Summarizing Student Responses to Reflection Prompts

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

We propose to automatically summarize student responses to reflection prompts and introduce a novel summarization algorithm that differs from traditional methods in several ways. First, since the linguistic units of student inputs range from single words to multiple sentences, our summaries are created from extracted phrases rather than from sentences. Second, the phrase summarization algorithm ranks the phrases by the number of students who semantically mention a phrase in a summary. Experimental results show that the proposed phrase summarization approach achieves significantly better summarization performance on an engineering course corpus in terms of ROUGE scores when compared to other summarization methods, including MEAD, LexRank and MMR.

1 Introduction

Educational research has demonstrated the effectiveness of *reflection prompts* (Boud et al., 2013) to enhance interaction between instructors and students (Van den Boom et al., 2004; Menekse et al., 2011). However, summarizing student responses to these prompts for large courses (e.g., introductory STEM, MOOCs) is an onerous task for humans and poses challenges for existing summarization methods. First, the linguistic units of student inputs range from single words to multiple sentences. Second, we assume that the concepts (represented as phrases) mentioned by more students should get more attention from the instructor. Based on this assumption, we introduce the notion of *student coverage*, defined as the number of students who semantically mention a particular phrase. The more student coverage a phrase has, the more important it is. To illustrate the new task, an example is shown in Table. 1.

Reflection prompt:
Describe what was confusing or needed more detail.

Student Responses:
S1: Graphs of attraction/repulsive & interatomic separation
S2: Property related to bond strength
S3: The activity was difficult to comprehend as the text fuzzing and difficult to read.
S4: Equations with bond strength and Hooke's law
S5: I didn't fully understand the concept of thermal expansion
S6: The activity (Part III)
S7: Energy vs. distance between atoms graph and what it tells us
S8: The graphs of attraction and repulsion were confusing to me...(*rest omitted, 53 student responses in total.*)

Human Reference Summary:
1) Graphs of attraction/ repulsive & atomic separation [10]
2) Properties and equations with bond strength [7]
3) Coefficient of thermal expansion [6]
4) Activity part III [4]

Table 1: An example reflection prompt, 53 student responses and a gold-standard summary. The numbers in the square brackets indicate the number of students who semantically mention each phrase (i.e., student coverage).

In this work, we propose a phrase summarization method that addresses the above challenges. First, our summaries are created from extracted phrases rather than from sentences. Phrases are easy to read and browse like keywords, and fit better on small devices when compared to sentences. For example, including phrases such as "*I didn't fully understand*" (S5) and "*were confusing to me*" (S8) in the summary is a waste of space. Second, we adopt a metric clustering paradigm with a semantic distance to estimate the student coverage of each phrase in the summary; a semantic metric allows similar phrases to be grouped together even if they are in different textual forms. Experimental

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1955–1960,
Lisbon, Portugal, 17-21 September 2015.

results demonstrate the utility of our approach.

Although not the focus of this paper, we have also built a mobile application called CourseMIR-ROR[1] that utilizes the proposed summarization algorithm (Luo et al., 2015). Fan et al. (2015) report a preliminary study about the usage of the application.

2 Related Work

While summarization systems that extract sentences are dominant, others have published in "summarization" at other levels besides the sentence. For example, Ueda et al. (2000) developed an "at-a-glance" summarization method with handcrafted rules. Recently, keyphrase extraction (Hasan and Ng, 2014; Liu et al., 2009; Medelyan et al., 2009; Wu et al., 2005) has received considerable attention, aiming to select important phrases from input documents, which is similar to phrase summarization. In this paper, we propose a general framework to adapt sentence summarization to phrase summarization.

Clustering has been used to score sentences and has shown good improvement in text summarization (Yang et al., 2012; Li and Li, 2014; Gung and Kalita, 2012). In this work, we are using a metric clustering with semantic similarity to estimate the student coverage at a phrase level. Similarly, both diversity-based summarization (Carbonell and Goldstein, 1998; Zhang et al., 2005; Zhu et al., 2007) and our proposed method aim to estimate and maximize student coverage by minimizing redundancy in the output phrases. Differently, our method performs the redundancy reduction at a cluster level (a group of phrases) rather than penalize redundancy with a greedy iterative procedure sentence by sentence, and not only the information content is considered, but also the information source.

