Automated Content Analysis: A Case Study of Computer Science Student Summaries

Yanjun Gao[1], **Patricia M. Davies**[2], and **Rebecca J. Passonneau**[1]

[1]Department of Computer Science and Engineering, Penn State University
[1]{*yug125,rjp49*}*@cse.psu.edu*
[2]Department of Computer Science, University of Wolverhampton
[2]*Patrica.Davies@wlv.ac.uk*

Abstract

Technology is transforming Higher Education learning and teaching. This paper reports on a project to examine how and why automated content analysis could be used to assess précis writing by university students. We examine the case of one hundred and twenty-two summaries written by computer science freshmen. The texts, which had been hand scored using a teacher-designed rubric, were autoscored using the Natural Language Processing software, PyrEval. Pearsons correlation coefficient and Spearman rank correlation were used to analyze the relationship between the teacher score and the PyrEval score for each summary. Three content models automatically constructed by PyrEval from different sets of human reference summaries led to consistent correlations, showing that the approach is reliable. Also observed was that, in cases where the focus of student assessment centers on formative feedback, categorizing the PyrEval scores by examining the average and standard deviations could lead to novel interpretations of their relationships. It is suggested that this project has implications for the ways in which automated content analysis could be used to help university students improve their summarization skills.

1 Situating Automated Content Analysis in Higher Education

Our present concerns are about CS students having difficulty summarizing or synthesizing texts accurately. Instead of staying focused, some tend to wander away from significant points in written reports. There are also issues relating to CS instructors wasting valuable time on badly written reports, especially in cases when class sizes are very large (with 150 to 250 students). This often results in students not receiving meaningful feedback that could help them to advance their learning. Increasing the availability and quality of timely feedback could significantly improve students' written-communication skills.

The focus of this study is to investigate how PyrEval (Gao et al., 2018a), an existing summary content analysis software tool, might be used to automate the assessment of student summaries, given a small set of reference summaries from which to construct a content model. Scores from an earlier implementation of automated pyramid scoring were shown to have high Pearson correlation of 0.83 with a main ideas rubric applied to 120 community college summaries (Passonneau et al., 2016); on the same summaries PyrEval has even higher correlation of 0.87. As such, the aim is not to examine its correctness here; instead, we seek to understand how it could be adapted for use within Higher Education (HE). In particular, we are interested in exploring how PyrEval might be used for formative, rather than summative, assessment of student work. With this view, the discussions here focus on PyrEval as a tool for helping students to improve written assignments prior to submission, thereby making the time instructors spend marking more beneficial.

Learning in HE, often described as constructivist, involves learners actively constructing knowledge and meaning based on prior experiences (Barr and Tagg, 1995; Bostock, 1998; Brockbank and McGill, 2007; Tess, 2013). In this approach, students create knowledge by connecting what they already know to new subject content encountered in lectures, texts and discussions. This shift in paradigm, from one where the learner retrieves information from the instructors, has prompted recently coined phrases such as, self-directed learning (Hiemestra, 1994) and student-centered learning (Lea et al., 2003). Unfortunately, assessing students' self-directed learning, and providing formative feedback in this learning

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 264–272
New Orleans, Louisiana, June 5, 2018.

approach, has not developed as rapidly.

Feedback is intended to provide students with information on their current state of learning and performance, and is essential for elevating students' motivation and confidence (Hyland, 2000). Rather than being an evaluation of performance on assigned tasks, formative feedback provides information to help students scaffold their knowledge and accelerate their learning (Sadler, 2010). Therefore, formative assessment applications play an important role by helping students take greater control of their own learning, and moves them towards becoming self-regulated learners.

Within HE, formative feedback is perceived as information communicated to the students about learning-oriented assignments (Race, 2001) such as essays. This feedback can be oral or written, and is often generated by the instructor. Providing feedback remains the responsibility of the instructor, and with much emphasis being placed on evaluating student learning at the end of an instructional unit, instructor feedback is often limited. Some even use custom software, such as E-rater®, used by the Educational Testing Service for automated scoring of essays, which provides a holistic score rather than a narrative. Our present concerns move beyond simply providing a score to examine how and why PyrEval could be used to provide formative feedback on students' summaries. It is distinctive in providing interpretable scores that can be justified by automated identification of important, unimportant and missing content (Passonneau et al., 2016). This study provides a conceptualization for the next steps in the development of the tool towards this end.

