Understanding Differences in Perceived Peer-Review Helpfulness using Natural Language Processing

Wenting Xiong
University of Pittsburgh
Department of Computer Science
Pittsburgh, PA, 15260
wex12@cs.pitt.edu

Diane Litman
University of Pittsburgh
Department of Computer Science &
Learning Research and Development Center
Pittsburgh, PA, 15260
litman@cs.pitt.edu

Abstract

Identifying peer-review helpfulness is an important task for improving the quality of feedback received by students, as well as for helping students write better reviews. As we tailor standard product review analysis techniques to our peer-review domain, we notice that peer-review helpfulness differs not only between students and experts but also between types of experts. In this paper, we investigate how different types of perceived helpfulness might influence the utility of features for automatic prediction. Our feature selection results show that certain low-level linguistic features are more useful for predicting student perceived helpfulness, while high-level cognitive constructs are more effective in modeling experts' perceived helpfulness.

1 Introduction

Peer review of writing is a commonly recommended technique to include in good writing instruction. It not only provides more feedback compared to what students might get from their instructors, but also provides opportunities for students to practice writing helpful reviews. While existing web-based peer-review systems facilitate peer review from the logistic aspect (e.g. collecting papers from authors, assigning reviewers, and sending reviews back), there still remains the problem that the quality of peer reviews varies, and potentially good feedback is not written in a helpful way. To address this issue, we propose to add a peer-review helpfulness model to current peer-review systems, to automatically predict peer-review helpfulness based on features mined from textual reviews using Natural Language Processing (NLP) techniques. Such an intelligent component could enable peer-review systems to 1) control the quality of peer reviews that are sent back to authors, so authors can focus on the helpful ones; and 2) provide feedback to reviewers with respect to their reviewing performance, so students can learn to write better reviews.

In our prior work (Xiong and Litman, 2011), we examined whether techniques used for predicting the helpfulness of product reviews (Kim et al., 2006) could be tailored to our peer-review domain, where the definition of helpfulness is largely influenced by the educational context of peer review. While previously we used the average of two expert-provided ratings as our gold standard of peer-review helpfulness[1], there are other types of helpfulness rating (e.g. author perceived helpfulness) that could be the gold standard, and that could potentially impact the features used to build the helpfulness model. In fact, we observe that peer-review helpfulness seems to differ not only between students and experts (example 1), but also between types of experts (example 2).

In the following examples, students judge helpfulness with discrete ratings from one to seven; experts judge it using a one to five scale. Higher ratings on both scales correspond to the most helpful reviews.

Example 1:

Student rating = 7, Average expert rating = 2 *The*

[1] Averaged ratings are considered more reliable since they are less noisy.

author also has great logic in this paper. How can we consider the United States a great democracy when everyone is not treated equal. All of the main points were indeed supported in this piece.

Student rating = 3, Average expert rating = 5 *I thought there were some good opportunities to provide further data to strengthen your argument. For example the statement "These methods of intimidation, and the lack of military force offered by the government to stop the KKK, led to the rescinding of African American democracy." Maybe here include data about how . . . (126 words)*

Example 2:

Writing-expert rating = 2, Content-expert rating = 5
Your over all arguements were organized in some order but was unclear due to the lack of thesis in the paper. Inside each arguement, there was no order to the ideas presented, they went back and forth between ideas. There was good support to the arguements but yet some of it didnt not fit your arguement.

Writing-expert rating = 5, Content-expert rating = 2
First off, it seems that you have difficulty writing transitions between paragraphs. It seems that you end your paragraphs with the main idea of each paragraph. That being said, . . . (173 words) As a final comment, try to continually move your paper, that is, have in your mind a logical flow with every paragraph having a purpose.

To better understand such differences and investigate their impact on automatically assessing peer-review helpfulness, in this paper, we compare helpfulness predictions using our many different possibilities for gold standard ratings. In particular, we compare the predictive ability of features across gold standard ratings by examining the most useful features and feature ranks using standard feature selection techniques. We show that paper ratings and lexicon categories that suggest clear transitions and opinions are most useful in predicting helpfulness as perceived by **students**, while review length is generally effective in predicting **expert** helpfulness. While the presence of praise and summary comments are more effective in modeling **writing-expert** helpfulness, providing solutions is more useful in predicting **content-expert** helpfulness.

