kappa	QWKappa
Baseline (B)	.693	.567	.794
B+E	.743*	.646*	.846*
B+C	.697*	.573*	.793
B+OA	.693	.567	.794
B+S	.708	.591	.790
B+T	.725*	.616*	.819*
Rubric (R)	.645*	.503*	.759*
ALL (B+R)	**.755***	**.661***	**.859***

Table 4: H1. 10-fold cross-validation within Spring 2011. **Baseline** includes WC and L features; **Rubric** features are **E+C+OA+S+T**; **All** uses both baseline and rubric features. '*' means significantly different from **Baseline** ($p < 0.05$).

Therefore, the hypothesis H1 holds. To judge which types of features help the baseline, we add each type of the rubric features to the baseline. The biggest improvement comes from the Existence (E) feature due to the fact that a large number of responses (31%) have a length of 0. A significant improvement can also be obtained by adding the Title (T) features. It makes sense because the WC and lexical features cannot capture whether a response is an exact repetition of a slide title. In addition, these features are complementary with each other since the model with **All** features are significantly better than all other combinations of features in Table 4, with the only exception that the improvement over **B+E** is not statistically significant for accuracy and Kappa.

H2

Although WC and Lexicon features are widely used in text-based scoring systems, they are criticized a lot due to poor generality. In contrast, the rubric features are derived only from the rubric decision tree and do not rely on ad-hoc words. Therefore, we hypothesize that the rubric features transfer better to different lectures, topics, and courses (H2), which can be divided into three sub-hypotheses.

First, the rubric features transfer better to different lectures (H2.a). In the experiment to examine H1, we used 10-fold cross-validation, that is, all the responses are randomly divided into 10 folds. 9 of them are used to train the model and the remaining one is used to test it. However, this setting might favor the baseline with the lexical features because students may use the same words in the same lecture. This is indeed true and that is one of the reasons why the baseline did well in the H1 setting. For example, in one lecture, 16 out of 53 students used the words "phrase diagram" because that is the most frequent confusing point in that lecture. Therefore, we perform leave-one-lecture-out evaluation (in total, there are 23 lectures) so that responses in the training and testing come from different lectures. The results are shown in Table 5. As we can see, the model of rubric features is comparable as the baseline now (they are not statistically significant different any more). At the same time, the state-

[3] Among responses coded as '0', 16 of them in Spring 2010 and 358 in Spring 2011 have a word count of 0 (the left path from the root of Fig. 2). We thus also explored a hybrid approach where machine-learning was done without the feature **E** and the empty responses were removed from the training data. During testing, if **E** was True, the score was assigned to be '0' in a rule-based manner, while in all other cases the score was predicted by the learned model. For all hypotheses, this hybrid approach yielded similar results as the approach described in the text (using all Rubric features and all data for training) so will not be described further.

ment that adding the rubric features improves the baseline significantly still holds under the cross-lecture setting.

Second, the rubric features transfer better to different topics (H2.b). Although the leave-one-lecture-out setting eliminates the lexicon overlap to some extent, the responses in different lectures still might be the same because different lectures may cover similar topics. For example, lectures 4, 5, 6, and 7 are all about "the structure of crystalline solids" and therefore the term "unit cells" often appears in students' responses in all four lectures. To solve this issue, we group the 23 lectures into 8 topics, and each topic corresponds to one of the chapters of the textbook (Callister and Rethwisch 2010). Then, we perform the leave-one-topic-out evaluation, as shown in Table 5. The model with rubric features performs better than the baseline in terms of Kappa and QWKappa (although not statistically significant). Again, adding the rubric features improves the baseline significantly. Note that, the relative gain[4] by adding rubric features over the baseline is larger (12% vs. 19% for QWKappa) than H2.a.