The next three sections present the following: background to the study through a review of existing literature; a summarization task given to CS students at a UK university along with a description of how it was assessed by the instructor, one of the authors PyrEval, an automated tool to analyze content of summaries that depends on a reference set of four or more expert summaries.

Section 5 presents our experiments to compare PyrEval scores of the students' summaries with scores assigned by the human scorer using a rubric. The findings show that PyrEval scores correlate moderately well with the rubric, but more importantly, the analysis led to reconsideration of scores for several summaries. Section 6 discusses the benefits and limitations of the automated tool, and our plans for future work.

2 Related Work

Summarization is an important pedagogical tool for teaching reading and writing strategies in elementary school (Kırmızı, 2009), middle school (Graham and Perin, 2007), community college (Perin et al., 2013), as part of blended instructional methods at the college level (Yang, 2014), and for English language learners (Babinski et al., 2017). Instruction in summarization strategies includes occasional forays into computer-based training (Sung et al., 2008), including intelligent tutoring systems that provide writing practice (Proske et al., 2012)(Roscoe et al., 2015).

Recent work built regression models to predict scores based on several rubrics for summaries from L2 business school students (Sladoljev Agejev and Šnajder, 2017). Features were automatically derived from Coh-Metrix (McNamara et al., 2014), BLEU scores (Papineni et al., 2002) and ROUGE scores (Lin, 2004). In (Srihari et al., 2008), OCR was used to digitize handwritten essays, which were then scored using various automated essay scoring methods, including latent semantic analysis and a feature-based approach. Essays are automatically scored in (Zupanc and Bosni, 2017) after constructing an ontology from model essays using information extraction and logic reasoning. PyrEval constructs a content model from a small set of reference summaries, using latent semantic vectors to represent meanings of phrases.

There has been recent interest in developing automated revision tools for students' written work but none have, hitherto, been reported in the literature. There is existing work on automated revision of short answers for middle school science writing (Tansomboon et al., 2017), and a corpus on automated revision of argumentation (Zhang et al., 2017). What is distinctive about our work is the feasibility of providing automated feedback on summary content, either for teachers or students, which could ultimately lead to the development of an automated revision tool.

3 Task and Educational Rubrics

3.1 The student setting and their task

At the start of this academic year, 159 CS students were enrolled in Academic Skills and Team-based Learning at Bakersview University (a pseudonym)

Scoring rubric

Item	Description of Idea
1	Dont take everything you read for granted. Always ask - says who? so what? what next?
2	Check if the references are accurate.
3	Check the authors' qualifications and experience (academic and practice), and what qualifies them to undertake this work. See if they have published any other works and if they have been cited by others.
4	If they have, it is worthwhile checking out some of those citations to see if they are positive or negative.
5	Check for evidence of how this information could have, or has, had an impact.
6	Ensure the data is provided to back up any arguments.
7	Understand how this information affects what you already know.
8	Check if there any consequences of this information that show the need for further research.
9	Critical thinking helps you identify potential strengths and weaknesses in the text.
10	Critical thinking helps you evaluate what you read and relate it to other information.

Figure 1: Scoring rubric for *Critical Thinking* task. Each of 10 items contributes 1 point.

in the UK. Bakersview is a non-selective university with an agenda to widen participation in higher education, and thus attracts students from a variety of learning backgrounds. Academic Skills and Team-based Learning is a core course taken by all CS freshmen. It aims to develop in students a range of written communication styles and approaches, and the critical reading skills, needed for academic and professional work. The goal is to give these students the opportunity to develop proficiencies and attitudes necessary for success at university and in employment.

The data being used for the present project came from student submissions for one of the assignments in the Academic Skills and Team-based Learning course. First, the students were asked to attend a workshop offered by the university's Library and Information Services. The focus of the workshop was finding information and critical thinking. Presentations and handouts were provided, and students were asked to make notes on the material covered. Following the workshop, they were asked to summarize, in no more than 200 words, what they learned during the workshop about critical thinking and its importance in HE.