2 Related Work

To our knowledge, no prior work on peer review from the NLP community has attempted to automatically predict peer-review helpfulness. Instead, the NLP community has focused on issues such as highlighting key sentences in papers (Sandor and Vorndran, 2009), detecting important feedback features in reviews (Cho, 2008; Xiong and Litman, 2010), and adapting peer-review assignment (Garcia, 2010). However, many NLP studies have been done on the helpfulness of other types of reviews, such as product reviews (Kim et al., 2006; Ghose and Ipeirotis, 2010), movie reviews (Liu et al., 2008), book reviews (Tsur and Rappoport, 2009), etc. Kim et al. (2006) used regression to predict the helpfulness ranking of product reviews based on various classes of linguistic features. Ghose and Ipeirotis (2010) further examined the socio-economic impact of product reviews using a similar approach and suggested the usefulness of subjectivity analysis. Another study (Liu et al., 2008) of movie reviews showed that helpfulness depends on reviewers' expertise, their writing style, and the timeliness of the review. Tsur and Rappoport (2009) proposed RevRank to select the most helpful book reviews in an unsupervised fashion based on review lexicons.

To tailor the utility of this prior work on helpfulness prediction to educational peer reviews, we will draw upon research on peer review in cognitive science. One empirical study of the nature of peer-review feedback (Nelson and Schunn, 2009) found that feedback implementation likelihood is significantly correlated with five feedback features. Of these features, *problem localization* —pinpointing the source of the problem and/or solution in the original paper— and *solution* —providing a solution to the observed problem— were found to be most important. Researchers (Cho, 2008; Xiong and Litman, 2010) have already shown that some of these constructs can be automatically learned from textual input using Machine Learning and NLP techniques. In addition to investigating what properties of textual comments make peer-review helpful, researchers also examined how the comments produced by students versus by different types of experts differ (Patchan et al., 2009). Though focusing on differences between what students and experts

produce, such work sheds light on our study of students' and experts' helpfulness ratings of the same student comments (i.e. what students and experts value).

Our work in peer-review helpfulness prediction integrates the NLP techniques and cognitive-science approaches mentioned above. We will particularly focus on examining the utility of features motivated by related work from both areas, with respect to different types of gold standard ratings of peer-review helpfulness for automatic prediction.

3 Data

In this study, we use a previously annotated peer-review corpus (Nelson and Schunn, 2009; Patchan et al., 2009) that was collected in an introductory college history class using the freely available web-based peer-review SWoRD (Scaffolded Writing and Rewriting in the Discipline) system (Cho and Schunn, 2007). The corpus consists of 16 papers (about six pages each) and 189 reviews (varying from twenty words to about two hundred words) accompanied by numeric ratings of the papers. Each review was manually segmented into idea units (defined as contiguous feedback referring to a single topic) (Nelson and Schunn, 2009), and these idea units were then annotated by two independent annotators for various coding categories, such as feedback type (*praise, problem, and summary*), *problem localization*, *solution*, etc. For example, the second case in Example 1, which only has one idea unit, was annotated as $feedbackType = problem$, $problemlocalization = True$, and $solution = True$. The agreement (Kappa) between the two annotators is 0.92 for FeedbackType, 0.69 for localization, and 0.89 for solution.[2]

Our corpus also contains author provided back evaluations. At the end of the peer-review assignment, students were asked to provide back evaluation on each review that they received by rating review helpfulness using a discrete scale from one to seven. After the corpus was collected, one writing expert and one content expert were also asked to rate review helpfulness with a slightly different scale from one to five. For our study, we will also compute the average ratings given by the two experts, yielding four types of possible gold-standard ratings of peer-review helpfulness for each review. Figure 1 shows the rating distribution of each type. Interestingly, we observed that expert ratings roughly follow a normal distribution, while students are more likely to give higher ratings (as illustrated in Figure 1).

