Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 83 (2013) 151 – 155

Procedia
Social and Behavioral Sciences

2nd World Conference on Educational Technology Researches – WCETR2012

Exploring the Assessment of Summaries: Using Latent Semantic Analysis to Grade Summaries Written by Spanish Students

León, J.A. [a] *, Olmos, R. [a], Escudero, I. [b] , Jorge-Botana, G. [b] & Perry, D. [c]

[a] *Universidad Autónoma de Madrid, Spain*
[b] *UNED, Madrid, Spain*
[c] *Universitat Politécnica de València, Spain*

Abstract

In this study we propose an integrated method to automatically assess summaries using LSA. The method is based on a regression equation calculated with a corpus of a hundred summaries (the training sample), and is validated on a different sample of summaries (the validation sample). The equation incorporates two parameters extracted from LSA: semantic similarity and vector length. A total of 396 students drawn from four stages of education participated in the study. The summaries of a short narrative text written by each participant were evaluated on a scale of 0-10 by four human graders and the scores compared to the evaluation of the summaries using LSA. The results supported that incorporating both parameters into the method resulted more successful than the traditional cosine measure, and that LSA showed a similar level of sensitivity to the quality of the summaries produced in different academic stages as that shown by the human graders.

Selection and/or peer-review under responsibility of Prof. Dr. Hafize Keser Ankara University, Turkey

Keywords: Summaries, Latent Semantic Analysis, Automatic assessment, Spanish students;

1. Introduction

Latent Semantic Analysis (LSA) is a computational technique that contains a mathematical representation of language. During the last twenty years its capacity to simulate aspects of human semantics has been widely demonstrated (e.g., Hu, Cai, Wiemer-Hastings, Graesser & McNamara, 2007; Landauer & Dumais, 1997; Olmos, León, Jorge-Botana, & Escudero, 2009). LSA is based on three fundamental ideas: (1) to begin to simulate human semantics of language we first obtain an occurrence matrix of terms contained in a document, (2) the dimensionality of this matrix is reduced using singular value decomposition, a mathematical technique that effectively makes the tool a latent semantic space, and (3) any word or text is represented by a vector in this new latent semantic space. An LSA application is based on automatic assessors (Foltz, Laham & Landauer, 1999). One very common method is when the source text consists of an expert summary (normally written by a grader or teacher), thus creating what is

* Corresponding author: Jose A. Leon. Tel.: +34-91-4975226
E-mail address: joseantonio.leon@uam.es

1877-0428 © 2013 The Authors. Published by Elsevier Ltd.
Selection and/or peer-review under responsibility of Prof. Dr. Hafize Keser Ankara University, Turkey
doi:10.1016/j.sbspro.2013.06.029

called a “golden summary” (León, Olmos, Escudero, Cañas & Salmerón, 2006). These tools automatically provide an essay score. Reliability tests for LSA can give surprisingly higher results (e.g. above .80), suggesting it can be just as reliable as expert human judges; i.e. trained graders or teachers.

The aim of this study was to use a simple, innovative LSA-based computational method to reliably evaluate summaries which are especially brief (approximately 50 words). The method incorporated the essential information from the latent semantic space: (1) a measure of semantic similarity using the cosine and (2) a measure of the vector length or extent of knowledge about the text. The method was tested on a sample of 396 students, at four different stages of education, who were required to read a narrative text and then write a summary. The accuracy of the LSA evaluations was compared to those of four specially trained judges who also evaluated each of the summaries. Within the general aim we sought three goals. First, to obtain reliable evaluations (at least > 0.70) combining both kinds of essential information (cosine and vector length) derived from the latent semantic space. Second, to show that LSA is as sensitive to the quality of summaries written by students at different academic levels as trained judges are. Third, to overcome possible limitations of working with such brief texts and show that LSA works with highly conceptualized summaries.

2. Method

To implement our method we used a database comprising 107 summaries of narrative text distributed across four grade levels. The sample used to adjust the method is called the training sample, This sample allows us to calculate the way we obtain the scores with LSA although it is not used to evaluate the reliability of the method. Each of these 107 summaries was graded independently by each of the four judges on a scale of 0 to 10. This scale included up to four points for content and up to six points for coherence of the summary. Blind scoring was used; in other words the graders were unaware of the students' academic level. An average score was obtained from the four graders' scores.

