Automated Summarization Evaluation (ASE) Using Natural Language Processing Tools

Scott A. Crossley[1], Minkyung Kim[1], Laura K. Allen[3], Danielle S. McNamara[2]

[1]Georgia State University

[2]Arizona State University

[3]Mississippi State University

Abstract

Summarization is an effective strategy to promote and enhance learning and deep comprehension of texts. However, summarization is seldom implemented by teachers in classrooms because the manual evaluation of students' summaries requires time and effort. This problem has led to the development of automated models of summarization quality. However, these models often rely on features derived from expert ratings of student summarizations of specific source texts and are therefore not generalizable to summarizations of new texts. Further, many of the models rely of proprietary tools that are not freely or publicly available, rendering replications difficult. In this study, we introduce an automated summarization evaluation (ASE) model that depends strictly on features of the source text or the summary, allowing for a purely textbased model of quality. This model effectively classifies summaries as either low or high quality with an accuracy above 80%. Importantly, the model was developed on a large number of source texts allowing for generalizability across texts. Further, the features used in this study are freely and publicly available affording replication.

Keywords: Natural language processing, Summarization, Discourse, Writing, Machine learning, Summary scoring

Automated Summarization Evaluation (ASE) Using Natural Language Processing Tools

Scott A. Crossley[1(✉)], Minkyung Kim[1], Laura Allen[3], and Danielle McNamara[2]

[1] Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA 30303, USA
{scrossley, mkim89}@gsu.edu
[2] Department of Psychology, Arizona State University, Tempe, AZ, USA
dmcnamara@asu.edu
[3] Department of Psychology, Mississippi State University, Starkville, MS 39762, USA
lka22@msstate.edu

Abstract. Summarization is an effective strategy to promote and enhance learning and deep comprehension of texts. However, summarization is seldom implemented by teachers in classrooms because the manual evaluation of students' summaries requires time and effort. This problem has led to the development of automated models of summarization quality. However, these models often rely on features derived from expert ratings of student summarizations of specific source texts and are therefore not generalizable to summarizations of new texts. Further, many of the models rely of proprietary tools that are not freely or publicly available, rendering replications difficult. In this study, we introduce an automated summarization evaluation (ASE) model that depends strictly on features of the source text or the summary, allowing for a purely text-based model of quality. This model effectively classifies summaries as either low or high quality with an accuracy above 80%. Importantly, the model was developed on a large number of source texts allowing for generalizability across texts. Further, the features used in this study are freely and publicly available affording replication.

Keywords: Natural language processing · Summarization · Discourse · Writing · Machine learning · Summary scoring

1 Introduction

There are a number of different strategies to teach students to comprehend and produce text, including text summarization [1]. Indeed, a recent meta-analysis indicated that summarization techniques improve comprehension in over 90% of studies [2]. Text summarization is unique in that it allows students to read and comprehend short texts and then demonstrate their understanding of those texts in writing. Thus, summarization taps into both reading and writing skills, essentially *writing to learn*. Research has

S. Isotani et al. (Eds.): AIED 2019, LNAI 11625, pp. 84–95, 2019.
https://doi.org/10.1007/978-3-030-23204-7_8

demonstrated that summarization effectively promotes learning, enhances deeper understanding of domain topics [3–5] and helps students practice critical academic and life skills including distinguishing relevant from irrelevant material and integrating new information with prior knowledge [6]. Written summaries also provide students with the opportunity to practice writing skills including writing objectively, communicating main ideas, paraphrasing, and developing cohesive structures [7–10].

While multiple studies have shown that summarization techniques benefit students, learning to write summaries requires practice. Students often write too much, write too little, copy verbatim, or fail to appropriately synthesize information [11]. Fortunately, summarization strategies can be taught and are effective for a wide range of students, including less skilled readers [12, 13], language learners [14], and students with learning disabilities [15]. Indeed, a meta-analysis reported that the average weighted effect size for summarization instruction for adolescent learners (Grades 4–12) was quite large [16].

