How Good Is Popularity? Summary Grading in Crowdsourcing

Haiying Li
Rutgers University
10 Seminary Place
New Brunswick, NJ 08901
1-848-932-0868
haiying.li@gse.rutgers.edu

Zhiqiang Cai
University of Memphis
365 Innovation Dr.
Memphis, TN 38152
1-901-678-2364
zcai@memphis.edu

Arthur C. Graesser
University of Memphis
365 Innovation Dr.
Memphis, TN 38152
1-901-678-2364
grasser@memphis.edu

ABSTRACT

In this paper, we applied the crowdsourcing approach to develop an automated popularity summary scoring, called wild summaries. In contrast, the golden standard summaries generated by one or more experts are called expert summaries. The innovation of our study is to compute LSA (Latent Semantic Analysis) similarities between target summary and wild summaries rather than expert summaries. We called this method CLSAS, i.e., crowdsourcing-based LSA similarity. We evaluated CLSAS by comparing it with other approaches, Coh-Metrix language and discourse features and LIWC psychometric word measures. Results showed that CLSAS alone could explain 19% of human summary score, which was equivalent to the variance explained by dozens of language and discourse features and/or the word features. Results also showed that adding language and/or word features to CLSAS increased small additional correlations. Findings imply that crowdsourcing-based LSA similarity approach is a promising method for automated summary assessment.

Keywords

Summary grading, Crowdsourcing, LSA Similarity, Coh-Metrix, LIWC

1. INTRODUCTION

The use of the summarization strategy enables to improve reading comprehension and production of expository texts for both L1 learners [1] and L2 learners [2]. Summarizing involves reading processes and reproducing processes. Reading process requires the learners to identify the main ideas and distinguish the important points from the unimportant points. Reproducing process requires the learners to restate the important ideas in a coherent, precise and accurate manner in their own words [3]. Learners' summarizing skill depends on the ability to construct a coherent mental model of the text, which is aligned with text discourse [4]. This ability consists of three knowledge components: rhetorical text structures and genres, propositional text content, and a coherent mental model for a variety of genres [4], which are important for reading comprehension [5]. Summarization strategy is an effective instructional strategy [6] to help students improve these abilities [7] and summary writing is therefore considered as a good measure of reading comprehension at a deep level.

Grading summaries are time-consuming and costly for teachers, so it is impossible for teachers to provide a real-time and instant summary score, let alone provide the instant feedback on the quality of summaries. Researchers thereby have developed the automated summary assessments with the techniques of natural language processing and machine learning [4,8]. These assessments are not practical for teachers because they require model building based on human expert summaries as the reference summaries and a large amount of human summary grading. Thus, model rebuilding is time-consuming and costly for teachers. Each time teachers need to repeat such complex steps as expert-written summaries as reference, human-scored summary as the training set, model training, and model evaluation. As summary writing is a weekly assignment for middle school and high school students, summary grading will be a common task for teachers. The present automated summary assessments will not reduce but increase the teachers' workload. These methods are impractical for teachers to use. Teachers need a more efficient and effective summary assessment with least efforts.

In this paper, we applied the crowdsourcing approach to develop an automated "popularity" summary scoring. Crowdsourcing enables a diverse and a large amount of population to generate abundant summaries, which are called "popularized summaries" or "wild summaries." In contrast, the golden standard summaries generated by one or more experts are called "expert summaries." The innovation of our study is to compute LSA (Latent Semantic Analysis) similarities between the target summary and the wild summaries instead of expert summaries. We called it CLSAS, namely, crowdsourcing-based LSA similarity. We proposed CLSAS was a robust measure for summary grading.

This study makes innovative contributions to the automated summary assessment for three reasons. First, it is efficient and effective, because the model was built based on one feature rather than dozens of features. Second, it is unnecessary for human experts to generate the golden summaries on each quality level. The model was built based on the wild summaries generated by all of the summary writers. Third, it is unnecessary for human experts to manually grade summaries for the model training.

