Automatic Summarization of Student Course Feedback

Wencan Luo† **Fei Liu**‡ **Zitao Liu**† **Diane Litman**†
†University of Pittsburgh, Pittsburgh, PA 15260
‡University of Central Florida, Orlando, FL 32716
{wencan, ztliu, litman}@cs.pitt.edu feiliu@cs.ucf.edu

Abstract

Student course feedback is generated daily in both classrooms and online course discussion forums. Traditionally, instructors manually analyze these responses in a costly manner. In this work, we propose a new approach to summarizing student course feedback based on the integer linear programming (ILP) framework. Our approach allows different student responses to share co-occurrence statistics and alleviates sparsity issues. Experimental results on a student feedback corpus show that our approach outperforms a range of baselines in terms of both ROUGE scores and human evaluation.

1 Introduction

Instructors love to solicit feedback from students. Rich information from student responses can reveal complex teaching problems, help teachers adjust their teaching strategies, and create more effective teaching and learning experiences. Text-based student feedback is often manually analyzed by teaching evaluation centers in a costly manner. Albeit useful, the approach does not scale well. It is therefore desirable to automatically summarize the student feedback produced in online and offline environments. In this work, student responses are collected from an introductory materials science and engineering course, taught in a classroom setting. Students are presented with prompts after each lecture and asked to provide feedback. These prompts solicit "*reflective feedback*" (Boud et al., 2013) from the students. An example is presented in Table 1.

Prompt
Describe what you found most interesting in today's class
Student Responses
S1: The main topics of this course seem interesting and correspond with my major (Chemical engineering)
S2: I found the group activity most interesting
S3: Process that make materials
S4: I found the properties of bike elements to be most interesting
S5: How materials are manufactured
S6: Finding out what we will learn in this class was interesting to me
S7: The activity with the bicycle parts
S8: "part of a bike" activity
... (*rest omitted, 53 responses in total.*)
Reference Summary
- group activity of analyzing bicycle's parts
- materials processing
- the main topic of this course

Table 1: Example student responses and a reference summary created by the teaching assistant. 'S1'–'S8' are student IDs.

In this work, we aim to summarize the student responses. This is formulated as an extractive summarization task, where a set of representative sentences are extracted from student responses to form a textual summary. One of the challenges of summarizing student feedback is its lexical variety. For example, in Table 1, "bike elements" (S4) and "bicycle parts" (S7), "the main topics of this course" (S1) and "what we will learn in this class" (S6) are different expressions that communicate the same or similar meanings. In fact, we observe 97% of the bigrams appear only once or twice in the student feedback corpus (§4), whereas in a typical news dataset (DUC 2004), it is about 80%. To tackle this challenge, we propose a new approach to summarizing student feedback, which extends the standard ILP framework by approximating the co-occurrence matrix using a low-rank alternative. The resulting system allows sentences authored by different students to share co-occurrence statistics. For example, "The activity with the bicycle parts" (S7) will be allowed to partially contain "bike elements" (S4) although the latter did not appear in the sentence. Experiments show that our approach produces better results on the student feedback summarization task in terms of both ROUGE scores and human evaluation.

2 ILP Formulation

Let $\mathcal{D}$ be a set of student responses that consist of M sentences in total. Let $y_j \in \{0,1\}$, $j = \{1, \cdots, M\}$ indicate if a sentence j is selected ($y_j = 1$) or not ($y_j = 0$) in the summary. Similarly, let N be the number of unique concepts in $\mathcal{D}$. $z_i \in \{0,1\}$, $i = \{1, \cdots, N\}$ indicate the appearance of concepts in the summary. Each concept i is assigned a weight of w_i, often measured by the number of sentences or documents that contain the concept. The ILP-based summarization approach (Gillick and Favre, 2009) searches for an optimal assignment to the sentence and concept variables so that the selected summary sentences maximize coverage of important concepts. The relationship between concepts and sentences is captured by a co-occurrence matrix $A \in \mathbb{R}^{N \times M}$, where $A_{ij} = 1$ indicates the i-th concept appears in the j-th sentence, and $A_{ij} = 0$ otherwise. In the literature, bigrams are frequently used as a surrogate for concepts (Gillick et al., 2008; Berg-Kirkpatrick et al., 2011). We follow the convention and use 'concept' and 'bigram' interchangeably in the paper.