Instruction and practice enhance summarization skills; however, it is challenging for teachers to implement summarization tasks because evaluating summaries requires effort and time [17, 18]. In response, a number of methods have been developed for automated summarization evaluation (ASE) in order assess textual elements related to summaries including integrated content, accuracy of content, language use, and text coherence [17, 19–24]. However, many of these approaches rely on specific information outside of the text to predict summarization quality (e.g., expert summarizations of the source text) or depend on text features that are not publicly available. Such limitations make these approaches less generalizable, difficult to replicate, and problematic to implement in a dynamic learning environment. Thus, the purpose of this study is to develop an automated linguistic model of text summarization quality that is founded on natural language processing (NLP) features that are publicly available and do not require topic specific data. We do so by analyzing over 1,000 summaries collected through crowd-sourced and traditional techniques. The summaries were written on over 30 prompts providing some assurance that an algorithm gleaned from them will be generalizable. In addition, the bases for our algorithm are linguistic features that are freely and publicly available to all researchers to facilitate replication of our results. Our goal is to develop an ASE model that can be used to inform on-line tutoring and feedback systems and in turn provide strategy instruction to students as well as formative and summative feedback to writers regarding the quality of their summaries.

1.1 Summarization

Good summaries provide a concise overview of the most important content in a given passage. To do so, individuals need to construct a coherent, accurate mental model of the passage and paraphrase it using concise statements [25, 26]. Thus, successful summarization depends on (at least) two processes: comprehending source material and reproducing the key elements of that material. Generally speaking, this is achieved through the processes of reading and writing. Successful summarization depends on identifying main ideas and their supporting ideas as well as detecting the rhetorical and organizational structures of the source text [3, 27]. Once these textual elements have been identified, the content of the source text needs to be reproduced by

communicating the main points of the text, generally in writing, by using inferences or generalizations, omitting unimportant details, rephrasing portions of the text, using objective language, and developing cohesive structures [8, 9, 11, 27].

Summary writing enhances reading comprehension [3, 4] by encoding and strengthening the retention of information [28–30]. Summarizing texts may also improve the retention of information to a greater degree than other educational tasks including argumentative writing [12], short-answer questions [31], and both fill-in-the-blank and multiple-choice questions [32]. Summary writing also helps develop writing skills, requiring students to convey information in a succinct manner in one's own words and build coherence through well connected sentences [25, 33].

1.2 Automated Summarization Evaluation (ASE)

Research has demonstrated the benefits of using summarizations in education settings, but the adoption of summarization tasks is not widespread. Manually scoring summaries is time-consuming and potentially subjective [17, 18]. Automatically assessing summary quality can help alleviate time constraints and, as a result, has become an important component of many educational technologies. Specifically, researchers are interested in how summarizations can be scored automatically using linguistic and semantic features of the summary and the source text and how these same features can be used to provide feedback to learners in online educational systems [17].

In assessing summary quality, the most important criterion is the inclusion of complete and accurate content that is consistent with the source text [17, 22–24]. There are a number of different approaches that have been used to automatically assess summary quality including overlap of key n-grams between source and summaries, examining lexical and semantic overlap between source texts and their summaries (or expert summaries), and the use of rhetorical devices in summaries including connectives. Perhaps the most common approach is the use of Latent Semantic Analysis (LSA) to assess overlap between summaries and their sources. LSA is a mathematical method to represent the meaning of words and text segments based on large text corpora, and in turn, the extent to which documents are semantically related to one another [20, 21]. Early work by Landauer et al. [34] focused on assessing students' reading skills by their ability to summarize short articles or excerpts from leveled readers within a system called Summary Street. Studies on Summary Street indicated that students receiving feedback from the system performed better than those in control groups [5, 35]. Other research has examined the potential for LSA to assess summarization quality using semantic overlap between student summaries and the source text as well as between student summaries and high-quality summaries written by peers and experts. These studies have shown that that semantic similarity with the expert rater summary [17, 19] and highly rated peer summaries [17] explain significant amounts of variance in human judgments of summarization quality.

