Natural Language Engineering **24** (6): 887–920. © Cambridge University Press 2018
doi:10.1017/S1351324918000323

*A novel ILP framework for summarizing content with high lexical variety**

WENCAN LUO[1], FEI LIU[2], ZITAO LIU[3]
and DIANE LITMAN[1]

[1]*University of Pittsburgh, Pittsburgh, PA 15260, U.S.*
e-mail: wencan@cs.pitt.edu; litman@cs.pitt.edu
[2]*University of Central Florida, Orlando, FL 32816, U.S.*
e-mail: feiliu@cs.ucf.edu
[3]*Pinterest, Inc., San Francisco, CA 94103, U.S.*
e-mail: zitaoliu@pinterest.com

(*Received 9 May 2017; revised 23 July 2018; accepted 24 July 2018; first published online 13 September 2018*)

Abstract

Summarizing content contributed by individuals can be challenging, because people make different lexical choices even when describing the same events. However, there remains a significant need to summarize such content. Examples include the student responses to post-class reflective questions, product reviews, and news articles published by different news agencies related to the same events. High lexical diversity of these documents hinders the system's ability to effectively identify salient content and reduce summary redundancy. In this paper, we overcome this issue by introducing an integer linear programming-based summarization framework. It incorporates a low-rank approximation to the sentence-word cooccurrence matrix to intrinsically group semantically similar lexical items. We conduct extensive experiments on datasets of student responses, product reviews, and news documents. Our approach compares favorably to a number of extractive baselines as well as a neural abstractive summarization system. The paper finally sheds light on when and why the proposed framework is effective at summarizing content with high lexical variety.

1 Introduction

Summarization is a promising technique for reducing information overload. It aims at converting long text documents to short, concise summaries conveying the essential content of the source documents (Nenkova and McKeown 2011). Extractive methods focus on selecting important sentences from the source and concatenating them to

*This research is supported by an internal grant from the Learning Research and Development Center at the University of Pittsburgh as well as by an Andrew Mellon Predoctoral Fellowship to the first author. We are grateful to Logan Lebanoff for helping with the experiments. We also thank Muhsin Menekse, the CourseMIRROR team, and Wenting Xiong for providing or helping to collect some of our datasets. We thank Jingtao Wang, Fan Zhang, Huy Nguyen, and Zahra Rahimi for valuable suggestions about the proposed summarization algorithm.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

form a summary, whereas abstractive methods can involve a number of high-level text operations such as word reordering, paraphrasing, and generalization (Jing and McKeown 1999). To date, summarization has been successfully exploited for a number of text domains, including news articles (Barzilay, McKeown and Elhadad 1999; Dang and Owczarzak 2008; Durrett, Berg-Kirkpatrick and Klein 2016; Grusky *et al.* 2018), product reviews (Gerani *et al.* 2014), online forum threads (Tarnpradab, Liu and Hua 2017), meeting transcripts (Liu and Liu 2013), scientific articles (Teufel and Moens 2002; Qazvinian *et al.* 2013), student course responses (Luo and Litman 2015; Luo *et al.* 2016b), and many others.

Summarizing content contributed by multiple authors is particularly challenging. This is partly because people tend to use different expressions to convey the same semantic meaning. In a recent study of summarizing student responses to post-class reflective questions, Luo *et al.* (2016b) observe that the students use distinct lexical items such as *bike elements* and *bicycle parts* to refer to the same concept. The student responses frequently contain expressions with little or no word overlap, such as *the main topics of this course* and *what we will learn in this class*, when they are prompted with 'describe what you found most interesting in today's class.' A similar phenomenon has also been observed in the news domain, where reporters use different nicknames, e.g., 'Bronx Zoo' and 'New York Highlanders,' to refer to the baseball team 'New York Yankees.' Luo *et al.* (2016b) report that about eighty per cent of the document bigrams occur only once or twice for the news domain, whereas the ratio is ninety-seven per cent for student responses, suggesting the latter domain has a higher level of lexical diversity. When source documents contain diverse expressions conveying the same meaning, it can hinder the summarization system's ability to effectively identify salient content from the source documents. It can also increase the summary redundancy if lexically distinct but semantically similar expressions are included in the summary.

Existing neural encoder–decoder models may not work well at summarizing such content with high lexical variety (Rush *et al.* 2015; Nallapati, Xiang and Zhou 2016; Paulus, Xiong and Socher 2017; See, Liu and Manning 2017). On one hand, training the neural sequence-to-sequence models requires a large amount of parallel data. The cost of annotating gold-standard summaries for many domains such as student responses can be prohibitive. Without sufficient labeled data, the models can only be trained on automatically gathered corpora, where an instance often includes a news article paired with its title or a few highlights. On the other hand, the summaries produced by existing neural encoder–decoder models are far from perfect. The summaries are mostly extractive with minor edits (See *et al.* 2017),[1] contain repetitive words and phrases (Suzuki and Nagata 2017) and may not accurately reproduce factual details (Cao *et al.* 2018; Song, Zhao and Liu 2018). We examine the performance of a state-of-the-art neural summarization model in Section 6.2.

[1] See *et al.* (2017) suggest that the pointer-generator model can copy thirty-five per cent of summary sentences from the source documents. Similar findings are also reported by Liao *et al.* (2018), where 99.6 per cent of the summary unigrams, 95.2 per cent of bigrams, and 87.2 per cent of trigrams generated by the pointer-generator networks appear in the source texts.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

In this work, we propose to augment the integer linear programming (ILP)-based summarization framework with a low-rank approximation of the cooccurrence matrix, and further evaluate the approach on a broad range of datasets exhibiting high lexical diversity. The ILP framework, being extractive in nature, has demonstrated considerable success on a number of summarization tasks (Gillick and Favre 2009; Berg-Kirkpatrick, Gillick and Klein 2011). It generates a summary by selecting a set of sentences from the source documents. The sentences shall maximize the coverage of important source content, while minimizing the redundancy among themselves. At the heart of the algorithm is a sentence-concept cooccurrence matrix, used to determine if a sentence contains important concepts and whether two sentences share the same concepts. We introduce a low-rank approximation to the cooccurrence matrix and optimize it using the proximal gradient method. The resulting system thus allows different sentences to share cooccurrence statistics. For example, 'The activity with the *bicycle parts*' will be allowed to partially contain '*bike elements*' although the latter phrase does not appear in the sentence. The low-rank matrix approximation provides an effective way to implicitly group lexically diverse but semantically similar expressions. It can handle out-of-vocabulary expressions and domain-specific terminologies well, hence being a more principled approach than heuristically calculating similarities of word embeddings.

Our research contributions of this work include the following:

- We present a novel ILP framework to summarize documents contributed by multiple authors and exhibiting high lexical variety. The framework is evaluated on eight different datasets, ranging from student course responses to product reviews and news articles. To the best of our knowledge, our work is one of the few summarization studies assessing the proposed approach on a broad spectrum of datasets.
- Through extensive experiments, we show that the new ILP framework compares favorably to both extractive baselines and a state-of-the-art neural abstractive summarization system. We further investigate the properties of both our system and various datasets to understand when and why the proposed approach is effective at summarizing content with high lexical variety.

In the following sections, we first present a thorough review of the related work (Section 2), then introduce our ILP summarization framework (Section 3) with a low-rank approximation of the cooccurrence matrix optimized using the proximal gradient method (Section 4). Experiments are performed on a collection of eight datasets (Section 5) containing student responses to post-class reflective questions, product reviews, peer reviews, and news articles. Intrinsic evaluation (Section 6.1) shows that the low-rank approximation algorithm can effectively group distinct expressions used in similar semantic context. For extrinsic evaluation (Section 6.2), our proposed framework obtains competitive results in comparison to the state-of-the-art summarization systems. Finally, we conduct comprehensive

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

studies analyzing the characteristics of the datasets and suggest critical factors that affect the summarization performance (Section 7).

2 Related work

Extractive summarization has undergone great development over the past decades. It focuses on extracting relevant sentences from a single document or a cluster of documents related to a particular topic. Various techniques have been explored, including maximal marginal relevance (Carbonell and Goldstein 1998), submodularity (Lin and Bilmes 2010), ILP (Gillick, Favre and Hakkani-Tür 2008; Almeida and Martins 2013; Li *et al.* 2013, 2014; Durrett, Berg-Kirkpatrick and Klein 2016), minimizing reconstruction error (He *et al.* 2012), graph-based models (Erkan and Radev 2004; Radev *et al.* 2004; Cohan and Goharian 2015, 2017), determinantal point processes (Kulesza and Taskar 2012), neural networks and reinforcement learning (Yasunaga *et al.* 2017; Narayan, Cohen and Lapata 2018) among others. Nonetheless, most studies are bound to a single dataset and few approaches have been evaluated in a cross-domain setting. In this paper, we propose an enhanced ILP framework and evaluate it on a broad range of datasets. We present an in-depth analysis of the dataset characteristics derived from both source documents and reference summaries to understand how domain-specific factors may affect the applicability of the proposed approach.

Neural summarization has seen promising improvements in recent years with encoder–decoder models (Rush *et al.* 2015; Nallapati *et al.* 2016). The encoder condenses the source text to a dense vector, whereas the decoder unrolls the vector to a summary sequence by predicting one word at a time. A number of studies have been proposed to deal with out-of-vocabulary words (See *et al.* 2017), improve the attention mechanism (Chen *et al.* 2016; Tan, Wan and Xiao 2017; Zhou *et al.* 2017), avoid generating repetitive words (See *et al.* 2017; Suzuki and Nagata 2017), adjust summary length (Kikuchi *et al.* 2016), encode long text (Celikyilmaz *et al.* 2018; Cohan *et al.* 2018), and improve the training objective (Ranzato *et al.* 2016; Paulus *et al.* 2017; Guo, Pasunuru and Bansal 2018). To date, these studies focus primarily on single-document summarization and headline generation. This is partly because training neural encoder–decoder models requires a large amount of parallel data, yet the cost of annotating gold-standard summaries for most domains can be prohibitive. We validate the effectiveness of a state-of-the-art neural summarization system (See *et al.* 2017) on our collection of datasets and report results in Section 6.2.