3.2 The rubric

One hundred and thirty-nine summaries were submitted. These were then scored by hand using a rubric developed from the presentation given during the workshop. The 10 main points identified in the presentation were used as checkpoints in the rubric, which is shown in Figure 1.

One point was assigned to each of the ideas listed in the rubric; however, the interpretation of what constituted an idea was open to the discretion of the instructor. Each student received a score out of 10 for the assignment. A handful of student summaries did not meet the word-count requirement, these were not included in the anonymized samples for testing the autoscoring software PyrEval. Thirteen summaries, which received scores of 9 and 10, were used as reference summaries to construct a content model for interpretable scores, and the score justification.

4 System Description

PyrEval constructs a pyramid content model that consists of sets of distinct summary content units (SCUs) found in a set of N reference summaries written by experts or more advanced students, for $4 \leq N \leq 6$. In pyramid summary content evaluation, originally a manual annotation method, an SCU is similar to a set of paraphrases, each paraphrase drawn from a distinct reference summary (Nenkova et al., 2007). A given SCU can be expressed in anywhere from 1 to N summaries, so will consist of from 1 to N contributors from distinct summaries. The number of contributors to an SCU is an importance weight that is assigned to ideas in a new summary being scored. The weights of SCUs in a new summary are summed, and the sum is normalized in different ways, as described further below. A pyramid content model thus consists of all the distinct ideas, or SCUs, in the reference summaries, along with their weights.

To construct the pyramid content model automatically, sentences are first decomposed into distinct clausal or phrasal segments, then each segment is converted to a dense vector representation using Weighted Text Matrix Factorization (WTMF) (Guo et al., 2014). These semantic vectors are then grouped into semantically similar sets to form the SCUs, using a restricted set partition algorithm, EDUA, as noted below (Gao et al., 2018b). A new summary is scored against this content model by first segmenting the sentences and vectorizing them, then matching them to the

content model using a weighted set cover algorithm (Sakai et al., 2003). The following subsections describe the preprocessing (segmentation and conversion to dense vectors), pyramid construction, and scoring.

4.1 Preprocessing

The preprocessing step uses a sentence decomposition parser we implemented to produce alternative covering segmentations of each sentence, and WTMF (see above) to produce the dense vector representations. This is a pre-trained process so as to make PyrEval a light-weight tool that can be applied easily to new summarization tasks. The decomposition parser output is derived from constituency parsing and dependency parsing, using Stanford CoreNLP tools (Chen and Manning, 2014). The decomposition parser first locates every tensed verb phrase (VP) in the constituency parse, then uses the subject dependencies from the dependency parser to find each VP subject. The leftover words are reinserted into segments, according to their positions in the original sentence. We use WTMF to convert each segment into a vector representation for semantic similarity evaluation. It has proved to have high accuracy in sentence similarity tasks.

Sentence Decomposition Example

Critical thinking also means you must approach everything you read with a certain level of scepticism and find out if the points that are being made are backed up with evidence.	
Segmentation 1	
Segment 1	that are being made
Segment 2	you read with a certain level of scepticism
Segment 3	if the points are backed up with evidence
Segment 4	you must approach everything and find out
Segment 5	Critical thinking also means .
Segmentation 2	
Segment 1	you Critical thinking also means you must approach everything read with a certain level of scepticism and find out if the,
Segment 2	points that are being made
Segment 3	points are backed up with evidence .

Figure 2: Sentence decomposition parser output showing two alternative segmentations of the same sentence. The full sentence is also considered as a default segmentation.

4.2 Pyramid Construction

The core of PyrEval is an algorithm, *Emergent Discovery of Units of Attraction (EDUA)*, for allocating segments into SCUs according to their semantic similarity.