Other researchers have relied on n-gram (i.e., multi-word phrases) overlap between summaries and source texts and linguistic features within summaries to predict summarization quality. For instance, Madnani et al. [22] successfully predicted summary scores using n-gram overlap between the summary and the source text, lexical and phrasal overlap between the student summary against a set of model (or reference)

summaries that received the highest score, the ratio of n-grams copied from the source text, number of sentences in the summary, and incidence of discourse connector terms in the summary. In a later study, Sladoljev-agejev et al. [24] examined automated scoring of summaries for two source texts in the assessment of college-level writing in English as a second language using six analytic rubrics related to accuracy, completeness, relevance, coherence, cohesion and text organization. They used linguistic features related to n-gram overlap between summaries and source texts and between summaries and reference summaries along with linguistic indices computed by the NLP tool Coh-Metrix [36, 37]. They also found that accuracy was related to n-gram overlap and that scores for the other analytic features were only predicted by indices related to connectives and referential cohesion.

1.3 Current Study

In the current study, our objective is to improve upon previous studies by developing a an AES model that depends solely on textual features within the source text or in the summary – in essence, any text any time. In practice, this means that none of the features used to predict summarization quality are based on sources outside of the text allowing us to develop a model that can be extended to new source texts. To make potential extensions more reliable, we develop our model using a wide variety of source texts (n = 30). Further, the text features we use to develop our model are freely and publicly available ensuring that replications are possible.

2 Method

2.1 Data

A total of 1,023 summaries were collected from adult participants in the United States. Among them, 792 summaries of 30 different source texts were collected using the Amazon Mechanical Turk (MTurk) online research service. The MTurk workers in this study were asked to each write one summary on three different topics. The remaining 231 summaries were produced by adults who read at less than ninth grade level (i.e., adult literacy population). These participants were asked to each write a summary on two different source texts. Source texts were on unrelated topics ranging from child safety to internet shopping. The source texts were given by the California Distance Learning Project (CDLP), with permission from the Sacramento County Office of Education. The CDLP texts are simplified news stories that are developed to be read by low-literate adults to improve their comprehension skills [18]. Each of the CDLP texts was between four and eight paragraphs and ranged from 128 to 452 words (SD = 73.900 words). Flesch-Kincaid grade level was between 4th and 8th grade (SD = 1.100) for all texts. On average, each source text was summarized by 31 participants (SD = 27.644). The minimum number of participant summaries per text was 20 while the maximum was 129. Participants were given instruction on how to write a summary, shown a text, and then asked to summarize the text. No demographic or individual difference information was collected for the MTurk workers, while this information was collected for the adult literacy population.

2.2 Summary Rating

The summaries were scored by two expert raters using an analytic scoring rubric that focused on inclusion of main ideas, accuracy of main ideas, and appropriate length. All analytic features were on a 0–3 scale. Before rating the summarization, the raters examined the original texts independently and identified main ideas. Through adjudication, the raters created a finalized list of main ideas for each source text. During scoring, the raters referenced this list of main ideas. For the current analysis, we only focus on the inclusion of main ideas in the summary. Raters gave the summary a score of 3 if the summary included all main ideas, a score of 2 if the summary included most of the main ideas, a score of 1 if the summary included some main ideas, and a score of 0 if the summary included no main ideas. Expert raters were first trained with an independent set of summaries that were not part of the final corpus of summary. Once the raters were normed, they then scored the entire corpus of summaries independently.

2.3 Linguistic Features

Four NLP tools were used to collect information about the lexical sophistication, syntactic complexity, lexical diversity, and cohesion of the summaries. These tools were the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [38], and the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) [39], the Tool for the Automatic Analysis of Lexical Diversity (TAALED, a beta version), and the Tool for the Automatic Analysis of Cohesion (TAACO) [40]. These tools were selected because they are freely available, open-source, and well documented: all features that allow for replicability. Each is briefly discussed below.

TAALES. TAALES calculates lexical and phrasal features, such as lexical frequency (i.e., how often a word occurs in a reference corpus), psycholinguistic word information (e.g., human ratings of familiarity, imageability, and concreteness), and *n*-gram (i.e., sequences of contiguous words) frequency (i.e., how often an *n*-gram occurs in a reference corpus). In calculating indices related to frequency, various reference corpora are used, such as the SUBTLEXus corpus of subtitles [41] and the Corpus of Contemporary American English (COCA) [42]. Indices are calculated using all words, content words, or function words.