The next section briefly reviews research on automated summary assessment, crowdsourcing approach, and three advanced text analysis tools, LSA similarity [9], Coh-Metrix [10] and LIWC (Linguistic Inquiry and Word Count) [11].

1.1 Automated Summary Assessment

Techniques of natural language processing and machine learning have been used to develop the automated summary assessment [4,8]. Diverse features used in the assessment range from semantic features measured by LSA [8] to language features exacted by BLEU (Bilingual Evaluation Understudy) [4], ROUGE (Recall-

Oriented Understudy for Gisting Evaluation) [12], TERp (Translation Error Rate Plus) [4], and N-gram [12]. Some features were used to detect plagiarism in summary (e.g., N-gram [4]), assess coherence of the summary (e.g., LSA [8] and N-gram [12]), evaluate content unit (e.g., unigram overlap [8]), or examine the length of summary [4]. These assessments were proved to robustly predict human summary grading [4,8] but had the following limitations.

First, all of these assessments need reference summaries that are generated by one or more human experts [4,8]. The reference summaries have different qualities, ranging from good to poor on multiple-point scales [4]. The student's summary is graded by comparing with the reference summaries. The similarities could be computed by similarities of LSA [8], a lexical and phrasal overlap (e.g., ROUGE) [8], N-gram overlap (e.g., BLEU) [4,8], summary length [4], or token count [4]. Second, the sufficient amount of human-graded summaries at each quality level is required to build the model for the supervised learning. Third, different language and discourse features and algorithms are tested in order to build a better fit model. As these assessments are not content independent, these three cycles are repeated if summaries' source text changes. These tasks definitely increase extra workload for teachers, so it is hard and impractical to spread these approaches. It is necessary to develop a summary assessment without expert reference summaries, human grading, and model rebuilding for a new source text. This study aims to explore a real-time and efficient summary assessment that requires the least efforts so that teachers can easily use it by themselves.

1.2 Crowdsourcing

Crowdsourcing refers to a process that mobilizes a huge amount of population (called crowd workers) to accomplish the complex, collaborative, and sustainable tasks on demand and at large scale, especially from an online community rather than traditional employees or suppliers [13]. Crowd workers can either be volunteers for collective projects such as Wikipedia or paid via platform such as Amazon's Mechanical Turk, one popular crowdsourcing platform [13]. Crowdsourcing is frequently used to generate ideas and break down creative tasks into smaller pieces [13-17]. The application of crowdsourcing is an emerging approach in research. For example, some researchers asked crowd workers to create or retrieve content for new stories [16,17], to generate a story [14] or summaries of social media events [15]. This collaborative work provides an author diverse ideas or contents quickly [13-17].

1.3 LSA Similarity

LSA [18] is a mathematical and statistical technique that represents knowledge about words, sentences, paragraphs, and documents on the basis of a large corpus of texts. LSA reduces a large corpus of texts to 100–300 dimensions using singular value decomposition technique. The conceptual similarity between two texts is computed as the geometric cosine between the vectors representing two texts. The cosine value varies from -1 to 1 [18,19], with the higher score representing higher similarity.

LSA is used to assess coherence in Coh-Metrix [10] and quality of essays [8, 20-22]. In addition, LSA has been utilized in the intelligent tutoring system (ITS) to assess the constructed response or the open response, such as AutoTutor [19]. These assessment systems for essay, summary, or open response requires expert reference summaries and human-graded summaries generated by human experts. Few studies do not use expert summaries as reference. Summarization in machine translation develops a fully automated approach to evaluate ranking systems that requires no expert summaries [8]. However, it requires a large amount of content annotations and is restricted to the ranking system, which it is not appropriate for teachers to use for summary grading. Cai et al. [9] explored the LSA similarity model without the golden standard reference for the open response assessment. Instead, the reference was all the responses written by students except the target response. We borrowed this approach in this study and use the learners' summaries as the reference summaries.