4 Features

Our features are motivated by the prior work introduced in Section 2, in particular, NLP work on predicting product-review helpfulness (Kim et al., 2006), as well as work on automatically learning cognitive-science constructs (Nelson and Schunn, 2009) using NLP (Cho, 2008; Xiong and Litman, 2010). The complete list of features is shown in Table 3 and described below. The **computational linguistic features** are automatically extracted based on the output of syntactic analysis of reviews and papers[3]. These features represent structural, lexical, syntactic and semantic information of the textual content, and also include information for identifying certain important cognitive constructs:

- **Structural features** consider the general structure of reviews, which includes review length in terms of tokens (*reviewLength*), number of sentences (*sentNum*), the average sentence length (*sentLengthAve*), percentage of sentences that end with question marks (*question%*), and number of exclamatory sentences (*exclams*).

- **Lexical features** are counts of ten lexical categories (Table 1), where the categories were learned in a semi-supervised way from review lexicons in a pilot study. We first manually created a list of words that were specified as signal words for annotating *feedbackType* and *problem localization* in the coding manual; then we supplemented the list with words selected by a decision tree model learned using a Bag-of-Words representation of the peer reviews.

[2]For Kappa value interpretation, Landis and Koch (1977) propose the following agreement standard: 0.21-0.40 = "Fair"; 0.41-0.60 = "Moderate"; 0.61-0.80 = "Substantial"; 0.81-1.00 = "Almost Perfect". Thus, while localization signals are more difficult to annotate, the inter-annotator agreement is still substantial.

[3]We used MSTParser (McDonald et al., 2005) for syntactic analysis.

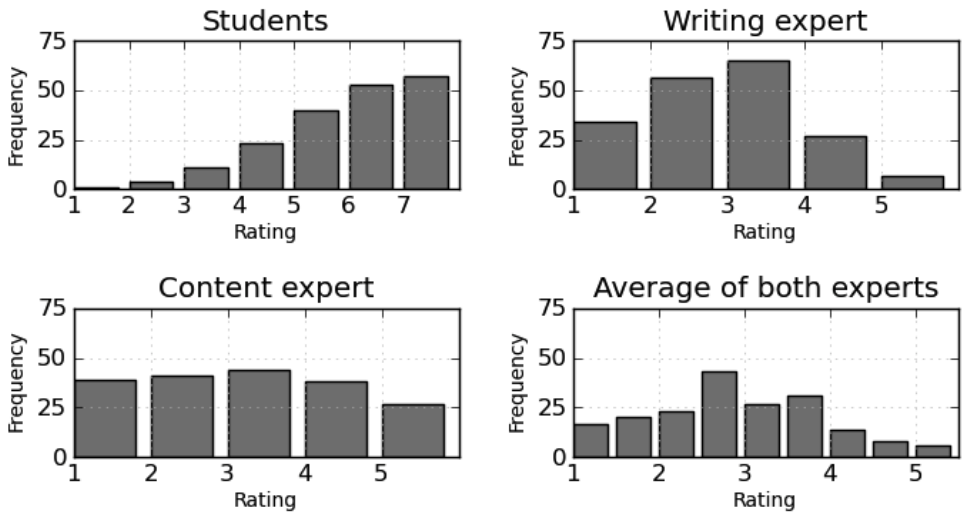

Figure 1: Distribution of peer-review helpfulness when rated by students and experts

Tag	Meaning	Word list
SUG	suggestion	should, must, might, could, need, needs, maybe, try, revision, want
LOC	location	page, paragraph, sentence
ERR	problem	error, mistakes, typo, problem, difficulties, conclusion
IDE	idea verb	consider, mention
LNK	transition	however, but
NEG	negative words	fail, hard, difficult, bad, short, little, bit, poor, few, unclear, only, more
POS	positive words	great, good, well, clearly, easily, effective, effectively, helpful, very
SUM	summarization	main, overall, also, how, job
NOT	negation	not, doesn't, don't
SOL	solution	revision specify correction

Table 1: Ten lexical categories

Compared with commonly used lexical unigrams and bigrams (Kim et al., 2006), these lexical categories are equally useful in modeling peer-review helpfulness, and significantly reduce the feature space.[4]

- **Syntactic features** mainly focus on nouns and verbs, and include percentage of tokens that are nouns, verbs, verbs conjugated in the first person (*1stPVerb%*), adjectives/adverbs, and open classes, respectively.