TAACO. TAACO computes indices related to textual features and text cohesion, such as type-token ratios (TTRs; the number of types [unique words] divided by the number of tokens [total running lemmas]), sentence overlap, paragraph overlap, and connectives (e.g., *moreover* and *nevertheless*). Overlap indices are calculated based on lemma overlap and semantic overlap (e.g., using LSA and vector-representation-of-words [word2vec]). TACCO also provides source-text similarity indices using LSA, latent dirichlet allocation (LDA), and word2vec.

TAALED. TAALED calculates lexical diversity indices, such as the measure of textual lexical diversity (MTLD; the mean length of sequential word strings in a text that maintain a given TTR value) [43] and Moving-average TTR (MATTR). Indices are calculated using all lemmas, content lemmas, function lemmas, or bi-grams.

TAASSC. TAASSC computes indices related to clausal and phrasal indices of syntactic complexity and indices related to complexity of verb-argument constructions (defined as a main verb plus all of its direct dependents). These features include clausal complexity (e.g., average numbers of particular structures per clause and dependents per clause) and noun-phase complexity (e.g., standard deviations of dependents per each noun phrase).

2.4 Statistical Analysis

Based on human ratings, the summary texts were grouped into two categories: a low-quality group in which summary texts were scored 0 or 1 (n = 432) and a high-quality group in which summary texts were scored 2 or 3 (n = 591). To par down the number of linguistic features and control for statistical assumptions, a series of pre-analytic pruning steps were undertaken. First, linguistic features for which correlations with summary scores were lower than |.20| were removed. Our threshold of .20 ensured that only variables with meaningful relations were included in the analysis. These linguistic indices were then controlled for multicollinearity (defined as $r > .700$).

To predict summary scores, a generalized linear mixed model (GLMM) was used. GLMMs combine linear mixed models (which handle both fixed and random effects) and generalized linear models (which address non-normal data, such as binomial distributions) to develop predictive model. In our GLMM model, the response/dependent variable was a binomial response defined as either high-quality or low-quality summarizations. The fixed effects in the analysis were the linguistic features calculated in each summary text. The random effects in the analysis quantified variation across source texts and participants. Thus, GLMMs can measure the effects of the linguistic features on the response variable (i.e., high-quality or low-quality of summary texts) while accounting for prompt effects and the repeated testing of the same participants. The GLMM model developed for this study using backward selection of the fixed effects, such that only significant fixed effects ($t > 1.96$ at a .05 significance level) were retained. We also tested interaction effects among the significant fixed effects. We then included random slope adjustments of source texts for each significant fixed effect because the effects of linguistic measures on summary scores are likely to differ depending on the source text.

The data were randomly divided into a training set and a test set using a 67/33 split [44]. The GLMM was created using the training set (n = 685), and then applied into the test set (n = 338) to evaluate how well the model classified an independent set of summaries. The test set contained approximately 33% of the summaries from each of the participant groups - 264 summaries from the MTurk participants (164 high quality; 100 low quality) and 74 summaries from the adult literacy participants (17 high quality; 57 low quality). Descriptive statistics indicated that the adult literacy participants produced more low quality summaries than the MTurk participants.

For data analysis, we used *R* (R Core Team, 2016) and the *lme4* package [45] to construct a GLMM model. We used the *LMERConvenienceFunctions* package [46] to perform backward selection of fixed effects and the *MuMIn* package [47] to calculate a marginal *r*-squared (i.e., variance explained by fixed effects only) and a conditional *r*-squared (i.e., variance explained by both fixed and random effects).

3 Results

After pruning the data using test for multicollinearity and effect size, 21 linguistic indices were retained and used to develop the GLMM model. Details on these features are reported in Table 1.