The two LSA measures, vector length and semantic similarity, were obtained as follows. Given that in LSA each document is represented by a vector, the vector length is simply calculated as the length of each summary vector. In the equation the vector length component represents how detailed the summary is - the greater the vector length the more detail, and the more familiar or relevant words appear in the semantic space. The measure of similarity is somewhat more difficult to obtain. For this reason we used a well-known method, habitually used in automated essay scoring with LSA: the Summary–Expert Summaries Method (Dikli, 2006; Foltz et al., 1999; Kintsch et al., 2007; León et al., 2006; Landauer & Dumais, 1997; Olmos et al., 2009). To obtain a measure of similarity, the summary is compared to a source text. This method uses 'golden' or 'ideal' summaries; that is, summaries written by experts which contain the essential information from the text and have very strong coherence. To this end, six teachers with expertise in comprehension were asked to write a summary of no more than fifty words. To obtain a measure of semantic similarity the cosine between the student summaries and each expert summary was calculated. Since there were six cosines, one for each expert summary, the average cosine was taken as the final measure of semantic similarity. In this way, vector length and semantic similarity values were obtained automatically for all 396 summaries.

2.1. Procedure and design

2.1.1. The Spanish LSA corpus

The generalist corpus contains material from on-line encyclopedias, newspapers, textbooks and several Internet sources. In total the corpus has 2,059,234 documents (i.e. paragraphs) and 1,661,954 different terms. A semantic space with 337 dimensions was used.

2.1.2. Material

A 402-word narrative text (The Legend of the Carob Tree) was used. No specialist background knowledge is required to understand this text.

2.1.3. Participants

396 students from four stages of education took part in this study. Student ages ranged from 10 to 23 years. The youngest group comprised 119 students from 6th grade, followed by 98 students from 8th grade, 100 students from 10th grade, and finally 79 undergraduate university students.

2.1.4. Procedure

Each student read the text at their own pace in a classroom. Before reading it they were told that it was important to understand the text in order to respond to a series of questions. After reading, students were allowed 15 minutes to write the summary.

2.1.5. The four judges' evaluations

Four PhD students were given four training sessions in evaluating summaries using a scale of 0 to 10. After training they graded the summaries on two main criteria. First the content of the summaries was evaluated on a scale of 0 (no content) to 4 (all key content). The text contained four main ideas that should be included in any summary of it (see León et al., 2006). Each main idea counted as one point. Secondly, coherence was evaluated on a scale of 0 (incoherent) to 6 (highly coherent). To assess coherence, the organization, causal relationships, use of connectives, extent of conceptualization of the summary and lack of redundancy were analyzed. The judges carried out the evaluations independently and without knowing the students' academic level.

2.1.6. Data analysis

The data were analyzed in the reliability between LSA and the judges was calculated using the validation sample, and an ANOVA was used to ascertain the sensitivity of the judges and LSA to differences in the quality of summaries from the different groups of students

3. Results

3.1. LSA-grader reliability

The LSA-grader reliabilities were calculated using a validation sample which consisted of 289 summaries of the narrative text. The LSA grades for this new sample of summaries were obtained as described in the section on method. This sample was set aside to avoid overfitting of reliabilities, and thus allow generalization of results to other summaries.

Table 1 shows the reliability of LSA compared to the score given by each judge and to the average judges' score (calculated with Pearson's correlation). The reliability of LSA ranged from .60 to .67 for the individual judges, and reached .68 for the average judges' score. These scores may be considered as fairly high.

Table 1. LSA-grader reliability of narrative text and human grader (**) Significant at the 1% level

	Grader				
Text	Grader 1	Grader 2	Grader 3	Grader 4	Average grader
Narrative text	.61(**)	.67(**)	.60(**)	.63(**)	.68(**)

3.2. Sensitivity to differences between grade levels

An ANOVA was carried out on to study the capacity of the human judges and LSA to detect differences in the quality of summaries written by students from the different grade levels (6th, 8th, 10th and university). Figure 1 shows the average human and LSA scores awarded to the summaries at each of the four grade levels. It can be seen that both the human judges and LSA detected differences in the quality of the summaries ($F(3,285) = 36.60$, $p < .05$ and $F(3,285) = 10.96$, $p < .05$ respectively), however, a post hoc test revealed that the human judges differentiated between three groups while LSA only two. For the judges, the summaries written by the university undergraduate group scored the highest (that is, were judged as being of better quality). These were followed by the summaries written by the 10th grade while the third (and lowest quality) group was made up of the summaries written by the 6th and 8th grade students (no statistically significant difference was found between these last two). LSA on the other hand, detected two groups of averages: again, the best summaries were those written by the university students while the other grade levels formed another group of lower-scoring summaries (no statistically significant differences were detected between them).