- **Semantic features** capture two important peer-review properties: their relevance to the main topics in students' papers, and their opinion sentiment polarities. Kim et al. (2006) extracted product property keywords from external resources based on their hypothesis that helpful product reviews refer frequently to certain product properties. Similarly, we hypothesize that helpful peer reviews are closely related to domain topics that are shared by all students papers in an assignment. Our domain topic set contains 288 words extracted from the collection of student papers using topic-lexicon extraction software[5]; our feature (*domainWord*)

[4]Lexical categories help avoid the risk of over-fitting, given only 189 peer reviews in our case compared to more than ten thousand Amazon.com reviews used for predicting product review helpfulness (Kim et al., 2006).

[5]The software extracts topic words based on topic signatures (Lin and Hovy, 2000), and was kindly provided by Annie Louis.

Feature	Description
regTag%	The percentage of problems in reviews that could be matched with a localization pattern.
soDomain%	The percentage of sentences where any domain word appears between the subject and the object.
dDeterminer	The number of demonstrative determiners.
windowSize	For each review sentence, we search for the most likely referred window of words in the related paper, and windowSize is the average number of words of all windows.

Table 2: Localization features

counts how many words of a given review belong to the extracted set. For sentiment polarities, we extract positive and negative sentiment words from the General Inquirer Dictionaries [6], and count their appearance in reviews in terms of their sentiment polarity (*posWord, negWord*).

- **Localization features** are motivated by linguistic features that are used for automatically predicting problem localization (an important cognitive construct for feedback understanding and implementation) (Nelson and Schunn, 2009), and are presented in Table 2. To illustrate how these features are computed, consider the following critique:

  > *The section of the essay on African Americans needs more careful attention to the timing and reasons for the federal governments decision to stop protecting African American civil and political rights.*

  The review has only one sentence, in which one regular expression is matched with "the section of" thus $regTag\% = 1$; no demonstrative determiner, thus $dDeterminer = 0$; "African" and "Americans" are domain words appearing between the subject "section" and the object "attention", so *soDomain* is true for this sentence and thus $soDomain\% = 1$ for the given review.

In addition to the low-level linguistic features presented above, we also examined **non-linguistic features** that are derived from the ratings and prior manual annotations of the corpus, described in Section 3.

- **Cognitive-science features** are motivated by an empirical study (Nelson and Schunn, 2009) which suggests significant correlation between certain cognitive constructs (e.g. *feedbackType*, *problem localization*, *solution*) and review implementation likelihood. Intuitively, helpful reviews are more likely to get implemented, thus we introduced these features to capture desirable high-level characteristics of peer reviews. Note that in our corpus these cognitive constructs are manually coded at the idea-unit level (Nelson and Schunn, 2009), however, peer-review helpfulness is rated at the review level.[7] Our cognitive-science features aggregate the annotations up to the review-level by reporting the percentage of idea-units in a review that exhibit each characteristic: the distribution of review types (*praise%*, *problem%*, *summary%*), the percentage of problem-localized critiques (*localization%*), as well as the percentage of solution-provided ones (*solution%*).

- **Social-science features** introduce elements reflecting interactions between students in a peer-review assignment. As suggested in related work on product review helpfulness (Kim et al., 2006; Danescu-Niculescu-Mizil et al., 2009), some social dimensions (e.g. customer opinion on related product quality) are of great influence in the perceived helpfulness of product reviews. Similarly, in our case, we introduced related paper ratings (*pRating*) — to consider whether and how helpfulness ratings are affected by the rating that the paper receives[8] — and the absolute difference between the rat-

[6] http://www.wjh.harvard.edu/ inquirer/homecat.htm

[7] Details of different granularity levels of annotation can be found in (Nelson and Schunn, 2009).