Table 1. Correlations between summary scores and computational indices

Index	Feature	Tool	*r*
Type-token ratio (AW)	Cohesion	TACCO	.469
Frequency (COCA spoken, AW)	Lexical sophistication	TAALES	–.435
Source similarity (word2vec)	Cohesion	TAACO	.406
Adjacent sentence similarity (word2vec)	Cohesion	TAACO	.362
MATTR (FW)	Lexical diversity	TAALED	–.324
Number of CW tokens	Lexical diversity	TAALED	.318
MTLD (FW)	Lexical diversity	TAALED	.309
MTLD bi-grams (FW)	Lexical diversity	TAALED	.303
MTLD (AW)	Lexical diversity	TAALED	.292
MTLD (CW)	Lexical diversity	TAALED	.292
Word frequency (COCA spoken, AW)	Lexical sophistication	TAALES	–.282
Repeated content lemmas and pronouns	Cohesion	TAACO	.282
Lexical density (Percentage of CWs)	Cohesion	TAACO	.254
Type-token ratio (CW)	Cohesion	TAACO	–.247
Frequency (COCA spoken, FW)	Lexical sophistication	TAALES	–.246
SD of dependents per nominal subject	Syntactic complexity	TAASSC	.232
Binary adjacent sentence overlap (FW)	Cohesion	TAACO	.229
Word frequency (SUBTLEXus, CW)	Lexical sophistication	TAALES	–.225
Word frequency (SUBTLEXus, AW)	Lexical sophistication	TAALES	–.216
SD of dependents per clause	Syntactic complexity	TAASSC	.215
SD of dependents per object of the preposition	Syntactic complexity	TAASSC	.204

AW = All words, CW = Content words, FW = Function words, SD = Standard Deviation

Using the training set (n = 685) to develop a baseline GLMM, a random intercept model was created including the prompts and the participants as random intercepts. This model explained 45.198% of the variance in the human ratings of summarization. Performing backward selection of fixed effects, the GLMM included four significant linguistic features (see Table 2). No significant interaction effects were revealed. The results indicated that high-quality summary texts tended to have higher type-token ratios of all words (i.e., less repetition of words/more unique words; t = 7.424, p < .001), greater similarity with source texts as measured by word2vec (t = 3.388, p < .001), words that occur less frequently in the COCA spoken corpus (t = –3.376, p < .001), and lower type-token ratios for content words (i.e., greater repetition of content words; t = −2.592, p < .01). This model explained 53.412% of the variance using the fixed factors (i.e., the linguistic features) and 80.756% of the variance using both fixed and random factors.

Table 2. Results of the generalized linear mixed model (GLMM)

Fixed effect	Feature	Estimate	Standard error	*t*	*p*
(Intercept)		−6.975	2.605	−2.678	<.010
Type-token ratio for all words	Cohesive	1.596	.246	6.493	<.001
Source similarity (word2vec)	Cohesive	7.424	2.191	3.388	<.001
Frequency for all words (COCA spoken)	Lexical	–.001	.001	−3.376	<.001
Type-token ratio for content words	Cohesive	–4.433	1.710	−2.592	<.010

The final GLMM was used to examine how accurately it classified high- and low-quality summaries in the training set (n = 685; see Table 3). The GLMM correctly allocated 597 of the 685 summaries in the training set for an accuracy of 87.153%. The precision scores (i.e., the ratio of correctly classified hits into the high-quality group to all hits classified into the high-quality group; 374/426) and recall scores (i.e., the ratio of correctly classified hits into the high-quality group to hits incorrectly classified into the low-quality group plus hits correctly classified into the high-quality group; 374/410) were .878 and .912, respectively. The combined accuracy of the model (F1) was .895.

Table 3. Confusion matrix for classifying high- and low-quality summaries in the training set

Actual group	Predicted group	
	Low-quality	High-quality
Low-quality	223	52
High-quality	36	374

The GLMM classification model was next extended to the test set (n = 338) to assess classification accuracy (see Table 4). The GLMM correctly allocated 276 of the 338 summaries in the test set for an accuracy of 81.657%. The precision scores (161/203) and recall scores (161/181) were .793 and .890, respectively. The F1 was .839. These results provide strong evidence that linguistic features as found in summaries can be used to classify summaries in terms of the inclusion of main ideas.