3 Data

Our data consists of student responses collected from 53 undergraduates enrolled in an introduction to materials science and engineering class. The students were asked to complete a survey at the end of each of 25 lectures during a semester, consisting of three carefully designed reflection prompts: 1) "Describe what you found most interesting in today's class." 2) "Describe what was confusing or needed more detail." 3) "Describe what you learned about how you learn."

	min	max	mean	std
Student-WC	1	91	9.2	7.3
TA-PWC	1	26	7.1	4.9
TA-WC	6	103	29.4	23.2
TA-PC	2	12	4.2	2.2

Table 2: **W**ord **C**ount (WC) in student responses (Student-WC), WC per phrase in TA's summary (TA-PWC), WC in TA's summary (TA-WC) and phrase count in TA's summary (TA-PC)

In total, more than 900 responses were collected for each prompt. Currently, gold-standard summaries of 12 out of 25 lectures are created by the teaching assistant for that course for each reflection prompt. The summaries include not only the important phrases, but also the number of students who mentioned them (i.e., student coverage). 4 lectures are randomly selected as a development set and the remaining data used as a test set, yielding 12 sets of development data and 24 sets of testing data, each with a prompt, the students' responses and the gold-standard summary. [2]

The statistics of the student responses and the TA's summary are shown in Table 2. The phrases summarized by the TA are significantly shorter than the student responses (7.1 vs. 9.2, $p<0.001$).

4 Proposed Method

We formulate our task as a standard extractive summarization problem. Unlike standard sentence-level extraction where the input and output are sentences, the input of our task ranges from words or phrases to full sentences. The output is a list of important phrases and the summary length (either # of phrases or words) is no more than L.

The proposed algorithm involves three stages: *candidate phrase extraction*, *phrase clustering*, and *phrase ranking*.

4.1 Candidate phrase extraction

We extract noun phrases (NPs) from the input using a syntax parser from the Senna toolkit (Collobert, 2011), preserving the most important con-

[1]Homepage: http://www.coursemirror.com; free download link in Google Play Store: https://play.google.com/store/apps/details?id=edu.pitt.cs.mips.coursemirror

[2]This data is publicly available at the CourseMIR-ROR website: http://www.coursemirror.com/download/dataset.

tent from the original responses without losing too much context information compared to keywords. For example, "the concept of thermal expansion" (S5) is extracted as a candidate phrase. Only NPs are considered because all reflection prompts used in the task are asking about "what", and knowledge concepts are usually represented as NPs.[3]

Due to the noisy data, malformed phrases are excluded, including single stop words (e.g. "it", "I", "there", "nothing") and phrases starting with a punctuation mark (e.g. "'t", "+ indexing").

4.2 Phrase clustering

Although phrases are more meaningful and less ambiguous compared to keywords, they suffer from the sparsity problem, especially in our data set when 89.9% of the phrases appeared only once. The challenge is the fact that students use different words for the same meaning (e.g., "bicycle parts" and "bike elements").

We use a clustering paradigm with a semantic distance metric to address this issue. Among different clustering algorithms, K-Medoids (Kaufman and Rousseeuw, 1987) fits well for our problem. First, it works with an arbitrary distance matrix between datapoints. It allows to use pairwise semantic similarity-based distance between phrases, yielding metric clustering. Second, it is robust to noise and outliers because it minimizes a sum of pairwise dissimilarities instead of squared Euclidean distances. It shows better performance than an LDA-based approach to group students' short answers for the purpose of semi-automated grading (Basu et al., 2013). Since K-Medoids picks a random set of seeds to initialize as the cluster centers (called medoids), the clustering algorithm runs 100 times and the cluster with the minimal within-cluster sum of the distances is retained to reduce random effects.

Distance metric. The semantic similarity is implemented using SEMILAR (Rus et al., 2013), using the *latent semantic analysis* trained on the Touchstone Applied Science Associates corpus (Ştefănescu et al., 2014). The distance matrix D is constructed from the similarity matrix S by applying the following transformation: $D = e^{-S}$, which is similar to the common heat kernel but without normalization[4].

[3]In our data, no advantage is observed by including other constituents like verb and prepositional phrases.

[4]This is not normalized to the range between 0 and 1 since we only care about the relative distance.

Number of clusters. For setting the number of clusters without tuning, we adopted a method from Wan and Yang (2008), by letting $K = \sqrt{V}$. where K is the number of clusters and V is the number of candidate phrases instead of the number of sentences.