[8] That is, to examine whether students give higher ratings to peers who gave them higher paper ratings in the first place.

ing and the average score given by all reviewers (*pRatingDiff*) — to measure the variation in perceived helpfulness of a given review.

5 Experiments

We take a machine learning approach to model different types of perceived helpfulness (student helpfulness, writing-expert helpfulness, content-expert helpfulness, average-expert helpfulness) based on combinations of linguistic and non-linguistic features extracted from our peer-review corpus. Then we compare the different helpfulness types in terms of the predictive power of features used in their corresponding models. For comparison purpose, we consider the linguistic and non-linguistic features both separately and in combination, which generates three set of features: 1) linguistic features, 2) non-linguistic features, and 3) all features. For each set of features, we train four models, each corresponding to a different kind of helpfulness rating. For each learning task (three by four), we use two standard feature selection algorithms to find the most useful features based on 10-fold cross validation. First, we perform Linear Regression with Greedy Stepwise search (stepwise LR) to select the most useful features when testing in each of the ten folds, and count how many times each features is selected in the ten trials. Second, we use Relief Feature Evaluation[9] with Ranker (Relief) (Kira and Rendell, 1992; Witten and Frank, 2005) to rank all used features based on their average merits (the ability of the given feature to differentiate between two example pairs) of ten trials.[10]

Although both methods are supervised, the wrapper is "more aggressive" because its feature evaluation is based on the performance of the regression model and thus the resulting feature set is tailored to the learning algorithm. In contrast, Relief does not optimize feature sets directly for classifier performance, thus it takes into account class information in a "less aggressive" manner than the Wrapper method. We use both methods in our experiment to provide complementary perspectives. While the former can directly tell us what features are most useful, the latter gives feature ranks which provide more detailed information about differences between features. To compare the feature selection results, we examine the four kind of helpfulness models for each of the three feature sets separately, as presented below. Note that the focus of this paper is comparing feature utilities in different helpfulness models rather than predicting those types of helpfulness ratings. (Details of how the average-expert model performs can be found in our prior work (Xiong and Litman, 2011).)

5.1 Feature Selection of Linguistic Features

Table 4 presents the feature selection results of computational linguistic features used in modeling the four different types of peer-review helpfulness. The first row lists the four sources of helpfulness ratings, and each column represents a corresponding model. The second row presents the most useful features in each model selected by stepwise LR, where "# of folds" refers to the number of trials in which the given feature appears in the resulting feature set during the 10-fold cross validation. Here we only report features that are selected by no less than five folds (half the time). The third row presents feature ranks computed using Relief, where we only report the top six features due to the space limit. Features are ordered in descending ranks, and the average merit and its standard deviation is reported for each one of the features.

The selection result of stepwise LR shows that reviewLength is most useful for predicting expert helpfulness in general, while specific lexicon categories (i.e. *LNK*, and *NOT*) and positive words (*posWord*) are more useful in predicting student helpfulness. When looking at the ranking result, we observe that transition cues (*LNK*) and *posWord* are also ranked high in the student-helpfulness model, although *question%* and suggestion words (*SUG*) are ranked highest. For expert-helpfulness models, *windowSize* and *posWord*, which are not listed in the selected features for expert helpfulness (although they are selected for students), are actually ranked high for modeling average-expert helpfulness. While exclamatory sentence number (*exclams*) and summarization cues are ranked top for the writing expert,

[9] Relief evaluates the worth of an attribute by repeatedly sampling an instance and changing the value of the given attribute based on the nearest instance of the same and different class.

[10] Both algorithms are provided by Weka (http://www.cs.waikato.ac.nz/ml/weka/).