$$\max_{\boldsymbol{y},\boldsymbol{z}} \quad \textstyle\sum_{i=1}^{N} w_i z_i \tag{1}$$

$$s.t. \quad \textstyle\sum_{j=1}^{M} A_{ij}\, y_j \geq z_i \tag{2}$$

$$A_{ij}\, y_j \leq z_i \tag{3}$$

$$\textstyle\sum_{j=1}^{M} l_j y_j \leq L \tag{4}$$

$$y_j \in \{0,1\}, z_i \in \{0,1\} \tag{5}$$

Two sets of linear constraints are specified to ensure the ILP validity: (1) a concept is selected if and only if at least one sentence carrying it has been selected (Eq. 2), and (2) all concepts in a sentence will be selected if that sentence is selected (Eq. 3). Finally, the selected summary sentences are allowed to contain a total of L words or less (Eq. 4).

3 Our Approach

Because of the lexical diversity in student responses, we suspect the co-occurrence matrix A may not establish a faithful correspondence between sentences and concepts. A concept may be conveyed using multiple bigram expressions; however, the current co-occurrence matrix only captures a binary relationship between sentences and bigrams. For example, we ought to give partial credit to "bicycle parts" (S7) given that a similar expression "bike elements" (S4) appears in the sentence. Domain-specific synonyms may be captured as well. For example, the sentence "I tried to follow along but I couldn't *grasp the* concepts" is expected to partially contain the concept "understand the", although the latter did not appear in the sentence.

The existing matrix A is highly sparse. Only 2.7% of the entries are non-zero in our dataset (§4). We therefore propose to *impute* the co-occurrence matrix by filling in missing values. This is accomplished by approximating the original co-occurrence matrix using a low-rank matrix. The low-rankness encourages similar concepts to be shared across sentences. The data imputation process makes two notable changes to the existing ILP framework. First, it extends the domain of A_{ij} from binary to a continuous scale $[0,1]$ (Eq. 2), which offers a better sentence-level semantic representation. The binary concept variables (z_i) are also relaxed to continuous domain $[0,1]$ (Eq. 5), which allows the concepts to be "partially" included in the summary.

Concretely, given the co-occurrence matrix $A \in \mathbb{R}^{N \times M}$, we aim to find a low-rank matrix $B \in \mathbb{R}^{N \times M}$ whose values are close to A at the observed positions. Our objective function is

$$\min_{B \in \mathbb{R}^{N \times M}} \frac{1}{2} \sum_{(i,j) \in \Omega} (A_{ij} - B_{ij})^2 + \lambda \|B\|_*, \tag{6}$$

where Ω represents the set of observed value positions. $\|B\|_*$ denotes the trace norm of B, i.e., $\|B\|_* = \sum_{i=1}^{r} \sigma_i$, where r is the rank of B and σ_i are the singular values. By defining the following projection operator P_Ω,

$$[P_\Omega(B)]_{ij} = \begin{cases} B_{ij} & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega \end{cases} \quad (7)$$

our objective function (Eq. 6) can be succinctly represented as

$$\min_{B \in \mathbb{R}^{N\times M}} \frac{1}{2}\|P_\Omega(A) - P_\Omega(B)\|_F^2 + \lambda\|B\|_*, \quad (8)$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Following (Mazumder et al., 2010), we optimize Eq. 8 using the proximal gradient descent algorithm. The update rule is

$$B^{(k+1)} = \text{prox}_{\lambda\rho_k}\Big(B^{(k)} + \rho_k\big(P_\Omega(A) - P_\Omega(B)\big)\Big), \quad (9)$$

where ρ_k is the step size at iteration k and the proximal function $\text{prox}_t(B)$ is defined as the singular value soft-thresholding operator, $\text{prox}_t(B) = U \cdot \text{diag}((\sigma_i - t)_+) \cdot V^\top$, where $B = U\text{diag}(\sigma_1, \cdots, \sigma_r)V^\top$ is the singular value decomposition (SVD) of B and $(x)_+ = \max(x, 0)$.

