Automatic Assessment of Student Reading Comprehension from Short Summaries

Lisa Mintz
University of Memphis
Institute for Intelligent Systems
365 Innovation Drive
Memphis, TN 38152
lmintz@memphis.edu

Dan Stefanescu
University of Memphis
Institute for Intelligent Systems
365 Innovation Drive
Memphis, TN 38152
dstfnscu@memphis.edu

Shi Feng
University of Memphis
Institute for Intelligent Systems
365 Innovation Drive
Memphis, TN 38152
sfeng@memphis.edu

Sidney D'Mello
University of Notre Dame
Notre Dame
IN 46556
sdmello@nd.edu

Arthur C. Graesser
University of Memphis
Institute for Intelligent Systems
365 Innovation Drive
Memphis, TN 38152
graesser@memphis.edu

ABSTRACT
This paper describes our research on automatically scoring students' summaries for comprehension using not only text specific quantitative and qualitative features, but also more complex features based on the computational indices of cohesion available via Coh-Metrix and on Information Content (IC, a measure of text informativeness). We assessed whether human rated summary scores could be predicted by indices of text complexity and IC. The IC metric of the summaries was a better predictor of human scores than word count or any of the Coh-Metrix text complexity dimensions. This finding may justify the implementation of IC in future automated summary rating tools to rate short summaries.

Keywords Information Content, summarization, reading

1. INTRODUCTION
Summarizing content after reading a text is a well-established method of assessing comprehension [1]. Assessing students' reading comprehension through summarization has many advantages over other methods, because summarization requires readers to actively reconstruct their mental representation of the text [2]. The purpose of the current study is to examine a method of automated summary grading using a small corpus of summaries written for a variety of texts. We explored the use of the computational linguistic tool Coh-Metrix [3], as well as informativeness of words to predict human rater's scores of summaries.

Coh-Metrix is a computational linguistic tool developed to measure hundreds of indices related to syntactic complexity, text cohesion, lexical diversity, and other features of language and discourse [3]. Coh-Metrix's five major dimensions of text complexity predict a number of psychological findings associated with comprehension, such as reading time and recall [4]. In this study we used CohMetrix to measure: *narrativity*, *syntactic simplicity*, *word concreteness*, *referential cohesion*, and *deep cohesion* (http://cohmetrix.memphis.edu) [4].

Information Content (IC) is a measure used by Resnik [5] to compute the informativeness of a concept in a hierarchical taxonomy such as WordNet [6]. IC relies on the assumption the informativeness of a concept is inversely dependent on its occurrence frequency: the more frequent a concept, the less informative it is. Resnik [5] computes the frequency of a concept c as the sum of the occurrence frequencies of the words defining the concept c and all the other words defining the subordinates. Once the occurrence frequency of a concept is defined, the IC value for each concept c is computed as the self-information measure of c:

$$IC(c) = \log\left(\frac{1}{P(c)}\right) = -\log(P(c)) - log\frac{\#c}{\sum_i \#c_i}$$

A method for transferring the IC values from concepts to words has been proposed [7]: a word is assigned the IC value corresponding to the most general concept that word can represent, which is the concept with the minimum IC value:

$$IC(w) = \min_{c|w \in c} IC(c)$$

This ensures that high IC values are only associated with informative words. We compute the IC of a text fragment as the sum of the IC values for the individual words occurring in that fragment. The resulting sum value can be used as a measure of informativeness of the entire text, or it can be normalized by the total number of words in that text. We experimented with both methods.

In this study, we asked human experts to rate a total of 225 summaries written after reading texts from different genres. Our goal was to use Coh-Metrix dimensions of text complexity and IC computed from WordNet to predict the human ratings beyond simple verbosity (word count). If successful, such an approach will allow us to estimate summary qualities without a gold standard or a large summary corpus. Thus, our algorithm would contribute to assessing summaries written for a variety of subject matters and text types.