EDUA builds a graph G where vertices are segments and edges are semantic similarity above a threshold t_{edge}. Similarity values are distributed differently for different sets of summaries, so we define t_{edge} in terms of a selected percentile over the range of observed cosine values for a given set of reference summaries; from past work through grid-search on development sets we use $t_{edge} = 0.83$. An SCU is a connected component of G with at most N vertices, where the average edge weight leads to a high quality pyramid. The quality of an individual SCU is the average similarity (or attraction A_C) of its edges. Given a connected component C with k edges, A_C is defined as:

$$A_C = \frac{\sum_{u,v \in C, u \neq v} similarity(u, v)}{k} \quad (1)$$

The global attraction over the pyramid is given as:

$$A_P = \max \sum_{1}^{n} \left(\frac{1}{|C_n|} \sum_{1}^{|C_n|} A_C \right) \quad (2)$$

where n here represents the number of reference summaries, which in turn corresponds to the different sizes of SCUs in the pyramid.

EDUA's objective is to find a set of connected components (SCUs) that achieve the highest A_P, while obeying the constraints that no two segments from the same reference summary can be in the same SCU. We have developed two versions of the algorithm: EDUA-Complete (EDUA-C) and EDUA-Greedy (EDUA-G). EDUA-C performs a Depth First Search in the graph to find the set of SCUs with maximum A_P. EDUA-G takes a greedy approach and imposes a constraint based on the observation that SCU annotation follows a Zipfian distribution (Nenkova et al., 2007): there are a few SCUs that occur in every reference summary (maximum weight), more that occur in all but one, and so on, with a long tail of SCUs that occur in only one reference summary (minimum weight). SCU weight forms a partition over the set of SCUs. EDUA-G finds the SCUs with maximum A_C at each iteration n from N to 1, and allocates them into equivalence class n until the capacity of that class is full, then moves on to the next n. A constraint on the relative size of the equivalence classes requires them to adhere to a Zipfian distribution. Both EDUA variants perform equally well on a machine summarization task (Gao et al., 2018b). However, EDUA-C is computationally expensive. Hence we conducted experiments using EDUA-G.

Pair	Pearson(P-v)	Spearman(P-v)	Pair	Pearson(P-v)	Spearman(P-v)
P_1,R	46.47 (6.88e-08)	44.27 (3.28e-07)	P_1,P_2	73.82 (9.09e-23)	73.50 (1.70e-22)
P_2,R	49.18 (8.75e-09)	46.13 (8.893-08)	P_1,P_t	68.02 (2.69e-18)	68.63 (1.02e-18)
P_t,R	45.85 (1.39e-07)	44.77 (2.95e-07)	P_2,P_t	75.97 (9.67e-25)	77.33 (4.21e-26)

Table 1: Pearson correlation ($\rho \times 100$) and Spearman rank correlation ($r_s \times 100$) of PyrEval scores with rubric R (left columns), and with other PyrEval scores (right columns) given different pyramids. P-values are in parentheses.

4.3 Scoring

For matching segments from a summary to a pyramid, PyrEval applies WMIN, a weighted independent set allocation algorithm (Sakai et al., 2003). The scoring algorithm has proven its reliability to have good correlation with human annotation (Passonneau et al., 2016).

The input to WMIN consists of the vector representations of all segmentations produced by the decomposition parser for each sentence in a new summary. Vertices in the WMIN graph are matches between an SCU and a segment from a new summary, weighted by the product of the SCU weight and the mean cosine similarity of the summary vector to the SCU vectors; we use 0.5 as the similarity threshold (Passonneau et al., 2016). The objective is to find an assignment of SCUs to the new summary that produces the highest sum of SCU weights. WMIN ensures that no SCU is allocated more than once to a summary, and that segments are not allocated from different segmentations of the same sentence.

Four scores are reported by PyrEval: Raw score, quality, coverage and comprehensive. Given a student summary, the raw score is calculated by the sum of all matched content units with their weights. For the quality score, the raw sum is normalized by the maximum sum that the pyramid could assign to the same number of SCUs, using each pyramid SCU no more than once. The coverage score normalizes the raw score by the maximum sum the pyramid could assign given the average number of SCUs in a reference summary. The comprehensive score is the average of the quality and coverage scores.