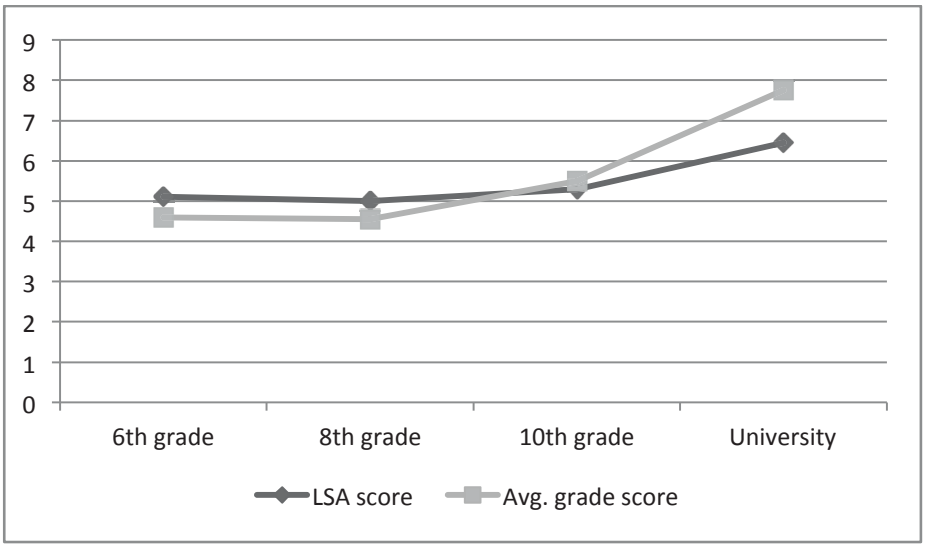

Figure 1. LSA and graders' mean score at each grade level for narrative text

Both the human judges and LSA detected differences in the quality of summaries ($F(3,293) = 89.97$, $p < .05$, and $F(3,293) = 47.88$, $p < .05$ respectively). This time the *post hoc* test (T-Student) showed that both the judges and LSA detected two groups of averages: the undergraduate group (with higher scoring summaries) and the remaining three grade levels (no statistically significant difference was found between their averages).

4. Conclusions

LSA has become one of the most widely-used computational tools of recent years, and one of the fastest-growing areas of application has been the field of education. Today, LSA is already a reality in some U.S. classrooms and it is gradually finding its way into more and more schools as a means of helping to improve students' writing and comprehension strategies (Dikli, 2006; E. Kintsch et al., 2007). Our research indicates that incorporating information on vector length together with semantic similarity provides a substantial improvement when using LSA to evaluate highly conceptualized summaries. The availability of automatic tools that evaluate reliably, help to

detect weak or strong points in the summaries, detect good and bad student strategies or capture the macrostructure and overly local information in texts, may take pressure off teachers, and at the same time provide the student with immediate and useful feedback on this type of task. However, this tool does require improvements. For example, it should be complemented with new algorithms to overcome some of its limitations (Kintsch, 2002), the use of the latent semantic space should be mathematically optimized (Hu et al., 2007), and it should be linked with psychological models such as semantic memory (Denhière, Lemaire, Bellissens & Jhean-Larose, 2007) or with other computational models of language (Steyvers & Griffiths, 2007). Once all of these contributions are added, the potential and the capability of LSA in the educational sector will be far greater.

Acknowledgements

This work was supported by Grant SEJ2006-09916 from the Spanish Ministry of Science and Technology and PSI 2009-31932 by the Spanish Ministry of Education.

References

Denhière, G., Lemaire, B., Bellissens, C. & Jhean-Larose, S. (2007). A Semantic Space Modeling Children's Semantic Memory. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *The Handbook of Latent Semantic Analysis* (pp. 143- 167). Mahwah, NJ: Erlbaum.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment, 5*(1). Retrieved from http://www.jtla.org.

Far, R., Pritchard, R. & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27,* 209-226.

Foltz, P. W., Laham, D. & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1.* Retrieved June 29, 2004, from http://knowledgetechnologies.com

Hu, X., Cai, Z., Wiemer-Hastings, Graesser, A. C. & McNamara, D.S. (2007). Strengths, limitations, and extensions of LSA. In T.K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *The Handbook of Latent Semantic Analysis* (pp. 401- 426). Mahwah, NJ: Erlbaum.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N. & Dooley, S. (2007). Summary Street: Computer-Guided Summary Writing. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.). *The Handbook of Latent Semantic Analysis* (pp. 263- 277). Mahwah, NJ: Erlbaum.

Landauer, T. K. & Dumais, S. T. (1997) A solution to Plato's problem: the Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

León, J. A., Olmos, R., Escudero, I., Cañas, J. J. & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and Latent Semantic Analysis in narrative and expository texts. *Behavior Research Methods, Instruments and Computers 38* (4), 616–627.

Olmos, R., León, J. A., Jorge-Botana, G. & Escudero, I. (2009). New algorithms assessing short summaries in expository texts using Latent Semantic Analysis. *Behaviour Research Methods, Instruments, and Computers, 41,* 944-950.

Steyvers, M. & Griffiths, T. (2007). Probabilistic Topic Models. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The Handbook of Latent Semantic Analysis* (pp. 427-448). Mahwah, NJ: Erlbaum.