Type	Features
Structural	reviewLength, sentNum, sentLengthAve, question%, exclams
Lexical	SUG, LOC, ERR, IDE, LNK, NEG, POS, SUM, NOT, SOL (Table 1)
Syntactic	noun%, verb%, 1stPVerb%, adj+adv%, opClass%
Semantic	domainWord, posWord, negWord
Localization	regTag%, soDomain%, dDeterminer, windowSize (Table 2)
Cognitive-science	praise%, problem%, summary%, localization%, solution%
Social-science	pRating, pRatingDiff

Table 3: Summary of features

Source	Students		Writing expert		Content expert		Expert average	
Stepwise LR	Feature	# of folds	Feature	# of folds	Feature	# of folds	Feature	# of folds
	LNK	9	reviewLength	8	reviewLength	10	reviewLength	10
	posWord	8			question%	6	sentNum	8
	NOT	6			sentNum	5	question%	8
	windowSize	6			1stPVerb%	5		
					POS	5		
Relief	Feature	Merit	Feature	Merit	Feature	Merit	Feature	Merit
	question%	$.019 \pm .002$	exclams	$.010 \pm .003$	question%	$.010 \pm .004$	exclams	$.010 \pm .003$
	SUG	$.015 \pm .003$	SUM	$.008 \pm .004$	ERR	$.009 \pm .003$	question%	$.011 \pm .004$
	LNK	$.014 \pm .003$	NEG	$.006 \pm .004$	SUG	$.009 \pm .004$	windowSize	$.008 \pm .002$
	sentLengthAve	$.012 \pm .003$	negWord	$.005 \pm .002$	posWord	$.007 \pm .002$	posWord	$.006 \pm .002$
	POS	$.011 \pm .002$	windowSize	$.004 \pm .002$	exclams	$.006 \pm .001$	reviewLength	$.004 \pm .001$
	posWord	$.010 \pm .001$	sentNum	$.003 \pm .001$	1stPVerb%	$.007 \pm .004$	sentLengthAve	$.004 \pm .001$

Table 4: Feature selection based on linguistic features

the percentage of questions (*question%*) and error cues (*ERR*) are ranked top for the content-expert. In addition, the percentage of words that are verbs conjugated in the first person (*1stPVerb%*) is both selected and ranked high in the content-expert helpfulness model. Out of the four models, *SUG* are ranked high for predicting both students and content-expert helpfulness. These observations indicate that both students and experts value questions (*question%*) and suggestions (*SUG*) in reviews, and students particularly favor clear signs of logic flow in review arguments (*LNK*), positive words (*posWord*), as well as reference of their paper content which provides explicit context information (*windowSize*). In addition, experts in general prefer longer reviews (*reviewLength*), and the writing expert thinks clear summary signs (*SUM*) are important indicators of helpful peer reviews.

5.2 Feature Selection of non-Linguistic Features

When switching to the high-level non-linguistic features (Table 5), we find that *solution%* is always selected (in all ten trials) as a most useful feature for predicting all four kind of helpfulness, and is also ranked high for content-expert and student helpfulness. Especially for the content-expert, *solution%* has a much higher merit (0.013) compared to all the other features ($\leq$ 0.03). This agrees with our observation in section 5.1 that SUG are ranked high in both cases. *localization%* is selected as one of the most useful features in the content-expert helpfulness model, which is also ranked top in the student model (though not selected frequently by stepwise LR). For modeling the writing-expert helpfulness, praise (*praise%*) is more important than problem and summary, and the paper rating (*pRating*) loses its predictive power compared to how it works in the other models. In contrast, *pRating* is both selected and ranked high for predicting students' perceived helpfulness.

5.3 Feature Selection of All Features

When considering all features together as reported in Table 6, *pRating* is only selected in the student-helpfulness model, and still remains to be the most important feature for predicting students' perceived helpfulness. As for experts, the structural feature

Source	Students		Writing expert		Content expert		Expert average	
Stepwise LR	Feature	# of folds	Feature	# of folds	Feature	# of folds	Feature	# of folds
	pRating	10	solution%	10	localization%	10	solution%	10
	solution%	10			solution%	10	pRating	10
	problem%	9			pRating	10	localization%	9
Relief	Feature	Merit	Feature	Merit	Feature	Merit	Feature	Merit
	localization%	$.012 \pm .003$	praise%	$.008 \pm .002$	solution%	$.013 \pm .005$	problem%	$.004 \pm .002$
	pRatingDiff	$.010 \pm .002$	problem%	$.007 \pm .002$	pRating	$.003 \pm .002$	localization%	$.004 \pm .006$
	pRating	$.007 \pm .002$	summary%	$.001 \pm .004$	praise%	$.001 \pm .002$	praise%	$.003 \pm .003$
	solution%	$.006 \pm .005$	localization%	$.001 \pm .005$	localization%	$.001 \pm .004$	solution%	$.002 \pm .004$
	problem%	$.004 \pm .002$	pRating	$.004 \pm .004$	problem%	$.001 \pm .002$	pRating	$.005 \pm .003$
	summary%	$.004 \pm .003$	pRatingDiff	$.007 \pm .002$	pRating	$.002 \pm .003$	pRatingDiff	$.006 \pm .005$