5 Experiments and Results

5.1 Correlations with Teacher Scores

To see how PyrEval performs in an educational context, we ran PyrEval on the student summaries and compared the resulting scores to those assigned by the instructor. Five of the 136 summaries had received a perfect score of 10 from the instructor; eight additional summaries were nearly as good, each with a score of 9. These, together with a model summary written by the instructor, were used in PyrEval to generate three different pyramid content models as follows: P_1 uses a random selection of six of the thirteen highest-scoring student summaries, and P_2 uses the remaining seven. P_t consists of the five student summaries with perfect scores combined with the instructor's summary. The remaining 122 student summaries are targets to PyrEval scoring.

As shown in Table 1, the highest Pearson correlation between PyrEval scores and the instructor's scores (P_n, R) is 49%, with an average of 47%. The highest Spearman rank correlation is 46%, with an average of 45%. Pyramid model P_t does not show a significant advantage over P_1 and P_2.

Rubrics Critical thinking helps you identify potential strengths and weaknesses in the text.
SCU1 5 Critical thinking is the exercise of questioning the material, identifying its strengths and weaknesses and understanding what this changes about your knowledge.
SCU1 5 thinking Critical is crucial for academic writing , as it ensures all text read is respected and understood , analysed in depth to identify strengths and weaknesses, evaluated and used to compare to other sources of information.
SCU1 5 Critical thinking is the process of understanding , interpreting and questioning the subject at hand, identifying potential strengths and weaknesses within the text.
SCU1 5 you To effectively identify the utility of the text , must not trust what is being said; instead evaluating and extracting its strengths , weaknesses and main points.
SCU1 5 Critical thinking is about finding your strength and weaknesses in a text, evaluating summarising.

Figure 3: An SCU (SCU1, Wt=5) matched with one rubric checkpoint (See textbox in the top). The SCU format is: SCU_{index}, $Weight$, $Segment$. The segments with the same index belong to the same SCU.

5.2 Quality of Pyramid and Scoring

We examined the quality of pyramid content models built by PyrEval by comparing the 10 ideas in the rubric with high-weighted SCUs from pyramid

P_t, since it includes the instructor's summary.

Rubrics Don't take everything you read for granted. Always ask —says who, so what? what next?		
SCU5	4	for you to evaluate
SCU5	4	Critical thinking involves objective analysis and evaluation.
SCU5	4	thinking involves the evaluation of sources and the ability to extract only the useful information from it
SCU5	4	that students must assess and criticise all work to determine its potential merits and shortcomings before deciding whether to include it in their own work
SCU6	4	In thinking critically , you should always ask: says who, so what and.
SCU6	4	The work in question must be evaluated in respect to one s.
SCU6	4	Critical thinkers should always find answers to these three questions:.
SCU6	4	they have a possible bias, and are these recent Another step in critical thinking

Figure 4: Two SCUs (SCU 5, SCU 6, Wt=4) conveying the same meaning as a rubric checkpoint.

According to the rubric, a perfect score would be 10. With six reference summaries in P_t, the highest weight for an SCU is 6. The important SCUs are those with weights in $[\frac{n}{2}$,n]. There are sixteen SCUs with weights greater than 2 generated by PyrEval. Table 2 shows the distribution

Weights	6	5	4	3
Number of SCUs	1	2	4	9

Table 2: Distribution of high-weighted SCUs.

of SCUs associated with different weights. The highest score one could obtain by mentioning all important ideas is 59.

SCU2	5	you read
SCU2	5	what you know
SCU2	5	what you know
SCU2	5	what you read and relate it to other information
SCU2	5	what A text should have information
SCU8	3	what next Check
SCU8	3	What next Are there any points uncovered, critical .
SCU8	3	what is next.

Figure 5: 2 SCUs (SCU 2 and SCU 8) that are less informative.

Next, we focus on comparing the SCUs to the rubric. As seen in Figure 3, PyrEval generates some SCUs that convey the same meaning as the rubric. Figure 4 shows an example SCU from pyramid P_t that corresponds to item 1 in the rubric (cf. Figure 1).

SCU3	4	Critical thinking helps .
SCU3	4	Critical thinking is crucial for academic writing.
SCU3	4	Critical Thinking is to not take everything .
SCU3	4	To be able to think critical

Figure 6: The content in SCU 3 is not included in rubric.