In this paper, we focus on the ILP-based summarization framework and propose enhancements to it to summarize text content with high lexical diversity. The ILP framework is shown to perform strongly on extractive summarization (Gillick and Favre 2009; Martins and Smith 2009; Berg-Kirkpatrick *et al.* 2011). It produces an optimal selection of sentences that (i) maximize the coverage of important concepts discussed in the source, (ii) minimize the redundancy in pairs of selected sentences, and (iii) ensure the summary length does not exceed a limit. Previous work has largely focused on improving the estimation of concept weights in the ILP framework (Galanis, Lampouras and Androutsopoulos 2012; Qian and Liu

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

2013; Boudin, Mougard and Favre 2015; Li, Liu and Zhao 2015; Durrett *et al.* 2016). However, distinct lexical items such as 'bike elements' and 'bicycle parts' are treated as different concepts and their weights are not shared. In this paper, we overcome this issue by proposing a low-rank approximation to the sentence-concept cooccurrence matrix to intrinsically group lexically distinct but semantically similar expressions, they are considered as a whole when maximizing concept coverage and minimizing redundancy.

Our work is also different from the traditional approaches using dimensionality reduction techniques such as non-negative matrix factorization and latent semantic analysis for summarization (Wang *et al.* 2008; Lee *et al.* 2009; Conroy *et al.* 2013; Conroy and Davis 2015; Wang *et al.* 2016a). In particular, Wang *et al.* (2008) use non-negative matrix factorization to group sentences into clusters; Conroy *et al.* (2013) explore non-negative matrix factorization and latent semantic analysis to obtain better estimates of term weights; Wang *et al.* (2016a) use low-rank approximation to cast sentences and images to the same embedding space. Different from the above methods, our proposed framework focuses on obtaining a low-rank approximation of the cooccurrence matrix embedded in the ILP framework, so that diverse expressions can share cooccurrence frequencies. Note that out-of-vocabulary expressions and domain-specific terminologies are abundant in our datasets, therefore simply calculating the lexical overlap (Rus *et al.* 2013) or cosine similarity of word embeddings (Goldberg and Levy 2014) cannot serve our goal well.

This manuscript extends our previous work on summarizing student course responses (Luo and Litman 2015; Luo, Liu and Litman 2016a; Luo *et al.* 2016b) submitted after each lecture via a mobile app named CourseMIRROR (Luo *et al.* 2015; Fan *et al.* 2015, 2017). The students are asked to respond to reflective prompts such as 'describe what you found most interesting in today's class' and 'describe what was confusing or needed more detail.' For large classes with hundreds of students, it can be quite difficult for instructors to manually analyze the student responses, hence the help of automatic summarization. Our extensions of this work are along three dimensions: (i) we crack the 'black-box' of the low-rank approximation algorithm to understand if it indeed allows lexically diverse but semantically similar items to share cooccurrence statistics; (ii) we compare the ILP-based summarization framework with the state-of-the-art baselines, including a popular neural encoder–decoder model for summarization; (iii) we expand the student feedback datasets to include responses collected from materials science and engineering, statistics for industrial engineers, and data structures. We additionally experiment with reviews and news articles. Analyzing the unique characteristics of each dataset allows us to identify crucial factors influencing the summarization performance.

With the fast development of Massive Open Online Courses platforms, more attention is being dedicated to analyzing educationally oriented language data. These studies seek to identify student leaders from Massive Open Online Courses discussion forums (Moon, Potdar and Martin 2014), perform sentiment analysis on student discussions (Wen, Yang and Ros 2014b), improve student engagement and reducing student retention (Wen, Yang and Rose 2014a; Rose and Siemens

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

2014), and using language generation techniques to automatically generate feedback to students (Gkatzia *et al.* 2013). Our focus of this paper is to automatically summarizing student responses so that instructors can collect feedback in a timely manner. We expect the developed summarization techniques and result analysis will further summarization research in similar text genres exhibiting high lexical variety.

3 ILP formulation

Let $\mathcal{D}$ be a set of documents that consist of M sentences in total. Let $y_j \in \{0,1\}$, $j = \{1,\ldots,M\}$ 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 $\mathcal{D}$. $z_i \in \{0,1\}$, $i = \{1,\ldots,N\}$ 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 ith concept appears in the jth 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 this paper.

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)).

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

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

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

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

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

$$z_i \in \{0,1\} \tag{6}$$

The above ILP can be transformed to matrix representation:

$$\max_{\boldsymbol{y},\boldsymbol{z}} \quad \boldsymbol{w}^{\top}\boldsymbol{z} \tag{7}$$

$$s.t. \quad \boldsymbol{A}\,\boldsymbol{y} \geq \boldsymbol{z} \tag{8}$$

$$\boldsymbol{A}\,\mathrm{diag}(\boldsymbol{y}) \leq \boldsymbol{Z} \tag{9}$$

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

$$\boldsymbol{\eta}^\top \boldsymbol{y} \leq L \tag{10}$$

$$\boldsymbol{y} \in \{0,1\}^M \tag{11}$$

$$\boldsymbol{z} \in \{0,1\}^N \tag{12}$$

We use boldface letters to represent vectors and matrices. $\boldsymbol{Z} = [\boldsymbol{z}, \ldots, \boldsymbol{z}] \in \mathbb{R}^{N \times M}$ is an auxiliary matrix created by horizontally stacking the concept vector $\boldsymbol{z} \in \mathbb{R}^N$ M times. Constraint set (Eq. (9)) specifies that a sentence is selected indicates that all concepts it carries have been selected. It corresponds to $N \times M$ constraints of the form $A_{i,j}\, y_j \leq z_i$, where $i \in [N], j \in [M]$.

As far as we know, this is the first-of-its-kind matrix representation of the ILP framework. It clearly shows the two important components of this framework, including (1) the concept-sentence cooccurrence matrix $\boldsymbol{A}$, and (2) concept weight vector $\boldsymbol{w}$. Existing work focus mainly on generating better estimates of concept weights ($\boldsymbol{w}$), while we focus on improving the cooccurrence matrix $\boldsymbol{A}$.

4 Our approach

Because of the lexical diversity problem, we suspect the cooccurrence matrix $\boldsymbol{A}$ may not establish a faithful correspondence between sentences and concepts. A concept may be conveyed using multiple bigram expressions; however, the current cooccurrence matrix only captures a binary relationship between sentences and bigrams. For example, we ought to give partial credit to 'bicycle parts' given that a similar expression 'bike elements' 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 $\boldsymbol{A}$ is highly sparse. Only 3.7 per cent of the entries are non-zero in the student response datasets on average (Section 5). We therefore propose to *impute* the cooccurrence matrix by filling in missing values (i.e., matrix completion (MC)). This is accomplished by approximating the original cooccurrence matrix using a low-rank matrix. The low-rankness encourages similar concepts to be shared across sentences.

The ILP with a low-rank approximation of the cooccurrence matrix can be formalized as follows:

$$\max_{\boldsymbol{y},\boldsymbol{z}} \quad \boldsymbol{w}^\top \boldsymbol{z} \tag{13}$$

$$s.t. \quad \hat{\boldsymbol{A}}\, \boldsymbol{y} \geq \boldsymbol{z} \tag{14}$$

$$\hat{\boldsymbol{A}}\, \mathrm{diag}(\boldsymbol{y}) \leq \boldsymbol{Z} \tag{15}$$

$$\boldsymbol{\eta}^\top \boldsymbol{y} \leq L \tag{16}$$

$$\boldsymbol{y} \in \{0,1\}^M \tag{17}$$

$$\boldsymbol{z} \in [0,1]^N \tag{18}$$

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

The low-rank approximation process makes two notable changes to the existing ILP framework.

- It extends the domain of A_{ij} from binary to a continuous scale [0, 1] (Eqs. (14) and (15)), which offers a better sentence-level semantic representation.
- The binary concept variables (z_i) are also relaxed to continuous domain [0, 1] (Eq. (18)), which allows the concepts to be 'partially' included in the summary.

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

$$\min_{\hat{A}\in\mathbb{R}^{N\times M}} \frac{1}{2}\sum_{(i,j)\in\Omega}(A_{ij}-\hat{A}_{ij})^2 + \lambda\left\|\hat{A}\right\|_*, \tag{19}$$

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

$$[P_\Omega(\hat{A})]_{ij} = \begin{cases} \hat{A}_{ij} & (i,j)\in\Omega \\ 0 & (i,j)\notin\Omega \end{cases} \tag{20}$$

our objective function (Eq. (19)) can be succinctly represented as

$$\min_{\hat{A}\in\mathbb{R}^{N\times M}} \frac{1}{2}\|P_\Omega(A)-P_\Omega(\hat{A})\|_F^2 + \lambda\|\hat{A}\|_* \tag{21}$$

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

Following Mazumder, Hastie and Tibshirani (2010), we optimize Eq. (21) using the proximal gradient descent algorithm. The update rule is

$$\hat{A}^{(k+1)} = \text{prox}_{\lambda\rho_k}\left(\hat{A}^{(k)} + \rho_k\left(P_\Omega(A)-P_\Omega(\hat{A})\right)\right) \tag{22}$$

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

Since the gradient of $\frac{1}{2}\|P_\Omega(A)-P_\Omega(\hat{A})\|_F^2$ 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.

5 Datasets

To demonstrate the generality of the proposed approach, we consider three distinct types of corpora, ranging from student response datasets from four different courses to three sets of reviews to one benchmark of news articles. The corpora are summarized in Table 1.

Student responses. Research has explored using reflection prompts/muddy cards/1-minute papers to promote and collect reflections from students (Wilson 1986;

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 1. *Selected summarization datasets*

	Statistics				
	Tasks	Docs/task	WC/task	WC/sen	Length
Student response (Eng)*	36	49	375.4	9.1	30
Student response (Stat2015)*	44	39	233.1	6.0	15
Student response (Stat2016)*	48	42	149.3	4.3	13
Student response (CS2016)	46	22	223.1	8.8	16
Reviews (camera)	3	18	1927.0	22.7	216
Reviews (movie)	3	18	8014.0	24.4	242
Reviews (peer)	3	18	1543.7	19.2	190
News articles (DUC04)*	50	10	5171.6	22.4	105

Publicly available datasets are marked with an asterisk (*). The statistics involve the number of summarization tasks (Tasks), average number of documents per task (Docs/task), average word count per task (WC/task), average word count per sentence (WC/sen), and average number of words in human summaries (Length).

Mosteller 1989; Harwood 1996). However, it is expensive and time consuming for humans to summarize such feedback. It is therefore desirable to automatically summarize the student feedback produced in online and offline environments, although it is only recently that a data collection effort to support such research has been initiated (Fan *et al.* 2015; Luo *et al.* 2015). In our data, one particular type of student response is considered, named *reflective feedback* (Boud *et al.* 2013), which has been shown to enhance interaction between instructors and students by educational researchers (Van den Boom *et al.* 2004; Menekse *et al.* 2011). More specifically, students are presented with the following prompts after each lecture and asked to provide responses: (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).