1.4 Coh-Metrix

Coh-Metrix (cohmetrix.com) is a computer-based tool that automates many language- and text-processing mechanisms over hundreds of measures of cohesion, language, and readability [10]. Coh-Metrix is developed based on a multilevel theoretical framework [23]. This framework specifies six theoretical levels: words, syntax, explicit textbase (e.g., explicit propositions, referential cohesion), situation model (also called mental model), discourse genre and rhetorical structure (the type of discourse and its composition), and the pragmatic communication level. The first five of these six levels have metrics captured in the Coh-Metrix automated text analysis tool [10].

The current version of Coh-Metrix [10] extracts 110 measures, which are categorized into genre (narrative versus informational), LSA space (e.g., text cohesion), word information (e.g., familiarity, concreteness, imageability, meaningfulness, age of acquisition), word frequency, part of speech, density score (e.g., density of pronouns), logic operators (e.g., *if-then*), connectives (e.g., *therefore*), type/token ratio, polysemy and hypernym, syntactic complexity (e.g., noun phrase density), readability (e.g., Flesch-Kincaid grade level), co-reference cohesion (e.g., noun overlap, argument overlap), along with five primary components extracted based on these features (e.g., narrativity, word concreteness, syntactic simplicity, referential cohesion, and deep cohesion).

1.5 LIWC

LIWC (Linguistic and Inquiry Word Count) [11] computes the percentage of words in a text that fit into the linguistic or psychological categories. The 2015 LIWC dictionary contains 6,400 words, word stems, and select emoticons. It generates 93 measures that are categorized into the following categories: word count, summary language variables (e.g., analytical thinking, authentic, emotional tone), linguistic dimensions (e.g., functional words, pronouns, conjunctions), other grammar (e.g., common verbs, interrogatives), psychological processes (e.g., affective, social, cognitive, informal language). The word count function of LIWC attempts to match each word in a given text to a word in the various categories.

The LIWC categories have been confirmed as valid and reliable markers of a variety of psychologically meaningful constructs [11]. The different categories of words would be expected to predict psychological dimensions. For example, negative emotion words would be diagnostic of gloomy texts. The function words (particularly pronouns) are diagnostic of social status, personality, and various psychological states. Differences in function word use can be reflected by gender, age, and social class. LIWC is used to measure the formal versus informal language formality [24,25].

This paper combined the crowdsourcing approach with the LSA similarity to assess summaries. This approach was evaluated by comparing the Coh-Metrix language and discourse features and

the LIWC word features with the human-graded summary scores as the criteria. Specially, seven models were trained and compared their predictability for the human summary scores: (1) CLSAS, (2) Coh-Metrix language feaures (94), (3) LIWC word features (93), (4) Coh-Metrix + LIWC, (5) CLSAS + Coh-Metrix, (6) CLSAS + LIWC, and (7) CLSAS + Coh-Metrix + LIWC. It is necessary to clarify that the human-graded summary scores were only used to evaluate but not build the model. We hypothesize that crowdsourcing-based LSA similarity is an efficient, effective, and reliable measure for summary grading for the following two reasons. First, LSA is a most robust feature for semantic meaning [11] than the language and word features. Second, the wild summaries as reference maximally represent diversity of students' summaries as compared with expert summaries.

2. METHOD

2.1 Participants

Crowd workers (N = 201) volunteered for 3-hour monetary compensation ($30) on Amazon Mechanical Turk (AMT), a trusted and commonly used data collection service [21]. The basic requirement for participation is that they have the goal to improve English summary writing. Participants were required to complete writing 8 summaries, but only 1,481 summaries were collected due to the technical issues. 71% participants were Asian, 16% white or Caucasian, 7% African American, 5% Hispanic, 2% other. Their average age was 33.50 (SD = 8.79), 57% were male, and 81% with bachelor degree or above.