There are some cases when PyrEval produces SCUs based on segments that are too short; these SCUs are actually less informative. In Figure 5, the content of SCU 2 serves as a object in sentences: *. . .a judgment should be made on if the new information has affected what you know or* . . . and SCU 8 is used as transition in the original statement.

We also identified one high-weight SCU not being matched with any checkpoints in rubric. See Figure 6.

Segment: 3 | Content Unit: 1 [Weight: 5] ***Segment:*** *..... so any strengths and weaknesses are identified through the evaluation of information and comparison with other sources*
Matched Content Unit: *Contributor*..... (1) are they cited by others within the selected discipline is to systematically identify weaknesses and strengths in the various sources ; Does Are they able to give testable verifiable evidence of *Contributor*..... (2) you identify potential strengths and weaknesses in the text *Contributor* ..... (3) Always it is important to identify the potential strengths and weaknesses from a text and to look for evidence . *Contributor* ..... (4) all text read is respected and understood , analysed in depth to identify strengths and weaknesses , evaluated and used to compare to other sources of information *Contributor* ..... (5) you To effectively identify the utility of the text , must not trust what is being said ; instead evaluating and extracting its strengths , weaknesses and main points .

Figure 7: An scoring example. The first line of top textbox indicates segment 3 from student summary matched with content unit 1 in pyramid, with weight as 5. The second line is the text of student summary segment. The textbox in the bottom shows a matched content unit from pyramid model composed by 5 contributors, denoted as $Contributor, \ldots, Index, Content$.

Finally, Figure 7 shows a match between a segment from a student summary and an SCU. The content in this SCU also corresponds to checkpoint 9 in rubric, as shown in Figure 3.

5.3 Revising the SCUs

The observations mentioned above lead us to question whether some type of post-processing on the

Removal	Pearson (P-v)	Spearman (P-v)
(2, 8)	49.37 (7.52e-09)	48.75 (1.24e-08)
(2, 3, 8)	53.39 (2.40e-10)	52.89 (3.80e-08)

Table 3: Pearson correlations and Spearman correlation of PyrEval scores with teachers' scores after removing the problematic SCUs. SCUs (2, 8) are the uninformative SCUs; Adding SCU 3 includes an irrelevant SCU.

Method	Below Avg	Avg	Above Avg	Total
H	21	76	25	122
P	27	75	20	122
Overlap	8	50	10	68

Table 4: Agreements between H - human using rubric, and P - PyrEval

pyramid models would improve the correlation scores. To test this supposition, we manually removed the three uninformative high-weighted SCUs identified above, and ran the scoring based on the resulting adjusted pyramids.

Table 3 shows that both Pearson and Spearman correlations are improved after removal of uninformative SCUs (49%), or both uninformative and irrelevant SCUs (53%). These slight increases suggest that post-processing, such as removing irrelevant and uninformative SCUs using entropy, could help to improve the quality of a pyramid.

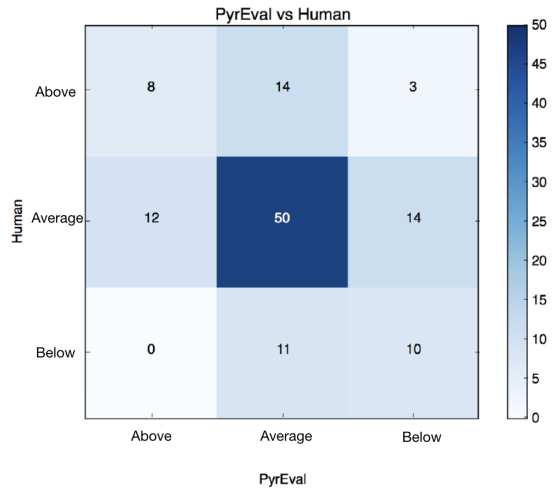

Figure 8: Confusion matrix of disagreements and agreements between human using rubric and PyrEval. Horizontal axis represents PyrEval and vertical axis represents human evaluation.