To test generality, we gathered student responses from four different courses, as shown in Table 1. The first one was collected by Menekse *et al.* (2011) using paper-based surveys from an introductory materials science and engineering class (henceforth **Eng**) taught in a major U.S. university, and a subset is made public by us (Luo and Litman 2015), available at the link: `http://www.coursemirror.com/download/dataset`. The remaining three courses are collected by us using a mobile application, CourseMIRROR (Luo *et al.* 2015; Fan *et al.* 2015) and then the reference summaries for each course are created by human annotators with the proper background. The human annotators are allowed to create abstract summaries using their own words in addition to selecting phrases directly from the responses. While the second and third datasets are from the same course, Statistics for Industrial Engineers, they were taught in 2015 and 2016, respectively (henceforth **Stat2015**

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 2. *Example prompt, student responses, and one human summary. 'S1'–'S10' are student IDs.*

Prompt
Describe what you found most interesting in today's class
Student responses
S1: Guilt analogy[r]
S2: Error bounding[g] is interesting and useful
S3: the idea of c and finding that error[g] looked great to me
S4: You stated that the concept of the error boundary[g] is abstract however i got it very well
S5: determining the probability of the error[g] while rejecting ho[b].
S6: The process of hypothesis testing[y]
S7: Determining the critical value for error[g]
S8: H1 and Ho conditions[b]
S9: Baydogan finally check the students in the class. But i think it must be in every lecture even in the PS
S10: Testing whether the information we have is true or not with hypothesis testing[y] method was interesting
...
Human summary 1
– Hypothesis testing (in general)[y]
– Error bounding[g]
– Guilt analogy helpful[r]
– Conditions for H1 and H0[b]
– Good use of examples[m]

The summary and phrase highlights are manually created by annotators. Phrases that bear the same color belong to the same issue. The superscripts of the phrase highlights are imposed by the authors to differentiate colors when printed in grayscale (y: yellow, g: green, r: red, b: blue, and m: magenta).

and **Stat2016**), at the Boğaziçi University in Turkey.[2] The course was taught in English while the official language is Turkish. The last one is from a fundamental undergraduate Computer Science course (data structures) at a local U.S. university taught in 2016 (henceforth **CS2016**).

Another reason we choose the student responses is that we have advanced annotation allowing us to perform an intrinsic evaluation to test whether the low-rank approximation does capture similar concepts or not. An example of the annotation is shown in Table 2, where phrases in the student responses that are semantically the same as the summary phrases are highlighted with the same color by human annotators. For example, 'error bounding' (S2), 'error boundary' (S4), 'finding that error' (S3), and 'determining the critical value for error' (S7) are

[2] Publicly available at `http://www.coursemirror.com/download/dataset2` (Luo *et al.* 2016a)

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 3. *Example movie reviews*

'Forrest Gump' is one of the best movies of all time, guaranteed.
I just love this movie.
It truly is amazing...
What an amazing story and moving meaning.
I really just love this movie and it has such a special place in my heart.
And anyone who hasn't seen it or who thinks that don't like it I seriously suggest seeing it or seeing it again.
That movie teaches you so much about life and the meaning of it.
This is one masterpiece of a movie that will not be forgotten about in a long time.
This is a powerful yet charming movie; fun for its special effects and profound in how it keeps you thinking long after it's over.
It may change your lifeOne hell of a movie; it will be close to my heart forever!
It is something to mull over for a long time.
The performances are just so unforgettable and never get out of your head.
I've watched the movie about once every two years since then.
The lines are so memorable, touching, and sometimes hilarious.We have Forrest Gump (Tom Hanks), not the sharpest tool in the box, his I.Q.
Well done, well acted, and well directed to pythagorean procision. A++
This story is beautiful and will inspire everyone to go the distance and see the world like Forrest did and will never give up on their dreams.10/10
A++
You 'd be a fool to miss it.Bottom Line : 4 out of 4 (own this movie)

semantically equivalent to 'Error bounding' in the human summary. Details of the intrinsic evaluation are introduced in 6.1.

Product and peer reviews. The review datasets are provided by Xiong and Litman (2014), consisting of three categories. The first one is a subset of product reviews from a widely used dataset in review opinion mining and sentiment analysis, contributed by Jindal and Liu (2008). In particular, it randomly sampled three set of reviews from a representative product (digital camera), each with eighteen reviews from an individual product type (e.g., summarizing eighteen camera reviews for Nikon D3200). The second one is movie reviews crawled from IMDB.com by the authors themselves. The third one is peer reviews collected in a college-level history class from an online peer-review reciprocal system, SWoRD (Cho 2008). The average number of sentences per review set is eighty-five for camera reviews, 328 for movie reviews, and eighty for peer review; the average number of words per sentence in the camera, movie, and peer reviews are twenty three, twenty-four, and nineteen, respectively. The human summaries were collected in the form of online surveys (one survey per domain) hosted by Qualtrics. Each human summary contains ten sentences from users' reviews. Example movie reviews are shown in Table 3.

News articles. Most summarization work focuses on news documents, as driven by the Document Understanding Conferences (DUC) and Text Analysis Conferences.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 4. *Example sentences from news*

Samaranch expressed surprise at allegations made by the IOC executive board member Marc Hodler of Switzerland that agents were offering to sell I.O.C. members' votes for payments from bidding cities.
Moving quickly to tackle an escalating corruption scandal, IOC leaders questioned Salt Lake City officials Friday in the first ever investigation into alleged vote-buying by an Olympic city.
Acting with unusual speed, the International Olympic Committee set up a special investigative panel that immediately summoned the organizers of the 2002 Salt Lake Games to address the bribery allegations.
It's the most serious case of alleged ethical misconduct investigated by the IOC since former U.S. member Robert Helmick was accused of conflict of interest in 1991.
This is the first time the IOC has ever investigated possible bribery by bidding cities, despite previous rumors and allegations of corruption in other Olympic votes.
Hodler said a group of four agents, including one IOC member, have been involved in promising votes for payment.
Samaranch Sunday ruled out taking the Games from Salt Lake City.
I can't be stronger in saying I don't consider it a possibility whatsoever of the games being withdrawn from Salt Lake City.
The chief investigator refused to rule out the possibility of taking the games away from Salt Lake City – though that scenario is considered highly unlikely.

For comparison, we select DUC 2004[3] to evaluate our approach (henceforth **DUC04**), which is widely used in the literature (Lin 2004; Hong *et al.* 2014; Ren *et al.* 2016; Takase *et al.* 2016; Wang *et al.* 2016b). It consists of fifty clusters of Text REtrieval Conference documents, from the following collections: AP newswire, 1998–2000; New York Times newswire, 1998–2000; Xinhua News Agency (English version), 1996–2000. Each cluster contained on average ten documents. The task is to create a short summary ($\leq$665 bytes) of each cluster. Example news sentences are shown in Table 4.

6 Experiments

In this section, we evaluate the proposed method intrinsically in terms of whether the cooccurrence matrix after the low-rank approximation is able to capture similar concepts on student response datasets, and also extrinsically in terms of the end task of summarization on all corpora. In the following experiments, summary length is set to be the average number of words in human summaries or less. For the MC algorithm, we perform grid search (on a scale of [0, 5] with stepsize 0.5) to tune the hyper-parameter λ (Eq. (19)) with a leave-one-lecture-out (for student responses) or leave-one-task-out (for others) cross-validation.

[3] `http://duc.nist.gov/duc2004/`

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 5. *Associated bigrams that do not appear in the sentence, but after Matrix Completion, yield a decent correlation (cell value greater than 0.9) with the corresponding sentence*

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 two *phase regions*	*phase diagram*
I tried to follow along but I could not *grasp the* concepts	*understand the*
No problems except for the specific equations used to determine properties from the stress – *strain graph*	*strain curves*
Why delete the first entry in the *linked bag* instead of just moving the pointers from the node before the deleted node to the node after	*linked list*
You make *a movie* that romanticizes the '50's, '60's, and '70's, and with enough publicity and a good enough soundtrack ...	*the film*
U.S. officials have said the construction ...	*united states*
American officials have said spy satellites ...	*united states*
It also sought to cast *Gates* as an obsessed man who feared the tiny Netscape Communications Corp. and its potential threat to his domination of the market for *Internet browsers*, the software used to navigate the World Wide Web.	*that microsoft*

6.1 Intrinsic evaluation

When examining the imputed sentence-concept cooccurrence matrix, we notice some interesting examples that indicate the effectiveness of the proposed approach, shown in Table 5.

We want to investigate whether the MC helps to capture similar concepts (i.e., bigrams). Recall that, if a bigram i is similar to another bigram in a sentence j, the sentence j should assign a partial score to the bigram i after the low-rank approximation. For instance, 'The activity with the bicycle parts' should give a partial score to 'bike elements' since it is similar to 'bicycle parts.' Note that, the cooccurrence matrix $\boldsymbol{A}$ measures whether a sentence includes a bigram or not. Without MC, if a bigram i does not appear in a sentence j, $A_{ij} = 0$. After MC, $\hat{A}_{ij}$ ($\hat{A}$ is the low-rank approximation matrix of A) becomes a continuous number ranging from 0 to 1 (negative values are truncated). Therefore, $\hat{A}_{ij} > 0$ does not

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

necessarily mean the sentence contains a similar bigram, since it might also give positive scores to non-similar bigrams. To solve this issue, we propose two different ways to test whether the MC really helps to capture similar concepts.

- H1.a: A bigram receives a higher partial score in a sentence that contains similar bigram(s) to it than a sentence that does not. That is, if a bigram i is similar to one of bigrams in a sentence j^+, but not similar to any bigram in another sentence j^-, then after MC, $\hat{A}_{ij^+} > \hat{A}_{ij^-}$.
- H1.b: A sentence gives higher partial scores to bigrams that are similar to its own bigrams than bigrams that are different from its own. That is, if a sentence j has a bigram that is similar to i^+, but none of its bigrams is similar to i^-, then, after MC, $\hat{A}_{i^+j} > \hat{A}_{i^-j}$.

In order to test these two hypotheses, we need to construct gold-standard pairs of similar bigrams and pairs of different bigrams, which can be automatically obtained with the phrase-highlighting data (Table 2). We first extract a candidate bigram from a phrase if and only if a single bigram can be extracted from the phrase. In this way, we discard long phrases if there are multiple candidate bigrams among them in order to avoid ambiguity as we cannot validate which of them match another target bigram. A bigram is defined as two words and at least one of them is not a stop word. We then extract every pair of candidate bigrams that are highlighted in the same color as similar bigrams. Similarly, we extract every pair of candidate bigrams that are highlighted as different colors as different bigrams. For example, 'bias reduction' is a candidate phrase, which is similar to 'bias correction' since they are in the same color.