2.2 Materials

Participants read 8 expository texts with different topics and text difficulties in the AutoTutor CSAL. CSAL is an intelligent tutoring system that teaches adult learners the summarization strategies in order to improve their reading comprehension [19]. Participants were required to write a summary with 50-100 words for each text. Four texts are on comparison-contrast text structure and another four on cause-effect text structure (See Table 1). The text difficulty was measured with the Coh-Metrix formality (z-score) at the multiple textural levels and Flesch-Kincaid grade level, sensitive to word length and sentence length [24]. These 8 texts were formal and above grade 8 to early college grades [24]. The balanced Latin-square designs were applied to control for order effects in terms of text difficulty, topics and text structures.

2.3 Summary Grading

The summaries were graded based on four components: topic sentence, content, grammar and mechanics, and signal words. Table 2 lists the detailed descriptions for three scales of each component, from 0 (minimum) to 2 (maximum) points. Thus, the total score ranged from 0 to 8. Four English native researchers graded summaries, 1 male and 3 female. There were three rounds of training for summary grading and after each grading, and then the disagreements were discussed. Before grading, they got familiar rubrics and then they started the three-round grading with one per week. Each round included 32 randomly-selected summaries (4 from each text and 8 texts in total). Inter-rater reliability was assessed by the intraclass correlation coefficient with a two-way random model and absolute agreement type. The average inter-rater reliability reached the threshold: Cronbach's a = .82, intraclass correlation coefficient = .80. As the average of reliabilities for three training sets were high, each grader graded summaries for two texts in the same text structure.

Table 1. Source Texts and the Number of Summaries (N).

Structure	Topics	Formality	FKGL	Words	N
Comparison	Butterfly and Moth	.12	8.6	255	183
	Hurricane	.20	9.4	222	185
	Walking and Running	.18	8.9	399	187
	Kobe and Jordan	.14	9.2	299	187
Causation	Floods	. 47	9.2	230	186
	Job Market	.62	10.9	240	181
	Effects of Exercising	.28	9.1	195	189
	Diabetes	.64	11.7	241	182

Table 2. Rubrics for Scoring Summary

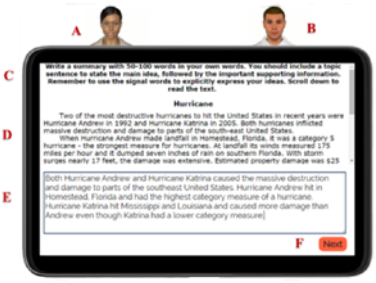
Categories	2 points	1 points	0 point
Topic Sentence	A clear topic sentence that states the main idea.	A topic sentence that touches upon the main idea.	The summary does not state the main idea.
Content	Major details stated economically and arranged in a logical order. No minor or unimportant details or reflections.	Some but not all major details stated and not necessarily in a logical order. Some minor or unimportant details or reflections.	Few major details stated and not necessarily in a logical order. Many minor or unimportant details or reflections.
Mechanics and Grammar	Few or no errors in mechanics, usage, grammar or spelling.	Some errors in mechanics, usage, grammar or spelling that to some extent interfere with meaning.	Serious errors in mechanics, usage, grammar or spelling, which make the summary difficult to understand.
Signal Words	Uses the clear and accurate signal words to connect information.	Uses several clear and accurate signal words to connect information.	Uses several clear signal words to connect information.

2.4 Measures

In this study, we employed three approaches to assess summaries: semantic meaning measured by LSA similarity, Coh-Metrix, and LIWC. The crowdsourcing-based LSA similarity score was the LSA cosine between a target summary and all the wild summaries from the corresponding source text. 94 language and discourse features were utilized to train and build the Coh-Metrix summary assessment model. All of 93 psychometric word features were utilized to train and build the LIWC summary assessment model.

2.5 Procedure

Participants took a demographic survey, a pretest (1 comparison and 1 causation), training (2 comparisons and 2 causations), and a posttest (1 comparison and 1 causation). On tests, participants wrote summaries by themselves. During training, two agents first interactively presented the importance of signal words for two text structures (comparison and causation) and how to use signal words to identify the corresponding text structure. Then participants interacted with the conversational agents to learn a summarizing strategy with adaptive scaffolding. Participants were required to write a summary with 50 to 100 words for each text. If the amount of words was beyond the range, the agents reminded the participants of the required length. If the participants copied the original sentences with 10 consecutive words, the agents reminded them of using their own words. Agents did not provide the adaptive feedback for their summary writing, but commented on three summary examples with good, medium, and bad qualities for each source text. The primary interface during training was shown in Figure 1.