We took another approach by binning the scores into three ranges: below average, average and above average. Table 4 presents two distributions obtained from both the human and PyrEval scores that are almost identical, and the agreements between two sets of scores. The human and PyrEval scores identify 21 versus 27 student summaries as below average, 25 versus 20 as above average. There are 76 summaries marked as around average by human and 75 by PyrEval. However, among 122 summaries, 68 of these (over 55%) overlap in terms of where they fall in these newly defined categories. Both agreements and disagreements are distributed as shown in Figure 8. In the extreme disagreements, none of the summaries judged as below average by human are evaluated as above average by PyrEval. Additionally, only three summaries PyrEval regards as below average are considered above average by human. PyrEval and educators easily agree on summaries that fall within the medium range, but tend to disagree on both below average summaries and above summaries.

6 Potential Uses and Developments

The three different pyramids returned very similar Pearson and Spearman correlation coefficients. Although they all indicated a moderately positive relationship between the human and PyrEval scores, the similarity in their values led us to consider a different approach for examining the relationships.

The above classification demonstrates how PyrEval could be used accurately to distinguish between good and bad student summaries. In other words, it is highly unlikely that summaries judged to be below average by a human scorer would be regard as above average by PyrEval, and vice versa. As such, the three groupings - below average, average and above average - provide scope for filtering submissions being uploaded to an online repository as follows. Summaries in Group C (below average) are rejected outright, with feedback on what needs improving; those in Group B (average) are accepted and scored by PyrEval but, in addition, given some indication on how the score could be improved; those in Group A (above average) are accepted as ready to be hand scored by the instructor.

What needs to be addressed next is the type of feedback PyrEval might provide each summary, and how. It is possible for the tool to list details of

SCUs missing from the summary, thereby providing the opportunity for students to improve their work. This would make Pyreval very effective as a formative feedback tool, especially if the revised summaries were then resubmitted and checked via the same process. A future project could involve devising a way to provide students with text-based feedback, aimed at helping them address specific areas of concern relating to missing content.

PyrEval's potential for advancing student learning is not limited to helping students write better. It could also be used in ways that significantly cut down on the amount of marking instructors have to do. Using the classification above could mean that papers in Group A are hand scored by the instructor, with an assurance that such papers will include a high percentage of all of the ideas present in the rubric. In certain situations, depending on the assessment criteria, high quality submissions might not need to be hand scored at all. Similarly, those in Group C could be rejected outright, with feedback on how to the text should be improved. Those in Group B could be accepted with a warning about the maximum score attainable, say 70 percent. There could also be an opportunity for the summary to be improved and resubmitted.

There is need to examine the three summaries which the human scorer rated above average but PyrEval classed as below average. Reading these texts over, this time checking for clues that could shed light on the discrepancies, revealed that the human scorer was lenient in all three cases. The reassessment showed that these papers were particularly well written (fluent), even though they did not strictly meet the requirements of the assignment. Reading them might have brought some relief to the human scorer; for example, following a spate of poorly written summaries. It is therefore possible that extra effort was made to match sections of these text to the checkpoints in the rubric, albeit that these matches were not warranted. Human are susceptible to emotion and fatigue, which can in turn affect their scoring behavior while automated scoring will be consistent.

7 Conclusion

The present research project extends current knowledge about the uses of NLP in building educational applications by discussing PyrEval as a formative assessment tool. The discovery of a new typology has enabled us to begin to understand how student self-directed learning could be developed and, indeed, measured. This could have a direct impact on the assessment practices and policies within institutions and, ultimately, on increasing retention and progression in university courses.

A long-term goal is to develop a web-based application, which uses PyrEval to provide formative assessment feedback on student summaries. The ultimate aim is to extend the thematic scope of the research to include other courses, particularly STEM.

Acknowledgments

This work was partly supported by Penn State University's Teaching and Learning with Technology.

References

Leslie M. Babinski, Steven J. Amendum, Steven E. Knotek, Marta Sánchez, and Patrick Malone. 2017. Improving young english learners language and literacy skills through teacher professional development: A randomized controlled trial. *American Educational Research Journal*, 55(1):117–143.

Robert B Barr and John Tagg. 1995. From teaching to learninga new paradigm for undergraduate education. *Change: The magazine of higher learning*, 27(6):12–26.