To test H1.a, given a bigram i, a bigram i^+ that is similar to it, and a bigram i^- that is different from it, we can select the bigram i, and the sentence j^+ that contains i^+, and the sentence j^- that contains i^-. We ignore j^- if it contains any other bigram that is similar to i to eliminate the compounded case that both similar and different bigrams are within one sentence. Note, if there are multiple sentences containing i^+, we consider each of them. In this way, we construct a triple $\langle i, j^+, j^- \rangle$, and test whether $\hat{A}_{ij^+} > \hat{A}_{ij^-}$. To test H1.b, for each pair of similar bigrams $\langle i, i^+ \rangle$ and different bigrams $\langle i, i^- \rangle$, we select the sentence j that contains i so that we construct a triple $\langle i^+, i^-, j \rangle$, and test whether $\hat{A}_{i^+j} > \hat{A}_{i^-j}$. We also filtered out j that contains similar bigram(s) to i^- to remove the compounded effect. In this way, we collected a gold-standard dataset to test the two hypotheses above as shown in Table 6.

The results are shown in Table 7. $\hat{A}_{ij^+} > \hat{A}_{ij^-}$ significantly on all three courses. That is, a bigram does receive a higher partial score in a sentence that contains similar bigram(s) to it than a sentence that does not. Therefore, H1.a holds. For H1.b, we only observe $\hat{A}_{i^+j} > \hat{A}_{i^-j}$ significantly on Stat2016 and there is no significant difference between $\hat{A}_{i^+j}$ and $\hat{A}_{i^-j}$ on the other two courses. First, the gold-standard dataset is still small in the sense that only a limited portion of bigrams in the entire dataset are evaluated. Second, the assumption that phrases annotated by different colors are not necessarily unrelated is too strong. For example, 'hypothesis testing' and 'H1 and Ho conditions' are in different colors in the example of Table 2, but one is a subtopic of the other. An alternative way to evaluate the hypothesis is

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 6. *A gold-standard data set was extracted from three student response corpora that have phrase-highlighting annotation*

Corpus	Bigrams	Similar pairs	Different pairs	$\langle i, j^+, j^- \rangle$	$\langle i^+, i^-, j \rangle$
Stat2015	516	198	698	404	279
Stat2016	1,673	638	1,928	1,188	228
CS2016	613	168	412	235	46

Statistics include: the number of bigrams, the number of pairs of similar bigrams and pairs of different bigrams, the number of tuples $\langle i, j^+, j^- \rangle$, and the number of $\langle i^+, i^-, j \rangle$. i is a bigram, j^+ is a sentence with a bigram similar to i, and j^- is a sentence with a bigram different from i. j is a sentence, i^+ is a bigram that is similar to a bigram in j, and i^- is a bigram that is different from any bigram in j.

Table 7. *Hypothesise testing: whether the matrix completion (MC) helps to capture similar concepts*

	Stat2015		Stat2016		CS2016	
H1.a	$\hat{A}_{ij^+}$ 0.122*	$\hat{A}_{ij^-}$ 0.056	$\hat{A}_{ij^+}$ 0.108*	$\hat{A}_{ij^-}$ 0.038	$\hat{A}_{ij^+}$ 0.238*	$\hat{A}_{ij^-}$ 0.089
H1.b	$\hat{A}_{i^+j}$ 0.147	$\hat{A}_{i^-j}$ 0.151	$\hat{A}_{i^+j}$ 0.132*	$\hat{A}_{i^-j}$ 0.074	$\hat{A}_{i^+j}$ 0.186	$\hat{A}_{i^-j}$ 0.149

* means $p < 0.05$ using a two-tailed paired t-test.

to let humans judge whether two bigrams are similar or not, which we leave for future work. Third, the gold standards are pairs of semantically similar bigrams, while MC captures bigrams that occurs in a similar context, which is not necessarily equivalent to semantic similarity. For example, the sentence 'graphs make it *easier to* understand concepts' in Table 7 is associated with 'hard to.'

6.2 Extrinsic evaluation

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) Pointer-Generator Networks (PGN) (See *et al.* 2017), a state-of-the-art neural encoder–decoder approach for abstractive summarization. The system was trained on the CNN/Daily Mail datasets (Hermann *et al.* 2015; Nallapati *et al.* 2016). (5) ILP (Berg-Kirkpatrick *et al.* 2011), a baseline ILP framework without MC.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

The PGN (See *et al.* 2017) describes a neural encoder–decoder architecture. It encourages the system to copy words from the source text via pointing, while retaining the ability to produce novel words through the generator. It also contains a coverage mechanism to keep track of what has been summarized, thus reducing word repetition. The PGN have not been tested for summarizing content contributed by multiple authors. In this study, we evaluate their performance on our collection of datasets.

For the ILP-based approaches, we use bigrams as concepts (bigrams consisting of only stopwords are removed[4]) and term frequency as concept weights. We leverage the cooccurrence statistics both within and across the entire corpus.[5] We also filtered out bigrams that appear only once in each corpus, yielding better ROUGE scores with lower computational cost. The results without using this low-frequency filtering are shown in the Appendix for comparison. In Table 8, we present summarization results evaluated by ROUGE (Lin 2004) and human judges.[6]

To compare with the official participants in DUC 2004 (Paul and James 2004), we selected the top-five systems submitted in the competition (ranked by R-1), together with the eight human annotators. The results are presented in Table 9.

ROUGE. It is a recall-oriented metric that compares system and reference summaries based on n-gram overlaps, which is widely used in summarization evaluation. In this work, we report ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4), and ROUGE-L (R-L) scores, which, respectively, measure the overlap of unigrams, bigrams, skip-bigram (with a maximum gap length of 4), and longest common subsequence. First, there is no winner for all datasets. MEAD is the best one on camera; SumBasic is best on Stat2016 and mostly on Stat2015; ILP is best on DUC04. The ILP baseline is comparable to the best participant (Table 9) and even has the best R-2. PGN is the worst, which is not surprising since it is trained on a different dataset, which may not generalize to our datasets. Our method ILP+MC is best on peer review and mostly on Eng and CS2016. Second, compared with ILP, our method works better on Eng, CS2016, movie, and peer.

These results show our proposed method does not always better than the ILP framework, and no single summarization system wins on all datasets. It is perhaps not surprising to some extent. The *no free lunch theorem* for machine learning (Wolpert 1996) states that, averaged overall possible data-generating distributions, every classification algorithm has the same error rate when classifying

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

[5] We construct one single matrix for each entire corpus except DUC04. For example, the cooccurrence matrix for Eng includes 1,492 distinct sentences and 9,239 unique bigrams, from all lectures and prompts. For DUC04, we construct a matrix for each document cluster instead of the entire corpus due to its high computational cost.

[6] The results on Eng are slightly different from the results published by Luo *et al.* (2016b) as we used leave-one-lecture-out cross-validation instead of three-fold cross-validation to select the parameter λ. We also changed the order of student responses by grouping same responses together, affecting the position feature in MEAD.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 8. *Summarization results evaluated by ROUGE and human judges*

	System	R-1	R-2	R-SU4	R-L	Human preference
Eng	MEAD	0.192*	0.052*	0.053*	0.183*	–
	LexRank	0.303*	0.093	0.098*	0.278*	–
	SumBasic	0.387	0.090*	**0.153**	0.370	26.9%
	PGN	0.237*	0.047*	0.066*	0.218*	–
	ILP	0.364*	0.123	0.135	0.347*	24.1%
	ILP+MC	**0.392**	**0.130**	0.150	**0.380**	**29.4%**
Stat2015	MEAD	0.225*	0.073*	0.065*	0.218*	–
	LexRank	0.334*	0.154	0.134*	0.326*	–
	SumBasic	**0.457***	**0.193**	**0.223***	0.444*	**30.7%**
	PGN	0.253*	0.096*	0.087*	0.241*	–
	ILP	0.405	0.186	0.165	**0.397**	29.2%*
	ILP+MC	0.401	0.183	0.173	0.391	29.6%
Stat2016	MEAD	0.364*	0.172*	0.140*	0.347*	–
	LexRank	0.397*	0.191	0.168	0.384*	–
	SumBasic	**0.554***	**0.295***	**0.292***	**0.538***	**32.9%**
	PGN	0.277*	0.106*	0.092*	0.269*	–
	ILP	0.482	0.262*	0.215	0.468	29.1%
	ILP+MC	0.457	0.214	0.198	0.441	28.0%
CS2016	MEAD	0.221*	0.056*	0.064*	0.207*	–
	LexRank	0.285*	0.085*	0.091*	0.272*	–
	SumBasic	**0.408**	0.141	0.164	**0.393**	31.5%
	PGN	0.204*	0.057*	0.056*	0.192*	–
	ILP	0.374	0.141	0.137*	0.356*	24.4%*
	ILP+MC	0.398	**0.154**	**0.158**	0.383	**32.7%**
camera	MEAD	**0.475**	**0.207**	**0.211**	0.421	–
	LexRank	0.439	0.181	0.183	0.398	–
	SumBasic	**0.475**	0.168	0.196	**0.439**	23.9%*
	PGN	0.106*	0.050	0.011*	0.101*	–
	ILP	0.457	0.165	0.181	0.427	**36.9%**
	ILP+MC	0.447	0.157	0.176	0.418	32.5%
movie	MEAD	0.394	0.131	0.146	0.341	–
	LexRank	0.434*	**0.147**	**0.186**	0.386	–
	SumBasic	**0.441**	0.098	0.176	0.401	27.6%*
	PGN	0.108*	0.047*	0.011*	0.101*	–
	ILP	0.435	0.091*	0.167	0.397	**36.9%**
	ILP+MC	0.436	0.106	0.169	**0.409**	21.8%
peer	MEAD	0.469	0.242	**0.212**	0.436	–
	LexRank	0.444	0.196	0.170	0.417	–
	SumBasic	0.473	0.154*	0.199	0.441	23.3%
	PGN	0.161	0.098	0.029	0.155	–
	ILP	0.466	0.199	0.183	0.445	**34.4%**
	ILP+MC	**0.491**	**0.261**	0.195	**0.469**	22.2%

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 8. *Continued.*

	System	R-1	R-2	R-SU4	R-L	Human preference
DUC04	MEAD	0.352	0.076	0.117	0.299	–
	LexRank	0.354	0.076	0.118	0.308	–
	SumBasic	0.364*	0.066	0.117	0.326*	24.9%*
	PGN	0.155*	0.025*	0.023*	0.137*	–
	ILP	**0.377***	**0.092***	**0.126***	**0.333***	27.3%*
	ILP+MC	0.342	0.072	0.109	0.308	**31.1%**

Best results are shown in **bold** for each data set. * indicates that the performance difference with ILP+MC is statistically significant ($p < 0.05$) using a two-tailed paired t-test. Underline means that ILP+MC is better than ILP.