Third, the rubric features transfer better to different courses (H2.c). In practice, we will deploy our model to different courses, and thus it is better to evaluate it on different courses beyond topics. Therefore, we train all the models with the data Spring 2011 and test them with Spring 2010 to simulate this situation. We did not perform the experiment vice versa, because the Spring 2010 data is much smaller. Strictly speaking, it is not an ideal evaluation data set because the Spring 2010 and Spring 2011 courses were taught by the same instructor using the same textbook. However, they were in different semesters and thus have different students. This is the best data we can have so far. We will address this issue in the future when we collect and annotate more data for other courses. The results are shown in Table 5. The model with only rubric features performs better than the baseline as does adding the rubric features to the baseline[5]. Interestingly, the rubric features alone now yield the best performance in terms of accuracy and Kappa.

To sum up, in general, the new rubric-based features transfer better to different lectures, topics, and courses than the baseline features. More importantly, it shows relatively stronger performance than the baseline for the setting from different lectures to different courses.

H3

Another interesting question is whether relationships between reflection quality and student learning gain will be the same using human and automatically-predicted quality scores. In this experiment we thus repeat the analysis in Menekse et al. (2011) described above, but now using automatically predicted scores instead of human-coded scores. The model is trained under the same setting used in H2.c. Then we do the Pearson correlation test between the total sum of predicted quality scores and the learning gain. The results are shown in Table 6. The model with all the features (which yielded the best QWkappa for H2.c) is the only prediction model that yields a statistically significant positive correlation. Therefore, H3 partially holds. The predicted quality of student responses still shows a positive learning gain using the model with all the features. It also suggests that in a real application, it is better to include both the baseline features and the rubric features.

[4]If the performances of two models are X and Y, the relative gain of Y over X is defined as $(Y - X)/X$.

[5]No significance tests since it is a held-out experiment.

	Feature	Accuracy	Kappa	QWKappa
H1	Baseline	.693	.567	.794
10-fold CV	Rubric	.645*	.503*	.759*
	ALL	**.755***	**.661***	**.859***
H2.a	Baseline	.657	.499	.745
CrossLecture	Rubric	.645	.494	.741
	All	**.731***	**.622***	**.834***
H2.b	Baseline	.634	.463	.699
CrossTopic	Rubric	.630	.473	.730
	All	**.729***	**.621***	**.835***
H2.c	Baseline	.315	.119	.361
CrossCourse	Rubric	**.472**	**.292**	.439
	All	.352	.151	**.515**

Table 5: H2. H2.a and H2.b use Spring 2011 only, including 23 lectures in H2.a and 8 topics in H2.b. For H2.c, Spring 2011 is the training set and Spring 2010 is the testing set. '*' means significantly different from **Baseline** ($p < 0.05$).

	Pearson Correlation	p
Human coded	0.473*	0.013
Baseline	0.341	0.082
Rubric	0.294	0.137
All	0.394*	0.042

Table 6: H3. '*' means the correlation is statistically significant ($p < 0.05$).

Conclusion and Future Work

In this work, we presented an approach to automatically predict the quality of a student reflective response that was motivated by a future goal of providing automatic feedback to students. In particular, we designed a new set of features derived from a quality coding rubric developed by domain experts, which we used in conjunction with baseline NLP features. Both intrinsic and extrinsic evaluations demonstrate the effectiveness of our approach.

In the future, we plan to try better ways to operationalize the rubric, for example, training our own model for the specificity decision point, using semi-supervised approach to extract the keyword features, and making use of other external resources such as textbooks. With the deployment of our mobile application, we are able to collect and annotate more data and revisit the third hypothesis with new and different courses. In addition, by incorporating our prediction model into our mobile application and using its output to generate feedback, we can see whether providing automatic feedback improves the quality of reflection or not, and whether students gain more with such feedback.

Acknowledgments

This research is supported by an internal grant from the LRDC at the University of Pittsburgh. We thank Muhsin Menekse for providing the data set.

References

Attali, Y.; Burstein, J.; Attali, Y.; Burstein, J.; Russell, M.; Hoffmann, D. T.; Attali, Y.; and Burstein, J. 2006. Automated essay scoring with e-rater® v.2.0. *Journal of Technology, Learning, and Assessment*.

Balfour, S. P. 2013. Assessing writing in moocs: Automated essay scoring and calibrated peer review. *Research & Practice in Assessment* 8(1):40–48.