Stephen J Bostock. 1998. Constructivism in mass higher education: a case study. *British journal of educational technology*, 29(3):225–240.

Anne Brockbank and Ian McGill. 2007. *Facilitating reflective learning in higher education*. McGraw-Hill Education (UK).

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750. Association for Computational Linguistics.

Yanjun Gao, Andrew Warner, and Rebecca J. Passonneau. 2018a. Pyreval: An automated method for summary content analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC) 2018*.

Yanjun Gao, Andrew Warner, Chen Sun, and Rebecca J. Passonneau. 2018b. Emergent discovery of content units in summaries. In submission.

Steve Graham and Dolores Perin. 2007. A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99(3):445–476.

Weiwei Guo, Wei Liu, and Mona T Diab. 2014. Fast tweet retrieval with compact binary codes. In *COLING*, pages 486–496.

R. Hiemestra. 1994. Self-directed learning. In T. Husen and T. N. Postlethwaite, editors, *The International Encyclopaedia of Education*, 2nd edition. Pergamon Press, Oxford.

Fiona Hyland. 2000. Esl writers and feedback: Giving more autonomy to students. *Language teaching research*, 4(1):33–54.

Fatma Susar Kırmızı. 2009. The relationship between writing achievement and the use of reading comprehension strategies in the 4th and 5th grades of primary schools. *Procedia - Social and Behavioral Sciences*, 1(1):230–234.

Susan J Lea, David Stephenson, and Juliette Troy. 2003. Higher education students' attitudes to student-centred learning: beyond 'educational bulimia'? *Studies in higher education*, 28(3):321–334.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, pages 74–81. Barcelona, Spain.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Ani Nenkova, Rebecca J. Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics ACL*, page 311318.

Rebecca J. Passonneau, Ananya Poddar, Gaurav Gite, Alisa Krivokapic, Qian Yang, and Dolores Perin. 2016. Wise crowd content assessment and educational rubrics. *International Journal of Artificial Intelligence in Education*, pages 1–27.

Dolores Perin, Rachel Hare Bork, Stephen T Peverly, and Linda H Mason. 2013. A contextualized curricular supplement for developmental reading and writing. *Journal of College Reading and Learning*, 43(2):8–38.

Antje Proske, Susanne Narciss, and Danielle S. McNamara. 2012. Computer-based scaffolding to facilitate students' development of expertise in academic writing. *Journal of Research in Reading*, 35(2):136–152.

Phil Race. 2001. A briefing on self, peer and group assessment. *LTSN generic centre assessment guides series*.

Rod D. Roscoe, Laura K. Allen, Jennifer L. Weston, Scott A. Crossley, and Danielle S. McNamara. 2015. The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, 34:39–59.

D Royce Sadler. 2010. Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5):535–550.

Shuichi Sakai, Mitsunori Togasaki, and Koichi Yamazaki. 2003. A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Applied Mathematics*, 126(2):313–322.

Tamara Sladoljev Agejev and Jan Šnajder. 2017. Using analytic scoring rubrics in the automatic assessment of college-level summary writing tasks in L2. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 181–186, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Sargur Srihari, Jim Collins, Rohini Srihari, Harish Srinivasan, Shravya Shetty, and Janina Brutt-Griffler. 2008. Automatic scoring of short handwritten essays in reading comprehension tests. *Artificial Intelligence*, 172(2):300 – 324.

Yao-Ting Sung, Kuo-En Chang, and Jung-Sheng Huang. 2008. Improving childrens reading comprehension and use of strategies through computer-based strategy training. *Computers in Human Behavior*, 24(4):1552–1571.

Charissa Tansomboon, Libby F. Gerard, Jonathan M. Vitale, and Marcia C. Linn. 2017. Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, 27(4):729–757.

Paul A Tess. 2013. The role of social media in higher education classes (real and virtual)–a literature review. *Computers in Human Behavior*, 29(5):A60–A68.

Yu-Fen Yang. 2014. Preparing language teachers for blended teaching of summary writing. *Computer Assisted Language Learning*, 27(3):185–206.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578, Vancouver, Canada. Association for Computational Linguistics.

Kaja Zupanc and Zoran Bosni. 2017. Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, 120:118 – 132.