Table 9. *A comparison with official participants in DUC 2004, including eight human annotators (1–8) and the top-five offical participants (A–E)*

	R-1	R-2	R-SU4	R-L
Human annotator 1	0.418	0.105	–	0.431
Human annotator 2	0.412	0.086	–	0.409
Human annotator 3	0.410	0.090	–	0.434
Human annotator 4	0.406	0.098	–	0.420
Human annotator 5	0.404	0.107	–	0.425
Human annotator 6	0.393	0.097	–	0.406
Human annotator 7	0.390	0.095	–	0.429
Human annotator 8	0.389	0.090	–	0.406
DUC04 participant A	**0.382**	0.071	–	0.313
DUC04 participant B	0.374	0.071	–	0.310
DUC04 participant C	0.374	0.073	–	0.312
DUC04 participant D	0.374	0.083	–	**0.374**
DUC04 participant E	0.371	0.048	–	0.305
ILP	0.377	**0.092**	0.126	0.333
ILP+MC	0.342	0.072	0.109	0.308

'–' means a metric is not available.

previously unobserved points. In other words, in some sense, no machine learning algorithm is universally any better than any other (Goodfellow *et al.* 2016).

Human evaluation. Because ROUGE cannot thoroughly capture the semantic similarity between system and reference summaries, we further perform a human evaluation. For each task, we present a pair of system outputs in a random order, together with one human summary to five Amazon turkers. If there are multiple human summaries, we will present each human summary and the pair of system outputs to turkers. For student responses, we also present the prompt. An example Human Intelligence Task (HIT) is illustrated in Figure 1.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

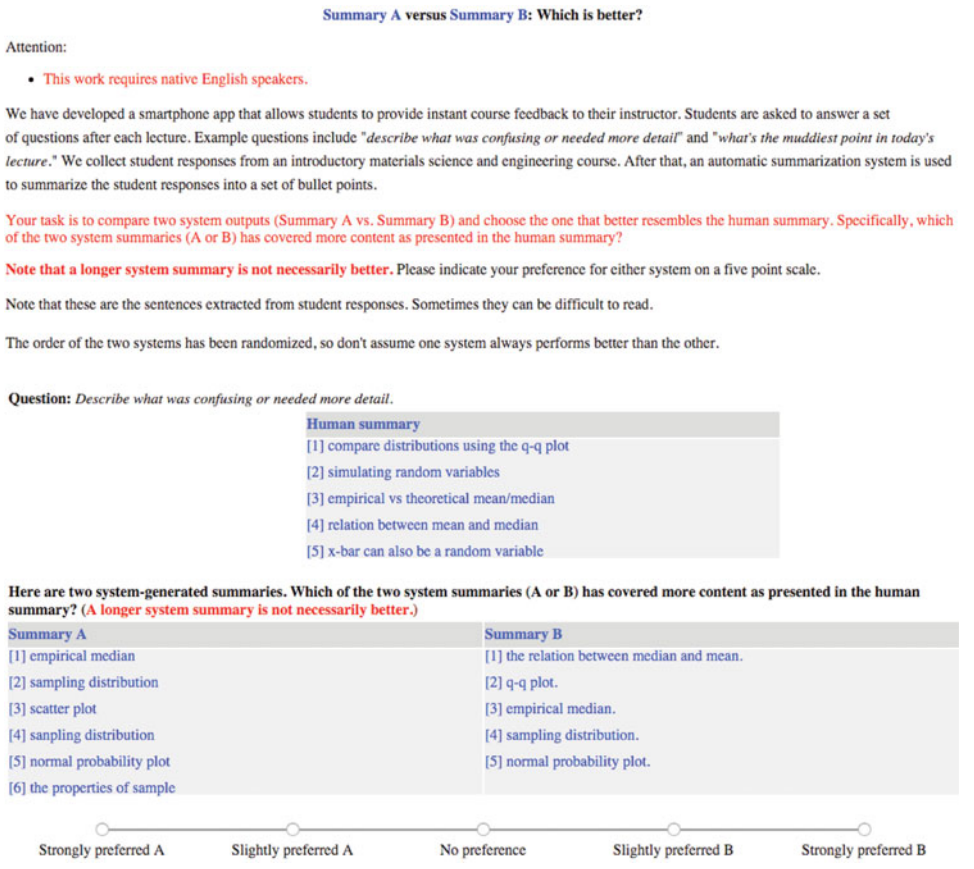


Fig. 1. (Colour online) An example HIT from Stat2015, 'System A' is ILP+MC and 'System B' is SumBasic.

The turkers are asked to indicate their preference for system A or B based on the semantic resemblance to the human summary on a five-Likert scale ('Strongly preferred A', 'Slightly preferred A', 'No preference', 'Slightly preferred B', 'Strongly preferred B'). They are rewarded \$0.04 per task. We use two strategies to control the quality of the human evaluation. First, we require the turkers to have a HIT approval rate of ninety per cent or above. Second, we insert some quality checkpoints by asking the turkers to compare two summaries of same text content but in 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 $\times$ 5 turkers $\times$ (36 tasks $\times$ 1 human summaries for Eng + 44$\times$2 for Stat2015 + 48$\times$2 for Stat2016 + 46$\times$2 for CS2016 + 3$\times$8 for camera + 3$\times$5 for movie + 3$\times$2 for peer + 50 $\times$ 4 for DUC04) = 8,355. The number of tasks for each corpus is shown in Table 1. To elaborate as an example, for Stat2015, there are twenty-two lectures lectures and two prompts for each lecture. Therefore, there are forty-four tasks (22$\times$2) in total. In addition, there are two human summaries for each task. We selected three competitive systems (SumBasic, ILP, and ILP+MC) and therefore we have three system–system pairs (ILP+MC vs.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 10. *Inter-annotator agreement measured by the percentage of individual judgments agreeing with the majority votes*

	ILP+MC versus ILP	ILP+MC versus SumBasic	SumBasic versus ILP
Eng	51.1%	49.4%	50.9%
Stat2015	49.9%* (ILP+MC)	50.0%	51.2%
Stat2016	48.0%	49.2%	51.2%
CS2016	51.3%* (ILP+MC)	51.5%	50.6%* (SumBasic)
Camera	49.2%	47.5%* (ILP+MC)	46.7%* (ILP)
Movie	45.3%	50.7%* (SumBasic)	44.0%* (SumBasic)
Peer	53.3%	43.3%	50.0%* (ILP)
DUC04	48.4%* (ILP)	46.4%* (ILP+MC)	44.0%

* means the human preference to the two systems are significantly different and the system in parenthesis is the winner. Underline means that it is lower than random choices (45.7%).

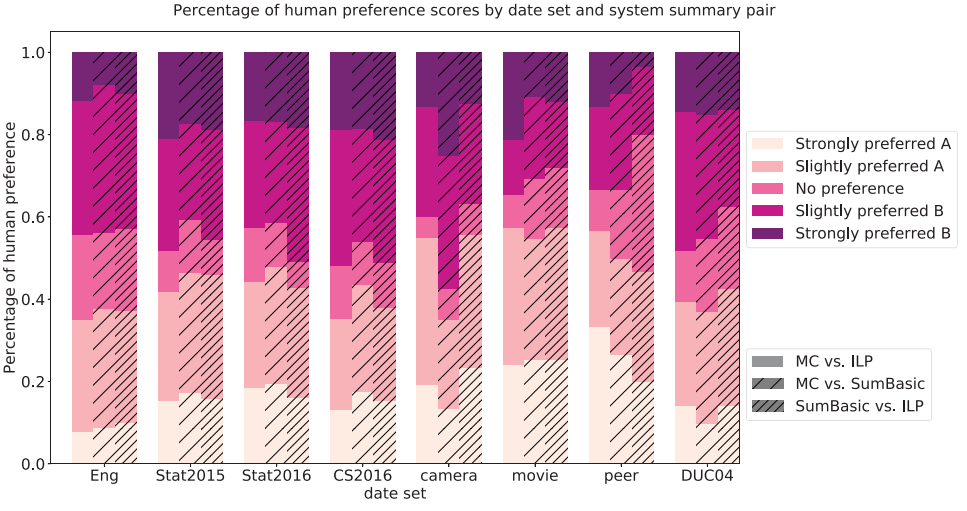


Fig. 2. (Colour online) Distribution of human preference scores.

ILP, ILP+MC vs. SumBasic, and ILP vs. SumBasic) for each task and each human summary. Therefore, we have 44×2×3=264 HITs for Stat2015. Each HIT will be done by five different turkers, resulting in 264×5=1,320 comparisons. In total, 306 unique turkers were recruited[7] and on average 27.3 of HITs were completed by one turker. The distribution of the human preference scores is shown in Figure 2.

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

[7] The turkers are anonymized.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

of Table 8.[8] ILP+MC is preferred significantly[9] more often than ILP on Stat2015, CS2016, and DUC04. There is no significant difference between ILP+MC and SumBasic on student response datasets. Interestingly, a system with better ROUGE scores does not necessarily mean it is more preferred by humans. For example, ILP is preferred more on all three review datasets. Regarding the inter-annotator agreement, we find 48.5 per cent of the individual judgments agree with the majority votes. The agreement scores decomposed by datasets and system pairs are shown in Table 10. Overall, the agreement scores are pretty low, compared to an agreement score achieved by randomly clicking (45.7 per cent).[10] It has several possibilities. The first one is that many turkers did click randomly (39 out of 160 failed our quality checkpoints). Unfortunately, we did not check all the turkers as we inserted the checkpoints randomly. The second possibility is that comparing two system summaries is difficult for humans, and thus it has a low agreement score. Xiong and Litman (2014) also found that it is hard to make humans agree on the choice of summary sentences. A third possibility is that turkers needed to see the raw input sentences which are not shown in a HIT.

