



## Introduction and Research Goal



- Peer feedback has mixed quality w.r.t helpfulness
- Helpful feedback should **identify problem** and **provide solution** [3]



- Instructor guidance and review prompts have limited impact
- Post-hoc identification of low quality-reviews is time consuming

**SWoRD**  
Peer Assessment

- We are enhancing SWoRD [1] peer review system to help reviewers improve their feedback
- Technologies (e.g. NLP, ML) enable automate prediction of feedback helpfulness and its signals (e.g. problem identification, solution providence) [6]

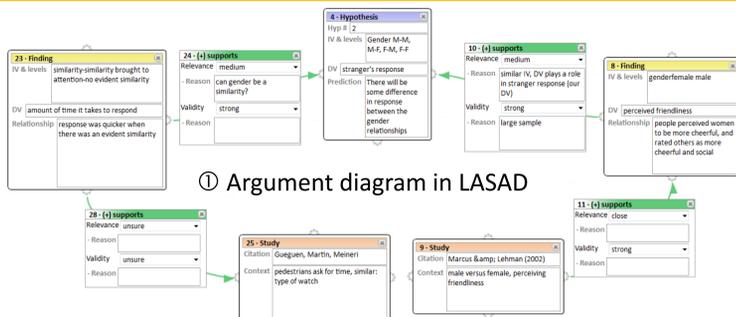
### Research goal

- Prediction models for peer review comments of different feedback types
- A prediction framework to support real-time tutoring about feedback quality [4]

### Approach: develop prediction models that work at sentence level

- Provide fine-grained prediction output for instructional feedback and visualization
- Comment labeling is improved by aggregating from predicted sentence labels

## Argument Diagram Peer Review Corpus



① Argument diagram in LASAD

② An annotated peer comment

③ Comment label dist.

④ Sentence label dist.

<IU><Pr>Not all of these field are filled out, which makes it hard to get a clear idea of how legit these studies are.</Pr></IU>  
<IU><Pr>Also, some are unclear. An example is 24-supports where the reason is a question.</Pr><SI>I think there should be a substantial reason there instead of a question to convince me why it is relevant.</SI></IU>

Makers: <IU> idea unit, <Pr> problem idea, <SI> solution idea

### Annotation scheme

- Comments are labeled regarding whether they have **problem** ideas, **solution** ideas, **combined** ideas (both problem and solution), or they are **non-criticism**
- Text spans that explain comment labels are highlighted
  - Obtain sentence labels: {**problem**, **solution**, **non-criticism**}

Label	#comments	Label	#sentences
Solution	178	Solution	389
Problem	194	Problem	458
Combined	135	Non-criticism	1061
Non-criticism	524	Total	1908
Total	1031		

## Example of Student Comments and System Intervention

### ① Review prompts and comments

**#1.** Are any parts of the diagram hard to understand because they are unclear? If so, describe any particularly confusing parts of the diagram and suggest ways to increase clarity.

The argument diagram was easy to follow. I was able to effortlessly go through the diagram and connect each part.

**#3.** Are the relevance, validity, and reason fields in the supportive arcs complete and convincing? If not, indicate where the argument for relevance or validity is missing or unclear. Suggest ways to make the validity or relevance argument more convincing or sensible.

*Not all of these field are filled out, which makes it hard to get a clear idea of how legit these studies are. Also, some are unclear. An example is 24-supports where the reason is a question. I think there should be a substantial reason there instead of a question to convince me why it is relevant.*

**#5.** Is at least one credible opposing Finding, Study, or Theory connected to each Hypothesis? If there is no opposition, suggest a spot for a potential counterargument. If there is opposition, is it credible? If the opposition is not credible,

*There is a good piece of credible opposition, though it is hard to tell from the diagram what the study exactly did.*

### ② System scaffolding intervention

*Your comments need to suggest solution.*

For every comment below, if you point out a problem make sure that you provide a solution to fix that problem.

Your comments which are highlighted in **GREEN** already have solutions provided, while the **RED** comments mention only problem. Examples of *problem* and *solution* text are formatted in *ITALIC* and **BOLD**.

### ③ Student response

I've revised my comments. Please check again.

Could you show me some examples of problem and solution comments?

My comments don't have the issue that you describe. Please submit comments.