Table 4. Confusion matrix for classification of high- and low-quality summaries in the test set

Actual group	Predicted group	
	Low-quality	High-quality
Low-quality	115	42
High-quality	20	161

To assess generalization of the model for the two groups, we separately examined the accuracy of the model on the MTurk and Adult Literacy participant populations. The GLMM correctly allocated 216 of the 264 summaries in the MTurk summaries in the test set for an accuracy of 81.818%. The precision and recall scores were .826 and .896, respectively. The F1 was .860. Similarly, the GLMM correctly allocated 60 of the 74 summaries in the adult literacy populations. This yielded an accuracy of 81.081% with precision and recall scores of .560 and .824, respectively. The F1 was .667. Overall, these results are similar to those of the combined test set, suggesting that the model was generalizable to the two populations.

4 Discussion

Summarization is considered an effective strategy to promote and enhance learning and deep comprehension of texts. However, this strategy is seldom implemented by teachers in classrooms because the manual evaluation of students' summaries requires time and effort. This problem has led to the development of automated models of summarization quality. However, these models often rely on features outside of the student or source texts and are therefore not generalizable to summarizations of new texts. Furthermore, many of the models rely of proprietary tools that are not freely or publicly available, rendering replications difficult.

In this study, we developed and presented an ASE model that depends strictly on features of the source text or the summary, allowing for a purely text-based model of quality. Summaries (n = 1,023) were collected from adult participants in the United States. Among them, 792 summaries of 30 different source texts were collected using the Amazon Mechanical Turk online research service while other summaries were collected from adult literacy participants with low reading skills. The summaries were scored by two expert raters using an analytic scoring rubric that focused on inclusion of main ideas, accuracy of main ideas, and appropriate length Additionally, four NLP tools were used to collect information about the lexical sophistication, syntactic complexity, lexical diversity, and cohesion of the summaries.

Our derived model was able to classify summaries as either low or high quality with an accuracy above 80%. Importantly, the model was developed on a large number of source texts allowing for generalizability across texts. Moreover, the model generalized well to summaries produced by two different populations (i.e., MTurk and Adult Literacy populations) for different text genres (i.e., science texts and adult literacy texts). However, results were stronger for the MTurk population indicating that the model is better suited for summaries written by writers that are not low literacy. Further, the features used in this study are freely and publicly available affording replication. These results provide strong evidence that linguistic features as found in summaries can be used to classify summaries in terms of the inclusion of main ideas. Specifically, the model reported that higher rated summaries had greater lexical repetition of all word types (i.e., both function and content words), contained more infrequent words, had greater semantic overlap with the source text, and contained fewer repetitions of content words. These features support the notion that more lexically

sophisticated summaries that have greater source overlap and more repetition of words in general (although fewer repetitions of content words) are scored higher.

5 Conclusion

The current study provides a strong foundation for future research on the automated scoring of student summaries and writing more broadly. Specifically, in terms of summarization scoring, we are in the process of collecting additional summaries from different prompts to extend the number and types of prompts in our database. We also plan to examine expert ratings that go beyond source integration and include text cohesion, the use of objective language, paraphrasing, and language sophistication. We are also in the process of refining newer NLP tools that may provide insight into how textual features are predictive of summarization quality.

The overarching objective of this work is to develop models of writing assessment that are open and free for educators, students, and researchers. These models will be integrated within the Writing Assessment Tool (WAT) which is currently under development. WAT will provide automated writing evaluation (AWE) for persuasive essays, source-based essays, and ASE for summaries. WAT will also provide access to validated linguistic and semantic features that characterize writing quality to researchers so that features such as those identified in this study are readily available to researchers. In doing so, we hope to increase access to tools that can improve writing research and writing education.

Acknowledgments. This research was supported in part by the Institute for Education Sciences (IES R305A180261). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We would also like to express thanks to Amy Johnson, Kristopher Kopp, and Cecile Perret for their help in collecting the data.