An interesting observation is that our approach produces summaries with more sentences, as shown in Table 11. The number of words in the summaries is approximately the same for all methods for a particular corpus, which is constrained by Eq. (16). For camera, movie, and peer reviews, the number of sentences in human summary is 10, and SumBasic and ILP+MC produce more sentences than ILP. It is hard for people to judge which system summaries is closer to a human summary when the summaries are long (216, 242, and 190 words for camera, movie, and peer reviews, respectively). For inter-annotator agreement, 50.3 per cent of judgments agree with the majority votes for student response data sets, 47.6 per cent for reviews, and only 46.3 per cent for news documents. We hypothesize that for these long summaries, people may prefer short system summaries, and for short summaries, people may prefer long system summaries. We leave the examination of this finding to future work.

Table 12 presents example system outputs. This offers an intuitive understanding of our proposed approach.

7 Analysis of influential factors

In this section, we want to investigate the impact of the low-rank approximation process to the ILP framework. Therefore, in the following experiments, we focus on the direct comparison with the ILP and ILP+MC and leave the comparison to other baselines as future work. The proposed method achieved better summarization

[8] The sum of the percentage is not 100% because there are 'no preference' choices.

[9] 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. Similar significant results can be observed if using a 3-point Likert scale ('preferred A', 'no preference', 'preferred B'), except that the difference between ILP and ILP+MC is not significant for Stat2015, but significant for CS2016 and movie.

[10] The random agreement score on a 5-Likert scale can be verified by a simulation experiment.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 11. *Number of sentences in the output summaries*

	Eng	Stat2015	Stat2016	CS2016	Camera	Movie	Peer	DUC04
MEAD	1.6*	1.3*	2.2*	1.1*	3.0*	1.7*	3.3*	2.5*
LexRank	2.8*	2.4*	3.0*	1.9*	7.0	5.3*	6.0*	3.4*
SumBasic	6.0	5.6	5.8*	4.2	14.7	19.7*	12.3*	7.7
ILP	4.8*	3.6*	3.7*	2.6*	14.0*	17.7	12.0*	5.2*
ILP+MC	6.4	5.6	5.3	4.3	17.3	31.3	16.7	7.5

* means it is significantly different to ILP+MC ($p < 0.05$) using a two-tailed paired t-test.

Table 12. *Example reference and system summaries*

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 (ILP+MC)
– 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

performance on Eng, CS2016, movie, and peer than the ILP baseline. Unfortunately, it does not work as expected on two courses for student responses (Stat2015 and Stat2016), review camera and news documents. This leaves the research question when and why the proposed method works better. In order to investigate what are key factors that impact the performance, we would like to perform additional experiments using synthesized datasets.

A variety of attributes that might impact the performance are summarized in Table 13, categorized into two types. The **input** attributes are extracted from the input original documents and the **summaries** attributes are extracted from human summaries and the input documents as well. Here are some important attributes we expect to have a big impact on the performance.

- $M * N$ is the size of the summarization task, represented by the size of the co-occurrence matrix A, as shown in Eqs. (2) and (3). Generally, the bigger

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 13. *Attributes description, extracted from the input and the human reference summaries*

	id	Description
Input	1	• **genre**: belonging to student response/review/news
	2	• **T**: number of tasks
	3	• **au**: number of authors
	4	• **M*N**: size of A
	5	• **M**: number of sentences in total
	6	• **N**: number of bigrams in total
	7	• **M/T**: number of sentences per task
	8	• **N/T**: number of bigrams per task
	9	• **N/M**: number of bigrams per sentence
	10	• **W/T**: number of words per task
	11	• **W/M**: number of words per sentence
	12	• **s**: sparsity ratio, ratio of 0 cells in A per task
	13	• **b=1**: ratio of bigrams appear only once
	14	• **b>1**: ratio of bigrams appear more than once
	15	• **H**: Shannon's diversity index, defined as $H = -\sum_i p_i \ln p_i$, where p_i is the frequency of bigram i divided by total number of bigrams in a task
Summaries	16	• **L**: length of human summaries in number of words
	17	• **hs**: number of human summaries per task
	18	• **r**: compression ratio, length of human summaries compared to length of input documents
	19	• $\alpha_{b>0}$: abstraction ratio, how many of bigrams in **human summaries** appeared in the original documents at least once
	20	• $\alpha_{b=0}$: ratio of bigrams in human summaries that are not in the input
	21	• $\alpha_{b=1}$: ratio of bigrams in human summaries that are in the input only once
	22	• $\alpha_{b>1}$: ratio of bigrams in human summaries that are in the input more than once
	23	• $\beta_{b=1}$: ratio of bigrams in the **input** appear only **once** but selected by human(s)
	24	• $\beta_{b=2}$: ratio of bigrams in the input appear **twice** and selected by human(s)
	25	• $\beta_{b=3}$: ratio of bigrams in the input appear **three times** and selected by human(s)
	26	• $\beta_{b=4}$: ratio of bigrams in the input appear **four times** and selected by human(s)
	27	• $\beta_{b>1}$: ratio of bigrams in the input appear **more than once** and selected by human(s)

the matrix, the more difficult it is to find an optimal solution of low-rank approximation as there are more parameters. Note, A is an $N \times M$ matrix, where N is the number of unique concepts, and M is the number of sentences.

- For the sparsity ratio s, if the matrix is too sparse, there will not be enough information within A to have a good estimate of the completed matrix after

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 14. *Attributes extracted from the input and the human reference summaries*

Id	Name	Eng	Stat2015	Stat2016	CS2016	Camera	Movie	Peer	DUC04
1	Genre	Response	Response	Response	Response	Review	Review	Review	News
2	T	36	44	48	46	3	3	3	50
3	au	37.7	39.3	42.2	22.4	18.0	18	18	10
4	M*N	13.8	10.8	7.2	7.4	0.9	15.7	0.7	2291.7
5	M	1492	1696	1660	1162	255	985	241	11566
6	N	9239	6366	4329	6409	3716	15934	2934	198140
7	M/T	41.4	38.5	34.6	25.3	85.0	328.3	80.3	231.3
8	N/T	256.6	144.7	90.2	139.3	1238.7	5311.3	978.0	3962.8
9	N/M	6.2	3.8	2.6	5.5	14.6	16.2	12.2	17.1
10	W/T	375.4	233.1	149.3	223.1	1927.0	8014.0	1543.7	5171.6
11	W/M	9.1	6.0	4.3	8.8	22.7	24.4	19.2	22.4
12	s	97.2%	96.6%	96.0%	95.4%	98.5%	99.6%	98.5%	99.4%
13	$b = 1$	90.3%	90.1%	87.6%	94.0%	94.7%	92.6%	91.1%	85.5%
14	$b > 1$	9.7%	9.9%	12.4%	6.0%	5.3%	7.4%	8.9%	14.5%
15	H	5.282	4.590	4.007	4.703	6.894	8.314	6.617	7.844
16	L	30	15	13	16	216	242	190	105
17	hs	1	2	2	2	8	5	2	4
18	r	8.8%	7.6%	10.9%	8.3%	13.1%	3.1%	13.5%	2.4%
19	$\alpha_{b>0}$	48.8%	46.5%	56.4%	45.8%	96.7%	97.6%	95.9%	37.0%
20	$\alpha_{b=0}$	51.2%	53.5%	43.6%	54.2%	3.3%	2.4%	4.1%	63.0%
21	$\alpha_{b=1}$	34.1%	18.1%	20.9%	25.6%	84.9%	76.4%	77.1%	15.9%
22	$\alpha_{b>1}$	14.7%	28.4%	35.5%	20.2%	11.8%	21.2%	18.8%	21.1%
23	$\beta_{b=1}$	3.3%	2.7%	4.3%	3.7%	45.8%	11.2%	23.3%	1.7%
24	$\beta_{b=2}$	8.5%	16.5%	28.2%	25.1%	65.3%	20.4%	40.3%	7.2%
25	$\beta_{b=3}$	12.5%	39.0%	58.8%	57.4%	79.3%	31.8%	53.8%	13.7%
26	$\beta_{b=4}$	33.3%	61.1%	76.9%	50.0%	90.9%	42.9%	50.0%	22.1%
27	$\beta_{b>1}$	12.3%	28.0%	45.2%	37.0%	70.0%	27.7%	46.0%	12.0%

The numbers in the row of $M * N$ are divided by 10^6. The description of each attribute is shown in Table 13.

imputation. In contrast, if the matrix is not sparse at all (e.g., all authors use the same term for a concept), there will be no benefit to performing low-rank approximation.

- The Shannon's diversity index H measures the degree of bigram diversity. The more diverse the bigram distribution, the smaller the corresponding Shannon entropy.
- The abstraction ratios $\alpha_{b=0}$, $\alpha_{b=1}$, $\alpha_{b>1}$ capture in what degree annotators use words from the input or use their own.
- $\beta_{b=1}$, $\beta_{b=2}$, $\beta_{b=3}$, $\beta_{b=4}$, $\beta_{b>1}$ intend to capture how humans create the summaries in terms of whether more frequent bigrams are more likely to be selected by humans.

The attributes extracted from the corpora are shown in Table 14. Note, a bigram that appears more often in original documents has a better chance to be included

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

in human summaries as indicated by $\beta_{b=1}$, $\beta_{b=2}$, $\beta_{b=3}$, and $\beta_{b=4}$. This verifies our choice to cut low-frequency bigrams.

According to the ROUGE scores, our method works better on Eng, CS2016, movie, and peer (Table 8). If we group each attribute into two groups, corresponding to whether ILP+MC works better, we do not find significant differences among these attributes. To further understand which factors impact the performance and have more predictive power, we train a binary classification decision tree by treating the four working corpora as positive examples and the remaining four as negative examples.

According to the decision tree model, there is only one decision point in the tree: $\alpha_{b=1}$, the ratio of bigrams in human summaries that are in the input only once. Generally, our proposed method works if $\alpha_{b=1} > 23.2$ per cent, except for camera. When $\alpha_{b=1}$ is low, it means that annotators either adopt concepts that appear multiple times or just use their own. In this case, the frequency-based weighting (i.e., w_i in Eq. (1)) can capture the concepts that appear multiple times. On the other hand, when $\alpha_{b=1}$ is high, it means that a big number of bigrams appeared only once in the input document. In this case, annotators have difficulty selecting a representative one due to the ambiguous choice. Therefore, we hypothesize

- **H2**: The ILP framework benefits more from low-rank approximation when $\alpha_{b=1}$ is higher.