Since the gradient of $\frac{1}{2}\|P_\Omega(A) - P_\Omega(B)\|_F^2$ is Lipschitz continuous with $L = 1$ (L is the Lipschitz continuous constant), we follow (Mazumder et al., 2010) to choose fixed step size $\rho_k = 1$, which has a provable convergence rate of $O(1/k)$, where k is the number of iterations.

4 Dataset

Our dataset is collected from an introductory materials science and engineering class taught in a major U.S. university. The class has 25 lectures and enrolled 53 undergrad students. The students are asked to provide feedback after each lecture based on three prompts: 1) "describe what you found most interesting in today's class," 2) "describe what was confusing or needed more detail," and 3) "describe what you learned about how you learn." These open-ended prompts are carefully designed to encourage students to self-reflect, allowing them to "recapture experience, think about it and evaluate it" (Boud et al., 2013). The average response length is 10±8.3 words. If we concatenate all the responses to each lecture and prompt into a "pseudo-document", the document contains 378 words on average.

The reference summaries are created by a teaching assistant. She is allowed to create abstract summaries using her own words in addition to selecting phrases directly from the responses. Because summary annotation is costly and recruiting annotators with proper background is nontrivial, 12 out of the 25 lectures are annotated with reference summaries. There is one gold-standard summary per lecture and question prompt, yielding 36 document-summary pairs[1]. On average, a reference summary contains 30 words, corresponding to 7.9% of the total words in student responses. 43.5% of the bigrams in human summaries appear in the responses.

5 Experiments

Our proposed approach is compared against a range of baselines. They are 1) MEAD (Radev et al., 2004), a centroid-based summarization system that scores sentences based on length, centroid, and position; 2) LEXRANK (Erkan and Radev, 2004), a graph-based summarization approach based on eigenvector centrality; 3) SUMBASIC (Vanderwende et al., 2007), an approach that assumes words occurring frequently in a document cluster have a higher chance of being included in the summary; 4) BASELINE-ILP (Berg-Kirkpatrick et al., 2011), a baseline ILP framework without data imputation.

For the ILP based approaches, we use bigrams as concepts (bigrams consisting of only stopwords are removed[2]) and sentence frequency as concept weights. We use all the sentences in 25 lectures to construct the concept-sentence co-occurrence matrix and perform data imputation. It allows us to leverage the co-occurrence statistics both within and across lectures. For the soft-impute algorithm, we perform grid search (on a scale of [0, 5] with stepsize 0.5) to tune the hyper-parameter λ. To make the most use of annotated lectures, we split them into three folds. In each one, we tune λ on two folds and test it on the other fold. Finally, we report the averaged results. In all experiments, summary length is set to be 30 words or less, corresponding to the

[1]This data set is publicly available at `http://www.coursemirror.com/download/dataset`.

[2]Bigrams with one stopword are not removed because 1) they are informative ("a bike", "the activity", "how materials'); 2) such bigrams appear in multiple sentences and are thus helpful for matrix imputation.

System	ROUGE-1 R (%)	P (%)	F (%)	ROUGE-2 R (%)	P (%)	F (%)	ROUGE-SU4 R (%)	P (%)	F (%)	Human Preference
MEAD	26.4	23.3	21.8	6.7	7.6	6.3	8.8	8.0	5.4	24.8%
LEXRANK	30.0	27.6	25.7	8.1	8.3	7.6	9.6	9.6	6.6	—
SUMBASIC	36.6	31.4	30.4	8.2	8.1	7.5	13.9	11.0	8.7	—
ILP BASELINE	35.5	31.8	29.8	11.1	10.7	9.9	12.9	11.5	8.2	69.6%
OUR APPROACH	**38.0**	**34.6**	**32.2**	**12.7**	**12.9**	**11.4**	**15.5**	**14.4**	**10.1**	**89.6%**

Table 2: Summarization results evaluated by ROUGE (%) and human judges. Shaded area indicates that the performance difference with OUR APPROACH is statistically significant ($p < 0.05$) using a two-tailed paired t-test on the 36 document-summary pairs.

average number of words in human summaries.