Note. (A) the teacher agent, Cristina (female), (B) the student agent, Jordan (male), (C) the instruction of the presented question, (D) the text presented with the scroll down button, (E) an input text-box for participants to enter and submit their summaries, and (F) the self-paced next button.

Figure 1. Screenshot of Learning Interface.

3. RESULTS

A series of linear regressions with 10-fold cross-validation in WEKA was performed on 7 models, respectively. Fisher z was used to compare the difference between two pairs of correlations (see Table 3). Results revealed that crowdsourcing-based LSA similarity robustly predicted human summary grading (r = .44; R2 = .19), as well as 55 Coh-Metrix measures (r = .43; R2 = .18), 57 LIWC measures (r = .47; R2 = .22), and 108 measures by Coh-Metrix (57) and LIWC (51) jointly (r = .46; R2 = .21). This indicates that the variance explained by one LSA similarity measure is equivalent to the variance explained by more than 55 language features or word features, and more than 100 language and word features jointly.

Adding 94 Coh-Metrix features to CLSAS added an additional variance (r = .51; R2 = .26) in explaining human grading scores. Adding 93 LIWC features also added an additional variance (r = .55; R2 = .30). Adding both Coh-Metrix and LIWC feature added an additional variance (r = .49; R2 = .24), but the increased variance was significantly lower than by adding either Coh-Metrix or LIWC features. Due to the limited pages and the significant predictors in the Coh-Metrix + LIWC model overlapped with those in the Coh-Metrix model or the LIWC model, we only reported the predominant predictors in the Coh-Metrix model and LIWC model as below.

The 55 Coh-Metrix measures consisted of 9 descriptive (e.g., word count, sentence length), 4 referential cohesions (e.g., noun overlap, argument overlap), 5 LSA overlap (e.g., adjacent sentences, LSA given, LSA new), 3 lexical diversity (e.g., type-token ratio), 5 connectives (e.g., logical, additive), 3 situation model (e.g., causal verbs and particles, LSA verb overlap), 5 syntactic complexity (e.g., minimal edit distance, sentence syntax similarity), 4 syntactic pattern density (e.g., noun phrase density, verb phrase density), 16 word information (e.g., noun, adjective, hypernymy for nouns), and 1 readability (e.g., Flesch Kincaid Grade Level).

The 57 LIWC features consisted of 3 summary variables (e.g., analytical thinking, authentic), 3 language metrics (e.g., sentence length, words with more than 6 letters), 11 function words (e.g., personal pronouns), 4 grammar other (e.g., regular verb, quantifiers), 4 affect words (e.g., emotion words, anger), 3 social words (e.g., friend, gender referents), 3 cognitive processes (e.g., tentativeness, certainty), 3 perceptual (e.g., seeing, hearing), 3 biological processes (e.g., body, health), 2 core drives and needs (affiliation and risk focus), 1 relativity, 4 personal concerns (e.g., religion, home), 2 informal speech (swear and filler), and 3 all punctuations (e.g., apostrophes, comma).

Table 3. Fisher's *z*: Comparisons of Correlations

Models	1	2	3	2+3	1+2	1+3
1 (*r*=.44)	---					
2 (*r*=.43)	-0.34	---				
3 (*r*=.47)	1.03	1.36	---			
2+3 (*r*=.46)	0.68	1.02	-0.35	---		
1+2 (*r*=.51)	2.46**	2.80**	1.43	1.78*	---	
1+3 (*r*=.55)	3.97***	4.31***	2.94**	3.29**	1.51	---
1+2+3 (*r*=.49)	1.74*	2.07*	0.71	1.05	-0.73	-2.24*

Note. 1 = LSA similarity; 2 = Coh-Metrix features; 3 = LIWC features. * $p < .05$. ** $p < .01$. *** $p < .001$.