4.3 Phrase ranking

In order to estimate the *student coverage*, phrases are clustered with the algorithm introduced above. We assume the phrases in a cluster are semantically similar to each other and any phrase in a cluster can represent it as a whole. Therefore the coverage of a phrase is assumed to be the same as the coverage of a cluster, which is a union of the students covered by each phrase in the cluster.

To select the most representative phrase in a cluster, LexRank (Erkan and Radev, 2004), a graph-based algorithm for computing relative importance of textual units (working for both sentences and phrases), is used to score the extracted candidate phrases. The top ranked phrase in the cluster is added to the output summary. This process starts from the cluster that has the most estimated student coverage and repeats for the next cluster until the length limit is reached.

Note that when the student coverage is the same between two clusters, the score of the top-ranked phrases in the clusters according to LexRank is used to break the tie: the higher, the better.

5 Experiments

We use the ROUGE evaluation metric (Lin, 2004) and report R-1 (unigrams), R-2 (bigrams), and R-SU4 (bigrams with skip distance up to 4 words), including the recall (R), precision (P) and F-Measure (F). These scores measure the overlap between human-generated summaries and a machine-generated summary.

We design and compare a number of other summarization methods to evaluate the proposed phrase summarization approach.

Keyphrase extraction. Maui (Medelyan et al., 2009) is selected as the baseline, which is one of the state-of-the-art keyphrase extraction methods.

Sentence to phrase summarization. Existing sentence summarization techniques can be used for phrase summarization by extracting candidate phrases and treating them as sentences. Within this framework, we adapt MEAD (Radev et al., 2004) and LexRank (Erkan and Radev, 2004) to

	R-1			R-2			R-SU4		
	R	P	F	R	P	F	R	P	F
Keyphrase	.171	.364	.211	.057	.134	.071	.039	.168	.049
OriMEAD	**.397**	.185	.219	.117	.069	.073	**.157**	.051	.045
MEAD	.341	.269	.265	.122	.102	.099	.126	.094	.072
MEAD+MMR	.360	.279	.277	**.130**	.106	.104	.142	.099	.078
LexRank	.325	.355	.307	.107	.110	.102	.120	.145	.098
LexRank+MMR	.328	.367	.312	.111	.126	.110	.117	.154	.098
Clustering+Medoid	.279	**.473**	.327	.078	.129	.091	.068	**.216**	.087
Proposed	.319	$.448^{\dagger *}$	$\mathbf{.340}^{\dagger}$	.122	$\mathbf{.176}^{\dagger *}$	**.134**	.112	$.205^{\dagger *}$	$\mathbf{.109}^{\dagger}$

Table 3: Summarization performance. The last row is our proposed approach. The highest score for each column is shown in bold. † indicates that the improvement over the MEAD+MMR baseline is statistically significant. * indicates that the improvement over LexRank+MMR is statistically significant.

our task. We also include the original MEAD[5] for comparison (named as OriMEAD).

Diversity-based summarization. We applied the MMR (Carbonell and Goldstein, 1998), a popular diversity-based summarization method as a post-processing step to the MEAD (**MEAD+MMR**) and LexRank (**LexRank+MMR**) baselines.[6]

Clustering+Medoid. To show the performance using the clustering alone, this baseline selects the medoid phrase instead of using LexRank to rank the phrases in a cluster to form the summary.

Results. The performance on the test set is shown in Table 3 with the length limit L as 4 phrases (the average phrase number in the TA's summary). Similar results can be observed when the length limit is based on the number of words, but cannot be reported here due to page limit.

First, our proposed method (last row), which clusters the extracted phrases and uses LexRank to score them, can outperform all the baselines over all three ROUGE scores in terms of F-measure. In addition, the proposed model performs better than the clustering and LexRank alone. Through a paired *t*-test, our model outperforms LexRank statistically in terms of precision for all three ROUGE scores and significantly improves Clustering+Medoid on all R-2 scores (except the precision with 0.06 p-value). We believe that the semantic similarity based clustering complements LexRank in two ways: 1) LexRank depends on the cosine similarity of TF-IDF vectors to build the graph while the clustering takes semantic similarity into account. 2) The clustering performed a global selection to form a summary by grouping similar phrases and ranking them by the number of covered students (similar to what the human did). Compared to LexRank, our approach captures the student coverage explicitly. While modifying LexRank by using semantic similarity is possible, estimating the student coverage is not straightforward.

Second, OriMEAD tends to select long sentences, resulting in a high recall but a low precision. The phrase version (MEAD) improves both the P and F scores by removing unnecessary parts in the original sentences.