Boud, D.; Keogh, R.; Walker, D.; et al. 2013. *Reflection: Turning experience into learning*. Routledge.

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

Burstein, J. 2003. The e-rater® scoring engine: Automated essay scoring with natural language processing.

Callister, W. D., and Rethwisch, D. G. 2010. *Materials science and engineering: An introduction*, volume 8. Wiley New York.

Caraballo, S. A., and Charniak, E. 1999. Determining the specificity of nouns from text. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 63–70.

Case, S. 2007. Reconfiguring and realigning the assessment feedback processes for an undergraduate criminology degree. *Assessment & Evaluation in Higher Education* 32(3):285–299.

Chi, M. T. 2009. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science* 1(1):73–105.

Fan, X.; Luo, W.; Menekse, M.; Litman, D.; and Wang, J. 2015. CourseMIRROR: Enhancing large classroom instructor-student interactions via mobile interfaces and natural language processing. In *Works-In-Progress of ACM Conference on Human Factors in Computing Systems*. ACM.

Ferguson, P. 2011. Student perceptions of quality feedback in teacher education. *Assessment & Evaluation in Higher Education* 36(1):51–62.

Hirschman, L.; Breck, E.; Light, M.; Burger, J. D.; and Ferro, L. 2000. Automated grading of short-answer tests.

Joachims, T. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, 137–142. London, UK: Springer-Verlag.

Lee, Y.-W.; Gentile, C.; and Kantor, R. 2008. Analytic scoring of toefl® cbt essays: Scores from humans and e-rater®. *ETS Research Report Series* 2008(1):i–71.

Li, J. J., and Nenkova, A. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*.

Luo, W., and Litman, D. 2015. Summarizing student responses to reflection prompts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1955–1960. Lisbon, Portugal: Association for Computational Linguistics.

Luo, W.; Fan, X.; Menekse, M.; Wang, J.; and Litman, D. 2015. Enhancing instructor-student and student-student interactions with mobile interfaces and summarization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 16–20.

Menekse, M.; Stump, G.; Krause, S. J.; and Chi, M. T. 2011. The effectiveness of students daily reflections on learning in engineering context. In *Proceedings of the American Society for Engineering Education (ASEE) Annual Conference*.

Nguyen, H.; Xiong, W.; and Litman, D. 2014. Classroom evaluation of a scaffolding intervention for improving peer review localization. In *Intelligent Tutoring Systems*, 272–282. Springer.

Paiva, F.; Glenn, J.; Mazidi, K.; Talbot, R.; Wylie, R.; Chi, M. T.; Dutilly, E.; Helding, B.; Lin, M.; Trickett, S.; et al. 2014. Comprehension seeding: Comprehension through self explanation, enhanced discussion, and inquiry generation. In *Intelligent Tutoring Systems*, 283–293. Springer.

Porter, M. F. 1997. Readings in information retrieval. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. chapter An Algorithm for Suffix Stripping, 313–316.

Rahimi, Z.; Litman, D. J.; Correnti, R.; Matsumura, L. C.; Wang, E.; and Kisa, Z. 2014. Automatic scoring of an analytical response-to-text assessment. In *Intelligent Tutoring Systems*, 601–610. Springer.

Reiter, N., and Frank, A. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, 40–49.

Rousseau, F.; Kiagias, E.; and Vazirgiannis, M. 2015. Text categorization as a graph classification problem. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1702–1712.

Ryu, P.-M., and Choi, K.-S. 2006. *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. Association for Computational Linguistics. chapter Taxonomy Learning using Term Specificity and Similarity, 41–48.

Sadler, D. R. 1989. Formative assessment and the design of instructional systems. *Instructional science* 18(2):119–144.

Van den Boom, G.; Paas, F.; Van Merrienboer, J. J.; and Van Gog, T. 2004. Reflection prompts and tutor feedback in a web-based learning environment: effects on students' self-regulated learning competence. *Computers in Human Behavior* 20(4):551 – 567.