2. METHOD
Seventy-five undergraduates from the University of Memphis participated in this study. We collected 73 texts of different genres

on different topics, containing between 1000 and 1500 words (M = 1301.3, SD = 186.0). There were 24 Informational, 24 persuasive and 25 Narrative texts selected from various websites on the internet. The texts were measured on different levels of textual complexity and Flesch-Kincaid readability. The texts were each separated into multiple pages (screens) of 75-100 words each, keeping the original paragraphs and always ending on a sentence. Each participant read three texts, one from each genre. Each text was randomly selected from each genre. After reading each text, participants wrote a 75 to 100 word summary of the text that they just read. Thus, each participant wrote 3 summaries, one per genre. Three expert raters independently rated the summaries on a 1-4 scale for comprehension. Chronbach's alpha scores suggested high inter-rater agreement of α = .802 (N = 225).

4. RESULTS AND DISCUSSION

A Pearson-correlation analysis was conducted between the summary rating score, word count, and IC. Word count was included because previous research suggests a strong positive relationship between word count and perceived quality of writing. IC strongly correlated with word count, r = .952, n = 224 p < .001. Word count and summary score were strongly correlated, r = .562, n = 224, p < .001, with an r^2 of .316. A linear regression revealed that approximately 31.3% of the variance in comprehension score can be accounted for by the variance in word count (β = .559, SE = .001. F = 100.635, p < .0001). We found a strong correlation between IC and summary score, score, r = .617, n = 224, p < .001, with an r^2 of .377. A linear regression revealed that approximately 37.7% of the variance in comprehension score could be accounted for by the variance in IC (β = .614, SE = .004, F = 133.715, p < .0001). We found that IC explained 5.5% more of the variance in comprehension score than word count.

It was interesting to note that Coh-Metrix's dimension of deep cohesion was significantly correlated with IC (r = .22, n = 224 p < .01), but not with word count (r = .078, n = 224, p = .248). However, a multiple regression using word count and deep cohesion as predictors did not show significant contribution of deep cohesion as a predictor. The result suggests that although IC is highly correlated with word count, it is a better predictor of comprehension than word count, which suggests that summary scores are more than mere summary length. In the future, IC could possibly be implemented in automated summary grading tools to increase their accuracy in scoring summaries.

Table 1. Correlations between summary score, IC, word count and Coh-Metrix's indices of text complexity

Measure	Comprehension	IC	Word Count
Summary score	-		
IC	.617**	-	
Word Count	.562**	.952**	-
Deep Cohesion	.121	.222*	.078
Referential Cohesion	.001	.057	-.103
Syntactic Simplicity	-.045	-.008	-.180*
Word Concreteness	.019	.099	-.076
Narrativity	.006	.095	-.056
p<.01* p<.001**			

5. CONCLUSION

In this study we attempted to use IC and the five dimensions of text complexity from Coh-Metrix to predict human ratings of summaries. Our results showed that surprisingly, the five dimensions of text complexity did not predict human ratings of comprehension from summarization. On the other hand, although IC was also highly correlated with word count, it explained more of the variance in comprehension score than word count. In future research we will explore using other linguistic indices as well as IC to predict summary scores on a larger corpus of summaries.

6. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (ITR 0325428, HCC 0834847, DRL 1235958) and Institute of Education Sciences (R305C120001). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. Our thanks to Eliana Silbermann, Megan Reed, Janay Stewart, and Mari Sanford for their assistance.

7. REFERENCES

[1] Madnani, N., Burstein, J., Sabatini, J., & O'Reilly, T. (2013). Automated Scoring of a Summary Writing Task Designed to Measure Reading Comprehension. In *Proceedings of the North American Association for Computational Linguistics Eighth Workshop Using Innovative NLP for Building Educational Applications* (pp. 163). Atlanta, Ga.

[2] Graesser, A. C., Wiemer-Hastings, P., & Wiemer-Hastings, K. (2001). Constructing inferences and relations during text comprehension. *Text representation: Linguistic and psycholinguistic aspects*, *8*, 249-271.

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

[4] Graesser, A.C., McNamara, D.S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223-234.

[5] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*. (IJCAI-95).

[6] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Cambridge, Mass.: MIT Press.

[7] Stefanescu, D., Banjade, R., Rus, V. (2014). *A Sentence Similarity Method based on Parsing and Information Content.* In Proceedings of CICLing 2014, Kathmandu, Nepal.