References

1. Marzano, R.J., Pickering, D.J., Pollock, J.E.: Classroom Instruction That Works: Research-Based Strategies for Increasing Student Achievement. Association for Supervision and Curriculum Development, Alexandria (2008)
2. Graham, S., Herbert, M.A.: Writing to Read: Evidence for How Writing Can Improve Reading: A Carnegie Corporation Time to Act Report. Alliance for Excellent Education, Washington (2010)
3. Spirgel, A.S., Delaney, P.F.: Does writing summaries improve memory for text? Educ. Psychol. Rev. **28**, 171–196 (2016)
4. van Dijk, T.A., Kintsch, W.: Strategies of Discourse Comprehension. Academic Press, New York (1983)
5. Wade-Stein, D., Kintsch, E.: Summary street: Interactive computer support for writing (2004). http://www.tandfonline.com/doi/abs/10.1207/s1532690xci2203_3
6. Rinehart, S.D., Stahl, S.A., Erickson, L.G.: Some effects of summarization training on reading and studying. Read. Res. Q. **21**, 422–438 (1986)

7. Brown, A.L., Campione, J.C., Day, J.D.: Learning to learn: on training students to learn from texts. Educ. Res. **10**, 14–21 (1981)
8. Brown, A.L., Day, J.D.: Macrorules for summarizing texts: the development of expertise. J. Verbal Learn. Verbal Behav. **22**, 1–14 (1983)
9. van Dijk, T.A., Kintsch, W.: Strategies of Discourse Comprehension. Academic, New York (1977)
10. Westby, C., Culatta, B., Lawrence, B., Hall-Kenyon, K.: Summarizing expository texts. Top. Lang. Disord. **30**(4), 275–287 (2010)
11. Jones, R.: Strategies for reading comprehension: Summarizing
12. Gil, L., Bråten, I., Vidal-Abarca, E., Strømsø, H.I.: Summary versus argument tasks when working with multiple documents: which is better for whom? Contemp. Educ. Psychol. **35**, 157–173 (2010)
13. Perin, D., Lauterbach, M., Raufman, J., Kalamkarian, H.S.: Text-based writing of low-skilled postsecondary students: relation to comprehension, self-efficacy and teacher judgments. Read. Writ. **30**, 887–915 (2017)
14. Chiu, C.-H.: Enhancing reading comprehension and summarization abilities of EFL learners through online summarization practice. J. Lang. Teach. Learn. **5**(1), 79–95 (2015)
15. Rogevich, M.E., Perin, D.: Effects on science summarization of a reading comprehension intervention for adolescents with behavior and attention disorders. Except. Child. **74**, 135–154 (2008)
16. Graham, S., Perin, D.: A meta-analysis of writing instruction for adolescent students. J. Educ. Psychol. **99**(3), 445–476 (2007)
17. Li, H., Cai, Z., Graesser, A.C.: Computerized summary scoring: crowdsourcing-based latent semantic analysis. Behav. Res. Methods **50**(5), 2144–2161 (2018)
18. Ruseti, S., et al.: Scoring summaries using recurrent neural networks. In: Nkambou, R., Azevedo, R., Vassileva, J. (eds.) ITS 2018. LNCS, vol. 10858, pp. 191–201. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91464-0_19
19. Jorge-Botana, G., Luzón, J.M., Gómez-Veiga, I., Martín-Cordero, J.I.: Automated LSA assessment of summaries in distance education: some variables to be considered. J. Educ. Comp. Res. **52**(3), 341–364 (2015)
20. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychol. Rev. **104**, 211–240 (1997)
21. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.: Handbook of Latent Semantic Analysis. Lawrence Erlbaum, Mahwah (2007)
22. Madnani, N., Burstein, J., Sabatini, J., O'reilly, T.: Automated scoring of a summary writing task designed to measure reading comprehension. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 163–168 (2013)
23. Mani, I.: Automatic Summarization. John Benjamins Publishing, Amsterdam (2001)
24. Sladoljev-agejev, T., Snajder, J., Analysis, T.: Using analytic scoring rubrics in the automatic assessment of college-level summary writing tasks in L2. In: Proceedings of the 8th International Joint Conference on Natural Language Processing, pp. 181–186 (2017)
25. Dole, J.A., Duffy, G.G., Roehler, L.R., Pearson, P.D.: Moving from the old to the new: research on reading comprehension instruction. Rev. Educ. Res. **61**(2), 239–264 (1991)
26. Kintsch, W., Van Dijk, T.A.: Toward a model of text comprehension and production. Psychol. Rev. **85**, 363–394 (1978)
27. Kintsch, W., Welsch, D., Schmalhofer, F., Zimny, S.: Sentence memory: a theoretical analysis. J. Mem. Lang. **29**, 133–159 (1990)
28. Hinze, S.R., Rapp, D.N.: Retrieval (sometimes) enhances learning: performance pressure reduces the benefits of retrieval practice. Appl. Cogn. Psychol. **28**(4), 597–606 (2014)