To test the predictive power of this attribute, we want to test it on new datasets. Unfortunately, creating new datasets with gold-standard human summaries is expensive and time-consuming, and the new dataset may not have the desired property within a certain range of $\alpha_{b=1}$. Therefore, we propose to manipulate the ratio and create new datasets using the existing datasets without additional human annotation. $\alpha_{b=1}$ can be represented as follows:

$$\alpha_{b=1} = \frac{\sum_i \sigma_i \cdot \phi_{w_i=1}}{\sum_i \sigma_i} \tag{23}$$

where

$$\sigma_i = \begin{cases} 1 & \text{if bigram } i \text{ appears in the human summary} \\ 0 & else \end{cases}$$

$$\phi_{w_i=1} = \begin{cases} 1 & \text{if } w_i = 1, w_i \text{ is the weight of the bigram i} \\ 0 & else \end{cases}$$

There are two different ways to control the ratio, both involving removing input sentences with certain constraints.

- To increase this ratio, we remove sentences with bigrams that appear multiple times so that there will be more bigrams that appear once (i.e., increase $\sigma_i \cdot \phi_{w_i=1}$) and thus increase the numerator. For example, if a bigram in a human summary appears in two input sentences (e.g., S1 and S2), we can randomly remove one of them (either S1 or S2) to make the bigram appear only once in the input. Note that we keep sentences that have bigrams

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

appearing multiple times and a bigram appearing only once as well, so that we guarantee that all the input sentences with a unique bigram in human summaries are kept and removing other sentences can only increase the ratio.
- To decrease this ratio, we remove the sentences with bigrams that appear only once in order to decrease the numerator. This will reduce the bigram frequency w_i from 1 to 0. Similarly, we keep sentences that contain bigrams appearing multiple times so that removing sentences will not increase the ratio.

In this way, we obtained different levels of $\alpha_{b=1}$ by deleting sentences. The ROUGE scores on the synthesized corpus are shown in Table 15.

Our hypothesis **H2** is partially valid. When increasing the ratio, ILP+MC has a relative advantage gain over ILP. For example, for Stat2015, ILP+MC is not significantly worse than ILP any more when increasing the ratio from 11.9 to 18.1. For camera, ILP+MC becomes better than ILP when increasing the ratio from 84.9 to 85.8. For Stat2016, CS2016, Eng, more improvements or significant improvements can be found for ILP+MC compared to ILP when increasing the ratio. However, for movie and peer review, ILP+MC is worse than ILP when increasing the ratio.

We have investigated a number of attributes that might impact the performance of our proposed method. Unfortunately, we do not have a conclusive answer when our method works better. However, we would like to share some thoughts about it.

First, our proposed method works better on two student responses courses (Eng and CS2016), but not the other two (Stat2015 and Stat2016). An important factor we ignored is that the students from the other two courses are not native English speakers, resulting in significantly shorter responses ($4.3 < 6.0 < 8.8$, 9.1, $p < 0.01$, Table 14, the row with id=11). With shorter sentences, there will be less context to leverage the low-rank approximation.

Second, our proposed method works better on movie and peer reviews, but not camera reviews. As pointed out by Xiong (2015), both movie reviews and peer reviews are potentially more complicated than the camera reviews, as the review content consists of both the reviewer's evaluations of the subject (e.g., a movie or paper) and the reviewer's references of the subject, where the subject itself is full of content (e.g., movie plot and papers). In contrast, such references in product reviews are usually the mentions of product components or properties, which have limited variations. This characteristic makes review summarization more challenging in these two domains.

8 Conclusion

We made the first effort to summarize student feedback using an ILP framework with a low-rank matrix approximation, and applied it to different types of datasets including news articles, product, and peer reviews. Our approach allows sentences to share cooccurrence statistics and alleviates sparsity issue. Our experiments showed that the proposed approach performs better against a range of baselines on the student response Eng and CS2016 on ROUGE scores, but not other courses.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 15. *ROUGE scores on synthesized corpora*

	$\alpha_{b=1}$	System	R-1	R-2	R-SU4	R-L
	26.5	ILP	0.341	**0.112**	0.121	0.329
		ILP+MC	**0.378**$^{+}$	**0.112**	**0.137**	**0.366**$^{+}$
Eng	34.1	ILP	0.364	0.123	0.135	0.347
		ILP+MC	**0.392**$^{+}$	**0.130**	**0.150**	**0.380**$^{+}$
	36.0	ILP	0.358	0.119	0.124	0.345
		ILP+MC	**0.397**$^{+}$	**0.126**	**0.155**	**0.379**
	11.9	ILP	0.401	0.183	0.160	0.393
		ILP+MC	0.379	0.167	0.158	0.369
Stat2015	18.1	ILP	0.405	0.186	0.165	0.397
		ILP+MC	0.401	0.183	**0.173**	0.391
	21.0	ILP	0.394	0.172	0.160	0.384
		ILP+MC	0.350^{-}	0.144^{-}	0.140	0.341^{-}
	13.2	ILP	0.467	0.252	0.202	0.454
		ILP+MC	0.461	0.209^{-}	0.200	0.443
Stat2016	20.9	ILP	0.482	0.262	0.215	0.468
		ILP+MC	0.457	0.214^{-}	0.198	0.441
	23.7	ILP	0.455	0.244	0.198	0.441
		ILP+MC	**0.476**	0.210^{-}	**0.203**	**0.460**
	11.0	ILP	0.362	0.138	0.134	0.346
		ILP+MC	**0.376**	0.135	**0.151**	**0.360**
CS2016	25.6	ILP	0.374	0.141	0.137	0.356
		ILP+MC	**0.398**	**0.154**	**0.158**$^{+}$	**0.383**$^{+}$
	34.2	ILP	0.296	0.091	0.087	0.281
		ILP+MC	**0.318**	0.085	**0.100**	**0.306**
	78.7	ILP	0.453	0.166	0.182	0.426
		ILP+MC	0.437	0.147	0.172	0.407
camera	84.9	ILP	0.457	0.165	0.181	0.427
		ILP+MC	0.447	0.157	0.176	0.418
	85.8	ILP	0.452	0.156	0.178	0.422
		ILP+MC	**0.454**	**0.159**	**0.183**	**0.427**
	71.9	ILP	0.439	0.107	0.172	0.397
		ILP+MC	0.417	0.103	0.159	0.392
movie	76.4	ILP	0.435	0.091	0.167	0.397
		ILP+MC	**0.436**	**0.106**$^{+}$	**0.169**	**0.409**
	76.8	ILP	0.435	0.109	0.170	0.387
		ILP+MC	0.411	0.102	0.155	0.383
	71.3	ILP	0.467	0.206	0.188	0.448
		ILP+MC	0.442	0.193	0.158	0.421
peer	77.1	ILP	0.466	0.199	0.183	0.445
		ILP+MC	**0.491**	**0.261**	**0.195**	**0.469**
	78.7	ILP	0.488	0.242	0.199	0.475
		ILP+MC	0.447	0.183	0.162	0.425

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Table 15. *Continued.*

	$\alpha_{b=1}$	System	R-1	R-2	R-SU4	R-L
	13.9	ILP	0.376	0.092	0.124	0.332
		ILP+MC	0.349^{-}	0.074^{-}	0.113^{-}	0.314^{-}
DUC04	15.9	ILP	0.377	0.092	0.126	0.333
		ILP+MC	0.342^{-}	0.072^{-}	0.109^{-}	0.308^{-}
	16.5	ILP	0.375	0.093	0.123	0.332
		ILP+MC	0.349^{-}	0.074^{-}	0.113^{-}	0.314^{-}

Bold scores indicate our approach ILP+MC is better than ILP. $^{+}$ and $^{-}$ mean a score is significantly better and worse, respectively ($p < 0.05$) using a two-tailed paired t-test.

ROUGE is often adopted in research papers to evaluate the quality of summarization because it is fast and is correlated well to human evaluation (Lin 2004; Graham 2015). However, it is also criticized that ROUGE cannot thoroughly capture the semantic similarity between system and reference summaries. Different alternatives have been proposed to enhance ROUGE. For example, Rankel (2016) proposed to use content-oriented features in conjunction with linguistic features. Similarly, Cohan and Goharian (2016) proposed to use content relevance. At the same time, many researchers supplement ROUGE with a manual evaluation. This is why we conduct evaluations using both ROUGE and human evaluation in this work.

However, we found that a system with better ROUGE scores does not necessarily mean it is more preferred by humans (Section 6.2). For example, ILP is preferred more on all three review datasets even if it got lower ROUGE scores than the other systems. It coincides with the fact that the ILP generated shorter summaries in terms of the number of sentences than the other two systems (Table 11).

We also investigated a variety of attributes that might impact the performance on a range of datasets. Unfortunately, we did not have a conclusive answer when our method will work better.

In the future, we would like to conduct a large-scale intrinsic evaluation to examine whether the low-rank matrix approximation captures similar bigrams or not and want to investigate more attributes, such as new metrics for diversity. We would like to explore the opportunities by combing a vector sentence representation learned by a neural network and the ILP framework.

References

Almeida, M., and Martins, A. 2013. Fast and robust compressive summarization with dual decomposition and multi-task learning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria: Association for Computational Linguistics, pp. 196–206.

Barzilay, R., McKeown, K. R., and Elhadad, M. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Berg-Kirkpatrick, T., Gillick, D., and Klein, D. 2011. Jointly learning to extract and compress. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Boud, D., Keogh, R., Walker, D., *et al.* (2013). *Reflection: Turning Experience into Learning*. Routledge, UK.

Boudin, F., Mougard, H., and Favre, B. 2015. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1914–1918.

Cao, Z., Wei, F., Li, W., and Li, S. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Carbonell, J., and Goldstein, J. 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)*, New York, NY, USA: ACM, pp. 335–336.

Celikyilmaz, A., Bosselut, A., He, X., and Choi, Y. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, pp. 1662–1675.

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., and Jiang, H. 2016. Distraction-based neural networks for document summarization. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*.

Cho, K. (2008). Machine classification of peer comments in physics. In *Educational Data Mining 2008 – 1st International Conference on Educational Data Mining, Proceedings*, pp. 192–196.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 615–621.

Cohan, A., and Goharian, N. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, pp. 390–400.

Cohan, A., and Goharian, N. 2016. Revisiting summarization evaluation for scientific articles. In N. C. C. Chair, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France: European Language Resources Association (ELRA).

Cohan, A., and Goharian, N. 2017. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, pp. 1–17.

Conroy, J., and Davis, S. 2015. Vector space models for scientific document summarization. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, Denver, Colorado: Association for Computational Linguistics, pp. 186–191.

Conroy, J., Davis, S. T., Kubina, J., Liu, Y.-K., O'Leary, D. P., and Schlesinger, J. D. 2013. Multilingual summarization: Dimensionality reduction and a step towards optimal term coverage. In *Proceedings of the MultiLing Workshop on Multilingual Multi-Document Summarization*, Sofia, Bulgaria: Association for Computational Linguistics, pp. 55–63.