4. DISCUSSION

This paper developed an effective and efficient automated summary assessment, called crowdsourcing-based LSA similarity (CLSAS). Crowdsourcing enables a diverse and a mass of people to produce abundant wild summaries. CLSAS used the wild summaries rather than the human expert summaries as the reference when computing LSA similarities. The CLSAS was validated by comparing with Coh-Metrix language features, LIWC word features, and both language and word measures together with human-scored summaries as the criteria. Results indicated that CLSAS measure predicted human summary grading as well as over 55 language measures, 57 word measures, and 108 language and word measures, respectively. Even though adding language features, word features, or both to CLSAS improved the predictability, the predictability of CLSAS alone is most robust with correlation coefficient above 6.74 in each model. Findings imply that crowdsourcing-based LSA similarity approach is a promising method and will have good popularity in automated summary assessment.

One possible explanation for the significant predictability of CLSAS is that the wild summaries generated by diverse populations display diverse qualities as compared with few expert summaries. These wild summaries maximally represent the target summary. On the hand, the wild summaries represent neutralized or averaged semantic meaning, which is called *centroid*. The centroid might better capture the semantic meaning represented in

the target summary. For example, the CLSAS model showed that LSA similarity had a very high coefficients, β = 8.60, which was substantially higher than other measures' in other models.

The Coh-Metrix measures are different from the crowdsourcing-based LSA similarity due to its nature on measuring cohesion, language, and readability rather than semantic meaning [10]. One semantic measure of LSA similarity between the target summary and the crowdsourcing-based summaries is equivalent to 55 Coh-Metrix language measures. Among these language measues, LSA overlap among all sentences in paragraph reached 5.43 for mean and 2.07 for standard deviation; LSA given/new -3.60 for mean and -2.39 for standard deviation; and LSA overlap between adjacent sentences, -1.20 for mean. The other measures showed very low coefficients, generally below 1.00. This implies that a range of language measures jointly plays a role in assessing summaries, but LSA measures are attributed more than others.

Besides the predominant role of LSA measures, other important Coh-Metrix measures included lexical diversity (β = 3.92) measured by type-token ratio. Type-token ration is widely used for both automated essay assessment [19] and automated summary assessment [4]. When the type-token ratio is high, namely, more unique words are used, the lexical diversity is high and the text is likely to be either very low in cohesion or very short. Oppositely, when the type-token ration is low, namely, more words are repeatedly used, the lexical diversity is low, but cohesion is high. Summarizing requires conciseness and briefness, so in one summary, repeatedly using the same word will lower the quality of summary. Another two crucial measures are sentence syntax similarity between adjacent sentences (β = 4.49) and across paragraphs (β = -4.71). The high syntax similarity between adjacent sentences suggests the uniformity and consistency of the syntactic construction. This implies that the whole summary is consistent in syntactic construction. However, the low syntax similarity across paragraphs results in greater syntactic variety.

Another two most robust predictors are paragraph count (β = -12.74) and word length (number of syllables; β = 4.32). These two measures are frequently used in the automated summary [4] and essay assessment [19]. Our study controlled the number of words of summaries, which explains why word count is not a robust predictor, as compared with the previous studies [9]. As the summary should be brief and concise, more paragraphs demonstrate the poor quality in conciseness. However, the high word length increases difficult to read and represents an academic or formal language style [25] in the summary.

The phenomena that the Coh-Metrix features were unevenly weighted did not occur in the LIWC features. Specifically, among Coh-Metrix measures, the measures such as cohesion, syntactic and lexical complexity are more robust than measures at the word level. LIWC measures are all at the word level, but go beyond the linguistic words. They expand to diverse psychometric words, such as analytical thinking, emotion, and social. All the LIWC measures are evenly weighted to predict human summary scores. This pattern occurs in the Coh-Metrix and LIWC joint model as well. These findings suggest that each type of words plays a small piece of role, as compared to language and semantic measures.