Lastly, the proposed method outperforms the MMR based baselines on the precision and F-measure of all three ROUGE scores. We observed that the MMR baselines suffer from the issue of diverse expressions used the students (e.g., "graphs" and "charts").

6 Conclusion

In this paper, we presented a novel application to summarize student feedback to reflection prompts by a combination of phrase extraction, phrase clustering and phrase ranking. It makes use of metric clustering to rank the phrases by their student coverage, taking the information source into account. Experimental results demonstrate the good effectiveness of the model. While the proposed method improved the performance against MMR, other summarization methods without an additional MMR component do exist, in-

[5]The default Length parameter in MEAD is changed to 1 from its default value 9 and the position feature is removed, yielding better performance.

[6]For each MMR based baseline, the parameter is optimized with a grid search on the development data set.

cluding SumBasic (Vanderwende et al., 2007), KLSUM and TopicSUM (Haghighi and Vanderwende, 2009). An initial experiment shows they do not yield better performance with default parameters. However, we will revisit it since these methods are meant for full sentences and are not optimized within the phrase framework.

In the future, we plan to have additional annotation to evaluate the relative importance using the student coverage numbers. We also deployed CourseMIRROR in a statistics class in Spring 2015 and have created gold-standard summaries, which will allow us to both replicate the intrinsic evaluation of this paper with a new and larger dataset as well conduct an extrinsic evaluation beyond ROUGE scores. Finally, we are interested in applying our summarization approach to other types of user-generated content from mobile-applications (e.g., review comments).

Acknowledgments

This research is partially 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 and Xiangmin Fan for developing the CourseMIRROR application and for valuable suggestions about the proposed summarization algorithm. We also thank anonymous reviewers for insightful comments and suggestions.

References

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

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. ACM.

Ronan Collobert. 2011. Deep learning for efficient discriminative parsing. In *International Conference on Artificial Intelligence and Statistics*, number EPFL-CONF-192374.

Dan Ştefănescu, Rajendra Banjade, and Vasile Rus. 2014. Latent semantic analysis models on wikipedia and tasa. In *The 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 26–31.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, December.

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.

James Gung and Jugal Kalita. 2012. Summarization of historical articles using temporal event clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 631–635.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado, June.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1262–1273, June.

Leonard Kaufman and Peter Rousseeuw. 1987. Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Method*, pages 405–416.

Yanran Li and Sujian Li. 2014. Query-focused multi-document summarization: Combining a topic model with graph-based semi-supervised learning. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1197–1207, August.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 257–266.

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 the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 16–20, June.

Olena Medelyan, Eibe Frank, and Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, EMNLP '09, pages 1318–1327.

Muhsin Menekse, Glenda Stump, Stephen J. Krause, and Michelene T.H. Chi. 2011. The effectiveness of students daily reflections on learning in engineering context. In *Proceedings of the American Society for Engineering Education (ASEE) Annual Conference*, Vancouver, Canada.

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

Vasile Rus, Mihai C Lintean, Rajendra Banjade, Nobal B Niraula, and Dan Stefanescu. 2013. Semilar: The semantic similarity toolkit. In *ACL (Conference System Demonstrations)*, pages 163–168.

Yoshihiro Ueda, Mamiko Oka, Takahiro Koyama, and Tadanobu Miyauchi. 2000. Toward the "at-a-glance" summary: Phrase-representation summarization method. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 878–884.

Gerard Van den Boom, Fred Paas, Jeroen JG Van Merrienboer, and Tamara Van Gog. 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.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.*, 43(6):1606–1618, November.

Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 299–306.

Yi-fang Brook Wu, Quanzhi Li, Razvan Stefan Bot, and Xin Chen. 2005. Domain-specific keyphrase extraction. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 283–284.

Rui Yang, Zhan Bu, and Zhengyou Xia. 2012. Automatic summarization for chinese text using affinity propagation clustering and latent semantic analysis. In *Proceedings of the 2012 International Conference on Web Information Systems and Mining*, WISM'12, pages 543–550.

Benyu Zhang, Hua Li, Yi Liu, Lei Ji, Wensi Xi, Weiguo Fan, Zheng Chen, and Wei-Ying Ma. 2005. Improving web search results using affinity graph. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 504–511. ACM.

Xiaojin Zhu, Andrew Goldberg, Jurgen Van Gael, and David Andrzejewski. 2007. Improving diversity in ranking using absorbing random walks. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 97–104, April.