Table 5: Feature selection based on non-linguistic features

Source	Students		Writing expert		Content expert		Expert average	
Stepwise LR	Feature	# of folds	Feature	# of folds	Feature	# of folds	Feature	# of folds
	pRating	10	reviewLength	10	reviewLength	10	reviewLength	10
	dDeterminer	7			problem%	8	problem%	6
	pRatingDiff	5						
	sentNum	5						
Relief	Feature	Merit	Feature	Merit	Feature	Merit	Feature	Merit
	pRating	$.030 \pm .006$	exclams	$.016 \pm .003$	solution%	$.025 \pm .003$	exclams	$.015 \pm .004$
	NOT	$.019 \pm .004$	praise%	$.015 \pm .003$	domainWord	$.012 \pm .002$	question%	$.012 \pm .004$
	pRatingDiff	$.019 \pm .005$	SUM	$.013 \pm .004$	regTag%	$.012 \pm .007$	LOC	$.007 \pm .002$
	sentNum	$.014 \pm .002$	summary%	$.008 \pm .003$	reviewLength	$.009 \pm .002$	sentNum	$.007 \pm .002$
	question%	$.014 \pm .003$	problem%	$.009 \pm .003$	question%	$.010 \pm .003$	reviewLength	$.007 \pm .001$
	NEG	$.013 \pm .002$	reviewLength	$.004 \pm .001$	sentNum	$.008 \pm .002$	praise%	$.008 \pm .004$

Table 6: Feature selection based on all features

reviewLength stands out from all other features in both the writing-expert and the content-expert models. Interestingly, it is the number of sentences (*sentNum*) rather than review length of structure features that is useful in the student-helpfulness model. And demonstrative determiners (*dDeterminer*) is also selected, which indicates that having a clear sign of comment targets is considered important from the students' perspective. When examining the model's ranking result, we find that more lexicon categories are ranked high for students compared to other kind of helpfulness. Specifically, *NOT* appears high again, suggesting clear expression of opinion is important in predicting student-helpfulness. Across four types of helpfulness, again, we observed that the writing expert tends to value praise and summary (indicated by both *SUM* and *summary%*) in reviews while the content-expert favors critiques, especially solution provided critiques.

5.4 Discussion

Based on our observations from the above three comparisons, we summarize our findings with respect to different feature types and provide interpretation: 1) review length (in tokens) is generally effective in predicting expert perceived helpfulness, while number of sentences is more useful in modeling student perceived helpfulness. Interestingly, there is a strong correlation between these two features ($r = 0.91, p \leq 0.001$), and why one is selected over the other in different helpfulness models needs further investigation. 2) Lexical categories such as transition cues, negation, and suggestion words are of more importance in modeling student perceived helpfulness. This might indicate that students prefer clear expression of problem, reference and even opinion in terms of specific lexicon clues, the lack of which is likely to result in difficulty in their understanding of the reviews. 3) As for cognitive-science features, solution is generally an effective indicator of helpful peer reviews. Within the three feedback types of peer reviews, praise is valued high by the writing expert. (It is interesting to notice that although praise is shown to be more important than problem and summary for modeling the writing-expert helpfulness, positive sentiment words do not appear to be more predictive than negative sentiments.) In contrast, problem is more desirable

from the content expert's point of view. Although students assign less importance to the problem themselves, solution provided peer reviews could be helpful for them with respect to the learning goal of peer-review assignments. 4) Paper rating is a very effective feature for predicting review helpfulness perceived by students, which is not the case for either expert. This supports the argument of social aspects in people's perception of review helpfulness, and it also reflects the fact that students tend to be nice to each other in such peer-review interactions. However, this dimension might not correspond with the real helpfulness of the reviews, at least from the perspective of both the writing expert and content expert.