Fisher's *z* comparisons CLSAS with Coh-Metrix measures, LIWC measures, and Coh-Metrix + LIWC measures demonstrated no differences in explained variance in human summary grading between CLSAS and Coh-Metrix, CLSAS and LIWC, and CLSAS and Coh-Metrix + LIWC. The findings supported our hypothesis that CLSAS could predict human summary grading as well as dozens of language measures and/or LIWC measures.

To further evaluate the validity of CLSAS, we added Coh-Metrix, LIWC, and Coh-Metrix + LIWC measures to CLSAS model with different combinations. Results showed adding each of these features increased the predictability. It is easier to explain the incremented model because the language and word features represent different aspects of summary assessment and enable to compensate the semantic feature. No matter what features were added to CLSAS, CLSAS is consistently the most significant feature in the models. Specifically, the correlation coefficient of LSA was 7.49, 6.74, and 6.80 when adding the Coh-Metrix language features, the LIWC word features, and both, respectively. Therefore, LSA similarity was a robust feature for summary assessment, no matter when it is used alone or jointly with other features.

5. CONCLUSION

These findings suggest that crowdsourcing-based LSA similarity (CLSAS) is a robust predictor of human summary grading and it is a reliable measure for the automated summary assessment, as compared with a range of language and word measures. As CLSAS has a powerful predictability for human summary score, the wild summaries are assumed as a promising and encouraging approach to replace the expert summaries for its time-saving and efficient. Opposed to the tedious and time-consuming manual summary grading, the wild summaries have no doubt for its popularity and practicability for teachers. This efficient and effective summary grading could dramatically encourage and motivate the teachers to instruct the summarization strategy. Consequently, this will enhance the students' summarization skills, especially summary writing. For example, when teachers need to grade the students' summaries, they could use all of the summaries that the students wrote as the reference. These summaries wildly generated by the students represent diverse qualities. For a particular target summary, the teacher only clicks the target summary and its CLSAS will be automatically computed with all of the summaries. Each time teachers need summary grading, they could repeat this cycle, no any human grading is needed. Based on the LSA similarity score, the summary score could be generated.

This crowdsourcing approach could be popularized and applied to the ITS learning and assessment environment as well. The current ITS assessment assesses the open response with a list of stored expectations and misconceptions [19]. Unfortunately, students' answers could not be assessed accurately due to the unmatched "golden" reference. To address this issue, the crowdsourcing generated responses could be adopted as the reference to replace the limited number of responses that the human expert generates. However, the reliability and validity of the wild open responses need to be evaluated in the future research.

The future study should concentrate on scaling crowdsourcing-based LSA similarity score into 3- or 5-point scales that teachers usually use for a better interpretation. The present study only showed its predominant role in summary assessment without specifying the extent to which LSA similarity score represents the different levels of summaries. The present study compared the CLSAS approach with dozens of measures, which may have an overfitting problem. The future study could select the most popular features that are used in the automated summary assessment and compared them with the CLSAS approach.

To sum up, this study proposed an innovative approach, crowdsourcing-based summary assessment, to the summary assessment from two perspectives. First, the summary reference could be a range of summaries that are wildly generated by a lot of population who are not necessary to be experts. Second, LSA similarity between the target summary and the wildly-generated summaries is a powerful predictor for human summary grading. This innovation will advance the development of automated assessment, especially automated assessment in the ITS.

6. ACKNOWLEDGMENTS

The research reported in this paper was supported by the National Science Foundation (0325428, 633918, 0834847, 0918409, 1108845) and the Institute of Education Sciences (R305A080594, R305G020018, R305C120001, R305A130030).

7. REFERENCES

[1] McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication, 27*, 57–86. DOI= 10.1177/0741088309351547.

[2] Crossley, S. A., Salsbury, T., McNamara, D. S., and Jarvis, S. 2010. Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 18*, 561-580. DOI= 10.1177/0265532210378031.