In Table 2, we present summarization results evaluated by ROUGE (Lin, 2004) and human judges. ROUGE is a standard evaluation metric that compares system and reference summaries based on n-gram overlaps. Our proposed approach outperforms all the baselines based on three standard ROUGE metrics.[3] When examining the imputed sentence-concept co-occurrence matrix, we notice some interesting examples that indicate the effectiveness of the proposed approach, shown in Table 3.

Because ROUGE cannot thoroughly capture the semantic similarity between system and reference summaries, we further perform human evaluation. For each lecture and prompt, we present the prompt, a pair of system outputs in a random order, and the human summary to five Amazon turkers. The turkers are asked to indicate their preference for system A or B based on the semantic resemblance to the human summary on a 5-Likert scale ('Strongly preferred A', 'Slightly preferred A', 'No preference', 'Slightly preferred B', 'Strongly preferred B'). They are rewarded \$0.08 per task. We use two strategies to control the quality of the human evaluation. First, we require the turkers to have a Human Intelligence Task (HIT) approval rate of 90% or above. Second, we insert some quality checkpoints by asking the turkers to compare two summaries of same text content but different sentence orders. Turkers who did not pass these tests are filtered out. Due to budget constraints, we conduct pairwise comparisons for three systems. The total number of comparisons is 3 system-system pairs $\times$ 12 lectures $\times$ 3 prompts $\times$ 5 turkers = 540 total pairs. We calculate the percentage of "wins" (strong or slight preference) for each system among all comparisons with its counterparts. Results are reported in the last column of Table 2. OUR APPROACH is preferred significantly more often than the other two systems[4]. Regarding the inter-annotator agreement, we find 74.3% of the individual judgements agree with the majority votes when using a 3-point Likert scale ('preferred A', 'no preference', 'preferred B').

Table 4 presents example system outputs. This offers intuitive understanding to our proposed approach.

Sentence	Assoc. Bigrams
the printing needs to better so it can be easier to read	*the graph*
graphs make it *easier to* understand concepts	*hard to*
the naming system for the 2 *phase regions*	*phase diagram*
I tried to follow along but I couldn't *grasp the* concepts	*understand the*
no problems except for the specific equations used to determine properties from the stress - *strain graph*	*strain curves*

Table 3: Associated bigrams do not appear in the sentence, but after Matrix Imputation, they yield a decent correlation (cell value greater than 0.9) with the corresponding sentence.

[3] F-scores are slightly lower than P/R because of the averaging effect and can be illustrated in one example. Suppose we have P1=0.1, R1=0.4, F1=0.16 and P2=0.4, R2=0.1, F2=0.16. Then the macro-averaged P/R/F-scores are: P=0.25, R=0.25, F=0.16. In this case, the F-score is lower than both P and R.

[4] For the significance test, we convert a preference to a score ranging from -2 to 2 ('2' means 'Strongly preferred' to a system and '-2' means 'Strongly preferred' to the counterpart system), and use a two-tailed paired t-test with $p < 0.05$ to compare the scores.

Prompt *Describe what you found most interesting in today's class*
Reference Summary - unit cell direction drawing and indexing - real world examples - importance of cell direction on materials properties
System Summary (ILP BASELINE) - drawing and indexing unit cell direction - it was interesting to understand how to find apf and fd from last weeks class - south pole explorers died due to properties of tin
System Summary (OUR APPROACH) - crystal structure directions - surprisingly i found nothing interesting today . - unit cell indexing - vectors in unit cells - unit cell drawing and indexing - the importance of cell direction on material properties

Table 4: Example reference and system summaries.

6 Related Work

Our previous work (Luo and Litman, 2015) proposes to summarize student responses by extracting phrases rather than sentences in order to meet the need of aggregating and displaying student responses in a mobile application (Luo et al., 2015; Fan et al., 2015). It adopts a clustering paradigm to address the lexical variety issue. In this work, we leverage matrix imputation to solve this problem and summarize student response at a sentence level.