### Fine-grained prediction enables detailed scaffolding intervention

- Prediction models determine whether comments contain *problem* or *solution* ideas
- Problem and solution text spans are italicized and bolded respectively to help students during scaffolding
- Students are asked to revise their comments to provide solution

## Experimental Setup

- Prediction task:** classify review comments regarding feedback types
- Compare two approaches**
  - Trained using comments (CTrain)** - learns prediction model using labeled review comments
  - Trained using sentences (STrain)** - learns prediction model using labeled sentences then aggregates sentence prediction output to create comment labels
- Prediction features**
  - Ngrams: *word and POS ngrams, word-POS pairs, punctuations, word count*
  - Dependency parse to capture structure cues
  - Ignore domain words and metadata: *diagram keywords, review prompts, comment/sentence orders*
- Evaluation:** multi-fold cross validation

**Hypothesis: STrain yields better predictive performance than Ctrain**

(proven in next two sections)

### References

- K. Cho and C. D. Schunn. *Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system*. Computers and Education, 48(3), 2007.
- A. B. Goldberg, N. Fillmore, D. Andrzejewski, Z. Xu, B. Gibson, and Xi. Zhu. *May all your wishes come true: A study of wishes and how to recognize them*. NAACL-HLT, 2009.
- M. M. Nelson and C. D. Schunn. *The nature of feedback: how different types of peer feedback affect writing performance*. Instructional Science, 37(4), 2009.
- H. Nguyen, W. Xiong, and D. Litman. *Classroom evaluation of a scaffolding intervention for improving peer review localization*. 12th ITS, 2014.
- J. Ramanand, K. Bhavsar, and N. Pedanekar (2010). *Wishful thinking: finding suggestions and "buy" wishes from product reviews*. NAACL-HLT Workshop, 2010.
- W. Xiong, D. J. Litman, and C. D. Schunn. *Natural language processing techniques for researching and improving peer feedback*. Journal of Writing Research, 4(2), 2012.

## Three Binary Classification Tasks

Model	Problem v. Non-problem		Solution v. Non-solution		Criticism v. Non-criticism	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Majority	0.68	-	0.70	-	0.51	-
CTrain	0.76	0.43	0.87	0.67	0.83	0.66
STrain	<b>0.81</b> (p<0.05)	<b>0.55</b> (p<0.05)	<b>0.88</b>	<b>0.71</b> (p<0.1)	<b>0.85</b>	<b>0.70</b>

- Problem v. Non-problem**
  - Problem class = problem + combined comments
- Solution v. Non-solution**
  - Solution class = solution + combined comments
- Criticism v. Non-Criticism**
  - Criticism class = problem + solution + combined comments

**Aggregating sentence labels improves comment feedback type prediction**

## Three-way Classification Task

Model	Accuracy	Kappa
Majority	0.51	-
Ctrain	0.76	0.60
STrain	<b>0.79</b>	<b>0.66</b> (p<0.05)

- Three-way classification** (following personal communication with the authors of [5])
  - To be implemented in SWoRD (enabling the our envisioning example in the previous section)
  - Solution-yes** class = solution + combined comments
  - Problem-only** class = problem – combined comments
  - Non-criticism** class = non-criticism comments
- Comment relabeling algorithm for **STrain** approach
  - Each sentence runs through 3 building-block prediction models as described above
  - For each sentence model, aggregates sentence labels to create comment labels.
  - For each comment that has three possible labels { (non-) criticism, (non-) solution, (non-)problem }
    - Labels **Solution** if aggregated so by **Sentence solution model**. Else:
    - Labels **Non-criticism** if aggregated so by **Sentence criticism model**. Else:
    - Labels **Problem** if aggregated so by **Sentence problem model**. Else:
    - Labels **Solution**.

**Sentence classification + comment relabeling are for multiclass feedback type classification**

## Conclusions and Future Work

- Proven advantages of sentence classification for peer feedback type prediction
  - Provide fine-grained output for instructional feedback and visualization
  - Improve peer feedback type prediction performance
- Proposed prediction features are generic and show potential generality
- Sentence-level annotation introduces minimal additional cognitive workload
  - Annotators have already read comment and noticed text spans

### Future work

- Analyze prediction features
- Human-engineered rules for solution and problem text [2, 5]
- Evaluate the approach in paper peer review data