6 Conclusion and Future Work

We have shown that the type of helpfulness to be predicted does indeed influence the utility of different feature types for automatic prediction. Low-level general linguistic features are more predictive when modeling students' perceived helpfulness; high-level theory supported constructs are more useful in experts' models. However, in the related area of automated essay scoring (Attali and Burstein, 2006), others have suggested the need for the use of validated features related to meaningful dimensions of writing, rather than low-level (but easy to automate) features. In this perspective, our work similarly poses challenge to the NLP community in terms of how to take into account the education-oriented dimensions of helpfulness when applying traditional NLP techniques of automatically predicating review helpfulness. In addition, it is important to note that predictive features of perceived helpfulness are not guaranteed to capture the nature of "truly" helpful peer reviews (in contrast to the perceived ones).

In the future, we would like to investigate how to integrate useful dimensions of helpfulness perceived by different audiences in order to come up with a "true" helpfulness gold standard. We would also like to explore more sophisticated features and other NLP techniques to improve our model of peer-review helpfulness. As we have already built models to automatically predict certain cognitive constructs (problem localization and solution), we will replace the annotated cognitive-science features used here with their automatic predictions, so that we can build our helpfulness model fully automatically. Finally, we would like to integrate our helpfulness model into a real peer-review system and evaluate its performance extrinsically in terms of improving students' learning and reviewing performance in future peer-review assignments.

Acknowledgments

This work was supported by the Learning Research and Development Center at the University of Pittsburgh. We thank Melissa Patchan and Chris Schunn for generously providing the manually annotated peer-review corpus. We are also grateful to Michael Lipschultz and Chris Schunn for their feedback while writing this paper.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment (JTLA)*, 4(3), February.

Kwangsu Cho and Christian D. Schunn. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education*, 48:409–426.

Kwangsu Cho. 2008. Machine classification of peer comments in physics. In *Proceedings of the First International Conference on Educational Data Mining (EDM2008)*, pages 192–196.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: A case study on Amazon.com helpfulness votes. In *Proceedings of WWW*, pages 141–150.

Raquel M. Crespo Garcia. 2010. Exploring document clustering techniques for personalized peer assessment in exploratory courses. In *Proceedings of Computer-Supported Peer Review in Education (CSPRED) Workshop in the Tenth International Conference on Intelligent Tutoring Systems (ITS 2010)*.

Anindya Ghose and Panagiotis G. Ipeirotis. 2010. Estimating the helpfunless and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 99.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference*

on Empirical Methods in Natural Language Processing (EMNLP2006), pages 423–430, Sydney, Australia, July.

Kenji Kira and Larry A. Rendell. 1992. A practical approach to feature selection. In Derek H. Sleeman and Peter Edwards, editors, *ML92: Proceedings of the Ninth International Conference on Machine Learning*, pages 249–256, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, volume 1 of *COLING '00*, pages 495–501, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yang Liu, Xiangji Guang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, pages 443–452, Los Alamitos, CA, USA.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 91–98, Stroudsburg, PA, USA. Association for Computational Linguistics.

Melissa M. Nelson and Christian D. Schunn. 2009. The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science*, 37(4):375–401.

Melissa M. Patchan, Davida Charney, and Christian D. Schunn. 2009. A validation study of students' end comments: Comparing comments by students, a writing instructor, and a content instructor. *Journal of Writing Research*, 1(2):124–152.

Agnes Sandor and Angela Vorndran. 2009. Detecting key sentences for automatic assistance in peer-reviewing research articles in educational sciences. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, pages 36–44.

Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM2009)*, pages 36–44.

IH Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, San Francisco, CA.

Wenting Xiong and Diane Litman. 2010. Identifying problem localization in peer-review feedback. In *Proceedings of Tenth International Conference on Intelligent Tutoring Systems (ITS2010)*, volume 6095, pages 429–431.

Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT)*, Portland, Oregon, June.