Dang, H. T., and Owczarzak, K. 2008. Overview of the TAC 2008 update summarization task. In *Proceedings of Text Analysis Conference (TAC)*.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Durrett, G., Berg-Kirkpatrick, T., and Klein, D. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, pp. 1998–2008.

Erkan, G., and Radev, D. R. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* **22**(1): 457–479.

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

Fan, X., Luo, W., Menekse, M., Litman, D., and Wang, J. 2017. Scaling reflection prompts in large classrooms via mobile interfaces and natural language processing. In *Proceedings of 22nd ACM Conference on Intelligent User Interfaces (IUI 2017)*.

Galanis, D., Lampouras, G., and Androutsopoulos, I. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of Computational Linguistics (COLING)*.

Gerani, S., Mehdad, Y., Carenini, G., Ng, R. T., and Nejat, B. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gillick, D., and Favre, B. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*. Association for Computational Linguistics, pp. 10–18.

Gillick, D., Favre, B., and Hakkani-Tür, D. 2008. The ICSI summarization system at TAC 2008. In *Proceedings of Text Analysis Conference (TAC)*.

Gkatzia, D., Hastie, H., Janarthanam, S., and Lemon, O. 2013. Generating student feedback from time-series data using reinforcement learning. In *Proceedings of European Workshop on Natural Language Generation (ENLG)*.

Goldberg, Y., and Levy, O. 2014. word2vec explained: Deriving Mikolov *et al.*'s negative-sampling word-embedding method. preprint arXiv:1402.3722.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`

Graham, Y. 2015. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, pp. 128–137.

Grusky, M., Naaman, M., and Artzi, Y. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), pp. 708–719. Association for Computational Linguistics, New Orleans, Louisiana.

Guo, H., Pasunuru, R., and Bansal, M. 2018. Soft, layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Harwood, W. S. 1996. The one minute paper: A communication tool for large lecture classes. *Journal of Chemical Education* **73**(3): 229.

He, Z., Chen, C., Bu, J., Wang, C., Zhang, L., Cai, D., and He, X. 2012. Document summarization based on data reconstruction. In *Proceedings of AAAI*.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pp. 1693–1701.

Hong, K., Conroy, J., Favre, B., Kulesza, A., Lin, H., and Nenkova, A. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno,

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

J. Odijk, and S. Piperidis (eds.), *Proceedings of LREC*, Reykjavik, Iceland: ACL Anthology Identifier: L14-1070, pp. 1608–1616.

Jindal, N., and Liu, B. 2008. Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Data Mining*. ACM, pp. 219–230.

Jing, H., and McKeown, K. 1999. The decomposition of human-written summary sentences. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., and Okumura, M. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.

Kulesza, A., and Taskar, B. (2012). Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning* **5**(2–3): 123–286.

Lee, J.-H., Park, S., Ahn, C.-M., and Kim, D. 2009. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management* **45**(1): 20–34.

Li, C., Liu, F., Weng, F., and Liu, Y. 2013. Document summarization via guided sentence compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: Association for Computational Linguistics, pp. 490–500.

Li, C., Liu, Y., Liu, F., Zhao, L., and Weng, F. 2014. Improving multi-document summarization by sentence compression based on expanded constituent parse tree. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, Doha, Qatar.

Li, C., Liu, Y., and Zhao, L. 2015. Using external resources and joint learning for bigram weighting in ILP-based multi-document summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado: Association for Computational Linguistics, pp. 778–787.

Liao, K., Lebanoff, L., and Liu, F. 2018. Abstract meaning representation for multi-document summarization. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, vol. 8. Barcelona, Spain.

Lin, H., and Bilmes, J. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 912–920.

Liu, F., and Liu, Y. 2013. Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression. *IEEE Transactions on Audio, Speech and Language Processing* **21**(7): 1469–1480.

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

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

Luo, W., Liu, F., and Litman, D. 2016a. An improved phrase-based approach to annotating and summarizing student course responses. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING-2016)*, Osaka, Japan. The COLING 2016 Organizing Committee, pp. 53–63.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Luo, W., Liu, F., Liu, Z., and Litman, D. 2016b. Automatic summarization of student course feedback. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, pp. 80–85.

Martins, A., and Smith, N. A. 2009. Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for NLP*, Boulder, Colorado, pp. 1–9.

Mazumder, R., Hastie, T., and Tibshirani, R. 2010. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research* **11**(Aug): 2287–2322.

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

Moon, S., Potdar, S., and Martin, L. 2014. Identifying student leaders from mooc discussion forums through language influence. In *Proceedings of EMNLP Workshop on Analysis of Large Scale Social Interaction in MOOCs*.

Mosteller, F. 1989. The 'muddiest point in the lecture' as a feedback device. *On Teaching and Learning: The Journal of the Harvard-Danforth Center* **3**: 10–21.

Nallapati, R., Xiang, B., and Zhou, B. 2016. Sequence-to-sequence RNNs for text summarization. *CoRR*, abs/1602.06023.

Narayan, S., Cohen, S. B., and Lapata, M. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, pp. 1747–1759.

Nenkova, A., McKeown, K., *et al.* 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval* **5**(2–3): 103–233.

Paul, O., and James, Y. 2004. An introduction to DUC-2004. In *Proceedings of the 4th Document Understanding Conference (DUC-2004)*.

Paulus, R., Xiong, C., and Socher, R. 2017. A deep reinforced model for abstractive summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qazvinian, V., Radev, D. R., Mohammad, S. M., Dorr, B., Zajic, D., Whidby, M., and Moon, T. 2013. Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research* **46**: 165–201.

Qian, X., and Liu, Y. 2013. Fast joint compression and summarization via graph cuts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: Association for Computational Linguistics, pp. 1492–1502.

Radev, D. R., Jing, H., Styś, M., and Tam, D. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management* **40**(6): 919–938.

Rankel, P. A. 2016. *Statistical Analysis of Text Summarization Evaluation*. PhD thesis, University of Maryland, College Park.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ren, P., Wei, F., Chen, Z., Ma, J., and Zhou, M. 2016. A redundancy-aware sentence regression framework for extractive summarization. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING-2016)*, Osaka, Japan. The COLING 2016 Organizing Committee, pp. 33–43.

Rose, C. P., and Siemens, G. 2014. Shared task on prediction of dropout over time in massively open online courses. In *Proceedings of EMNLP Workshop on Analysis of Large Scale Social Interaction in MOOCs*.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Rus, V., Lintean, M. C., Banjade, R., Niraula, N. B., and Stefanescu, D. 2013. SEMILAR: The semantic similarity toolkit. In *ACL (Conference System Demonstrations)*, pp. 163–168.

Rush, A. M., Chopra, S., and Weston, J. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, pp. 379–389.

See, A., Liu, P. J., and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, pp. 1073–1083.

Song, K., Zhao, L., and Liu, F. 2018. Structure-infused copy mechanisms for abstractive summarization. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA.

Suzuki, J., and Nagata, M. 2017. Cutting-off redundant repeating generations for neural abstractive summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Takase, S., Suzuki, J., Okazaki, N., Hirao, T., and Nagata, M. 2016. Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, pp. 1054–1059.

Tan, J., Wan, X., and Xiao, J. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tarnpradab, S., Liu, F., and Hua, K. A. 2017. Toward extractive summarization of online forum discussions via hierarchical attention networks. In *Proceedings of the 30th Florida Artificial Intelligence Research Society Conference (FLAIRS)*, Marco Island, Florida, pp. 288–292.

Teufel, S., and Moens, M. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics* **28**(4): 409–445.

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

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. 2007. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management* **43**(6): 1606–1618.

Wang, D., Li, T., Zhu, S., and Ding, C. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 307–314.

Wang, W. Y., Mehdad, Y., Radev, D. R., and Stent, A. 2016a. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 58–68.

Wang, X., Nishino, M., Hirao, T., Sudoh, K., and Nagata, M. 2016b. Exploring text links for coherent multi-document summarization. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING-2016)*, Osaka, Japan. The COLING 2016 Organizing Committee, pp. 213–223.

Wen, M., Yang, D., and Rose, C. P. 2014a. Linguistic reflections of student engagement in massive open online courses. In *Proceedings of International Conference on Web and Social Media (ICWSM)*.

Wen, M., Yang, D., and Rosé, C. P. 2014b. Sentiment analysis in MOOC discussion forums: What does it tell us. In *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press

Wilson, R. C. 1986. Improving faculty teaching: Effective use of student evaluations and consultants. *Journal of Higher Education* **57**(2): 196–211.

Wolpert, D. H. 1996. The lack of a priori distinctions between learning algorithms. *Neural computation* **8**(7): 1341–1390.

Xiong, W. 2015. *Helpfulness Guided Review Summarization*. PhD thesis, University of Pittsburgh.

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

Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan, K., and Radev, D. 2017. Graph-based neural multi-document summarization. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, Vancouver, Canada.

Zhou, Q., Yang, N., Wei, F., and Zhou, M. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Appendix A. Results without removing low-frequency bigrams

Table A1. *Summarization results without removing low-frequency bigrams*

	System	R-1	R-2	R-SU4	R-L
Eng	ILP	0.351	0.108	0.126	0.337
	ILP+MC	**0.355**	**0.111**	**0.130**	**0.347**
Stat2015	ILP	0.403	0.205	0.163	0.395
	ILP+MC	**0.448**$^{+}$	**0.248**$^{+}$	**0.214**$^{+}$	**0.433**
Stat2016	ILP	**0.470**	**0.249**	**0.207**	**0.455**
	ILP+MC	0.423^{-}	0.209^{-}	0.170^{-}	0.410^{-}
CS	ILP	0.374	0.138	0.139	0.354
	ILP+MC	**0.380**	**0.144**	**0.146**	**0.362**
peer	ILP	**0.470**	**0.228**	**0.176**	**0.449**
	ILP+MC	0.452	0.175	0.165	0.428
camera	ILP	**0.456**	**0.168**	**0.179**	**0.426**
	ILP+MC	0.440	0.146	0.168	0.410
movie	ILP	0.426	**0.109**	0.163	0.382
	ILP+MC	**0.430**	0.102	**0.165**	**0.397**
DUC04	ILP	**0.377**	**0.092**	**0.126**	**0.333**
	ILP+MC	0.337^{-}	0.071^{-}	0.106^{-}	0.305^{-}

That is, all bigrams are used in the matrix approximation process. Compared to Table 8, by using the cut-off technique, both ILP and ILP+MC get better.

https://doi.org/10.1017/S1351324918000323 Published online by Cambridge University Press