29. Butler, A.C., Karpicke, J.D., Roediger III, H.L.: The effect of type and timing of feedback on learning from multiple-choice tests. J. Exp. Psychol. Appl. **13**(4), 273–281 (2007)
30. Stewart, T.L., Myers, A.C., Culley, M.R.: Enhanced learning and retention through "writing to learn" in the psychology classroom. Teach. Psychol. **37**(1), 46–49 (2009)
31. Shokrpour, N., Fotovatian, S.: Effects of consciousness raising of metacognitive strategies on EFL students' reading comprehension. ITL – Int. J. Appl. Linguist. **157**, 75–92 (2009)
32. Mok, W.S.Y., Chan, W.W.L.: How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? Instruct. Sci. **44**(6), 567–581 (2016)
33. Delaney, Y.A.: Investigating the reading-to-write construct. J. Engl. Acad. Purp. **7**, 140–150 (2008)
34. Landauer, T.K., Lochbaum, K.E., Dooley, S.: A new formative assessment technology for reading and writing. Theor. Pract. **48**(1), 44–52 (2009)
35. Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., Dooley, S.: Summary street: computer support for comprehension and writing. J. Educ. Comput. Res. **33**, 53–80 (2005)
36. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: analysis of text on cohesion and language. Behav. Res. Meth. Ins. C. **36**, 193–202 (2004)
37. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, Cambridge (2014)
38. Kyle, K., Crossley, S., Berger, C.: The tool for the automatic analysis of lexical sophistication (TAALES) version 2.0. Behav. Res. Methods **50**(3), 1030–1046 (2018)
39. Kyle, K.: Measuring syntactic development in L2 writing: fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. Doctoral Dissertation (2016). http://scholarworks.gsu.edu/alesl_diss/35
40. Crossley, S.A., Kyle, K., McNamara, D.S.: The tool for the automatic analysis of text cohesion (TAACO): automatic assessment of local, global, and text cohesion. Behav. Res. Methods **48**(4), 1227–1237 (2016)
41. Brysbaert, M., New, B.: Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav. Res. Methods **40**(4), 977–990 (2009)
42. Davies, M.: The 385+ million word Corpus of Contemporary American English (1990–2008+): design, architecture, and linguistic insights. Int. J. Corpus Linguist. **14**, 159–190 (2009)
43. McCarthy, P.M., Jarvis, S.: MTLD, Vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behav. Res. Methods **42**(2), 381–392 (2010)
44. Witten, I.A., Frank, E., Hall, M.A.: Data mining: Practical Machine Learning and Techniques. Elsevier, San Francisco, CA (2011)
45. Bates, D., Maechler, M., Bolker, B., Walker, S.: lme4: linear mixed-effects models using Eigen and S4. R Packag. Version **1**(7), 1–23 (2014)
46. Tremblay, A., Ransijn, J.: LMERConvenienceFunctions: a suite of functions to back-fit fixed effects and forward-fit random effects, as well as other miscellaneous functions. R Packag. Version **2**, 919–931 (2013)
47. Barton, K., Barton, M.K.: Package MuMIn. Model selection and model averaging based on information criteria (2018)