The integer linear programming framework has demonstrated substantial success on summarizing news documents (Gillick et al., 2008; Gillick et al., 2009; Woodsend and Lapata, 2012; Li et al., 2013). Previous studies try to improve this line of work by generating better estimates of concept weights. Galanis et al. (2012) proposed a support vector regression model to estimate bigram frequency in the summary. Berg-Kirkpatrick et al. (2011) explored a supervised approach to learn parameters using a cost-augmentative SVM. Different from the above approaches, we focus on the co-occurrence matrix instead of concept weights, which is another important component of the ILP framework.

Most summarization work focuses on summarizing news documents, as driven by the DUC/TAC conferences. Notable systems include maximal marginal relevance (Carbonell and Goldstein, 1998), submodular functions (Lin and Bilmes, 2010), jointly extract and compress sentences (Zajic et al., 2007), optimize content selection and surface realization (Woodsend and Lapata, 2012), minimize reconstruction error (He et al., 2012), and dual decomposition (Almeida and Martins, 2013). Albeit the encouraging performance of our proposed approach on summarizing student responses, when applied to the DUC 2004 dataset (Hong et al., 2014) and evaluated using ROUGE we observe only comparable or marginal improvement over the ILP baseline. However, this is not surprising since the lexical variety is low (20% of bigrams appear more than twice compared to 3% of bigrams appear more than twice in student responses) and thus less data sparsity, so the DUC data cannot benefit much from imputation.

7 Conclusion

We make the first effort to summarize student feedback using an integer linear programming framework with data imputation. Our approach allows sentences to share co-occurrence statistics and alleviates sparsity issue. Our experiments show that the proposed approach performs competitively against a range of baselines and shows promise for future automation of student feedback analysis.

In the future, we may take advantage of the high quality student responses (Luo and Litman, 2016) and explore helpfulness-guided summarization (Xiong and Litman, 2014) to improve the summarization performance. We will also investigate whether the proposed approach benefits other informal text such as product reviews, social media discussions or spontaneous speech conversations, in which we expect the same sparsity issue occurs and the language expression is diverse.

Acknowledgments

This research is supported by an internal grant from the Learning Research and Development Center at the University of Pittsburgh. We thank Muhsin Menekse for providing the data set. We thank Jingtao Wang, Fan Zhang, Huy Nguyen and Zahra Rahimi for valuable suggestions about the proposed summarization algorithm. We also thank anonymous reviewers for insightful comments and suggestions.

References

Miguel B. Almeida and Andre F. T. Martins. 2013. Fast and robust compressive summarization with dual decomposition and multi-task learning. In *Proceedings of ACL*.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of ACL*.

David Boud, Rosemary Keogh, David Walker, et al. 2013. *Reflection: Turning experience into learning*. Routledge.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal Artificial Intelligence Research*, 22(1).

Xiangmin Fan, Wencan Luo, Muhsin Menekse, Diane Litman, and Jingtao Wang. 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.

Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of COLING*.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of NAACL*.

Dan Gillick, Benoit Favre, and Dilek Hakkani-Tur. 2008. The icsi summarization system at tac 2008. In *Proceedings of TAC*.

Dan Gillick, Benoit Favre, Dilek Hakkani-Tur, Berndt Bohnet, Yang Liu, and Shasha Xie. 2009. The ICSI/UTD summarization system at TAC 2009. In *Proceedings of TAC*.

Zhanying He, Chun Chen, Jiajun Bu, Can Wang, Lijun Zhang, Deng Cai, and Xiaofei He. 2012. Document summarization based on data reconstruction. In *Proceedings of AAAI*.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1070.

Chen Li, Xian Qian, and Yang Liu. 2013. Using supervised bigram-based ILP for extractive summarization. In *Proceedings of ACL*.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Proceedings of NAACL*.

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

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

Wencan Luo and Diane Litman. 2016. Determining the quality of a student reflective response. In *Proceedings 29th International FLAIRS Conference*, Key Largo, FL, May.

Wencan Luo, Xiangmin Fan, Muhsin Menekse, Jingtao Wang, , and Diane Litman. 2015. Enhancing instructor-student and student-student interactions with mobile interfaces and summarization. In *Proceedings of NAACL (Demo)*.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*.

Dragomir R. Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Management*, 43(6):1606–1618.

Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of EMNLP*.

Wenting Xiong and Diane Litman. 2014. Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1985–1995, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

David Zajic, Bonnie J. Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management*.