[3] Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction, 7*(3), 161-195. DOI= 10.1207/s1532690xci0703_1.

[4] Madnani, N., Burstein, J., Sabatini, J. and O'Reilly, T., 2013. Automated scoring of a summary writing task designed to measure reading comprehension. NAACL/HLT 2013, 163.

[5] Kintsch, W., 1998. Comprehension: A paradigm for cognition. Cambridge university press.

[6] Friend, R., 2001. Effects of strategy instruction on summary writing of college students. *Contemporary Educational Psychology, 26*(1), 3-24. DOI= 10.1006/ceps.1999.1022.

[7] G. Yu. 2003. Reading for summarization as reading comprehension test method: Promises and problems. *Language Testing Update, 32*:44–47.

[8] Passonneau, R. J., Chen, E., Guo, W. and Perin, D. 2013. Automated pyramid scoring of summaries using distributional semantics. In *ACL* (Sofia, Bulgaria, August 4-9, 2013), 143-147.

[9] Cai, Z., Graesser, A.C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., Halpern, D. and Butler, H., 2011. Trialog in ARIES: User input assessment in an intelligent tutoring system. In *Proceedings of the 3rd IEEE International Conference on Intelligent Computing and Intelligent Sy*stems (San Francisco, C.A., August 4-6, 2011), 429-433

[10] McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. New York: Cambridge University Press.

[11] Pennebaker, J.W., Boyd, R.L., Jordan, K. and Blackburn, K., 2015. *The development and psychometric properties of LIWC2015*. UT Faculty/Researcher Works.

[12] Lin, C.Y. and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (Edmonton, Canada, May 27-June 1, 2003). Association for Computational Linguistics, 71-78

[13] Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M. and Horton, J. 2013. . The future of crowd work. In *Proceedings of the Conference on Computer Supported Cooperative work* (Antonio, TX, February23-27, *2013*), ACM, 1301-1318

[14] Kim, J., Cheng, J. and Bernstein, M.S. 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative work & Social Computing* (Vancouver, BC, March 14-18, 2014). ACM, 745-755

[15] Kim, J. and Monroy-Hernandez, A., 2015. *Storia: Summarizing social media content based on narrative theory using crowdsourcing*. arXiv preprint arXiv:1509.03026.

[16] Matias, J.N. and Monroy-Hernandez, A., 2014, . NewsPad: Designing for collaborative storytelling in neighborhoods. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (Totonto, Canada, April 26-May 01, 2014). ACM, 1987-1992.

[17] Agapie, E. and Monroy-Hernandez, A., 2015. Eventful: Crowdsourcing Local News Reporting. arXiv preprint arXiv:1507.01300.

[18] Landauer, T. K., McNamara, D., Dennis, S., and Kintsch, W. (Eds.). (2007). Handbook of latent semantic analysis. Mahwah, NJ: Erlbaum.

[19] Li, H., Shubeck, K., and Graesser, A. C. (2016). Using technology in language assessment. In D. Tsagari and J. Banerjee (Eds.), Contemporary second language assessment. London, UK: Bloomsbury Academic.

[20] Landauer, T.K., Laham, D. and Foltz, P.W. 2003. Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*(3), 295-308.

[21] Burstein, J. 2003. The E-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum, 113-122.

[22] Nenkova, A. and Passonneau, R., 2004. Evaluating content selection in summarization: The pyramid method.

[23] Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*, 371-398. DOI: 10.1111/j.1756-8765.2010.01081.x.

[24] Graesser, A.C., McNamara, D.S., Cai, Z., Conley, M., Li, H. and Pennebaker, J., 2014. Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115(2)*, 210-229. DOI= 10.1086/678293.

[25] Li, H., Graesser, A.C., Conley, M., Cai, Z., Pavlik Jr, P.I. and Pennebaker, J.W., 2015. A New Measure of Text Formality: An Analysis of Discourse of Mao Zedong. *Discourse Processes*, 1-28. DOI= 10.1080/0163853X.2015.10101.