Towards a Method for Assessing Summaries in Spanish using LSA

René Venegas

Pontificia Universidad Católica de Valparaíso, Chile
Av. Brasil 2830, piso 9, Valparaíso, Chile
rene.venegas@ucv.cl

Abstract

One of the most important goals in Intelligent Tutoring is to create applications that can evaluate the quality of a text in a human-like manner. The aim of this study is to compare three methods of using Latent Semantic Analysis (LSA) to evaluate the quality of summaries written by students in Spanish. The sample is made up by 226 summaries written by Chilean students based on both expository and narrative texts. Each summary was first assessed by human judges in order to compare the results with the scoring provided by three different LSA methods: a) comparing the summaries with the original text divided in paragraphs, b) comparing the summaries with the text as one unit, and c) comparing the summaries with other summaries written by four human experts. Results show that comparison between each student's summary and the text as a one unit constitutes the method which most closely resembles human evaluation.

Introduction

Recent studies have revealed that Latin American students have reading comprehension problems (Peronard et al. 1998; PISA 2007). In this study, we assess reading comprehension through one of its products, summaries. Summaries are a useful tool in the evaluation of reading comprehension as they represent the global semantic content of a source text (Van Dijk 1983). Through a summary it is possible to track a comprehender's inferential processes during the reading comprehension task. At present, some computer programs can simulate the behavior of human experts when assessing written texts. One such program uses Latent Semantic Analysis (LSA) for the assessment of summaries written in English (Landauer et al. 2007). So far, the applications for assessment of summaries written in Spanish based on LSA have been very few (Pérez et al. 2005; León et al. 2005; Venegas 2007). Therefore, the aim of this study is to compare three methods of using LSA to evaluate the quality of summaries written by students in Spanish.

Methods

In a first stage, 226 summaries were selected. These summaries were written by students from vocational schools (≅16 years of age). Those summaries (average 80 words) were scored by a team of linguists according to a 30 point scale (Parodi 2007). The evaluation criteria included the presence and quality of core ideas from the source texts that were included in the summary. Next, a relative threshold was set, namely 60%, which allowed for the division of the set of summaries into two groups: 98 summaries were classified as high quality achievement (≥ 18 points) and 128 were classified as low quality achievement (≤ 17 points). Details according to the predominant discourse organization mode in the source text are shown in Table 1:

Discourse modality	High achievement	Low achievement	Total
Expository	41	96	137
Narrative	57	32	89
Total	**98**	**128**	**226**

Table 1. Total number of scored summaries.

In the second stage, the same summaries were automatically scored with LSA, according to three methods: a) by comparing each summary with the original text divided into paragraphs, b) by comparing each summary with the text as one unit, and c) by comparing each summary with other summaries written by four experts (experienced Spanish instructors). In order to extract the semantic similarity, a semantic space of 297 dimensions was built. This semantic space was constructed by using a thematic diversified corpus (over 10 million words). This decision is justified by findings which show that topic dependency may hinder the efficiency of automatic assessment using LSA (Olmos et al. 2009). Next, using the cosine values obtained through each method we carried out a supervised binary classification. For each method, a relative threshold to the maximum cosine scoring (60%) was determined, according to which summaries were divided between high achievement and low achievement and then compared to the hand-scoring

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
* Founded by Fondecyt 11070225

results. Finally, precision, recall and F1-measure values were determined for each method.

Results

The results of the comparison of the three methods will be presented in two stages:

A) Comparison of the three methods without considering discourse organization mode of source texts. Table 2 shows the results that we obtained:

Comparison between methods	Method 1 Text divided in paragraphs	Method 2 Text as a unit	Method 3 Summaries written by experts
Threshold* (60%)	.42	.52	.48
High achievement (n=98)	34.7%	82.7%	75.50%
Low achievement (n=128)	77.30%	40.6%	50%
Recall	.35	.83	.76
Precision	.54	.52	.54
F1-measure	.42	.64	.63

Table 2. Comparison between the three methods without considering discourse organization mode. *Values are cosines.

According to our results, the second method is the one that most closely resembles (82.7%) evaluation by human judges when assessing a summary as of high quality. This result is confirmed by the high balance between precision and recall (F1= .64). On the other hand, the first method is the one which most closely resembles human assessment when judging a summary as of low quality, although the F1 value is very low.

B) Comparison of the three methods according to the discourse organization mode of source texts. As shown in Table 3, in the case of expository texts the second method is the one in which there is the highest level of agreement between human scoring and automatic scoring in terms of assessing a summary as of high quality.

Comparison	**Method 1**		**Method 2**		**Method 3**	
discourse organization mode	**Exp**	**Narr**	**Exp**	**Narr**	**Exp**	**Narr**
Threshold (60%)	.41	.31	.53	.52	.48	.40
High achievement n (exp)=41; n (narr)=57	22%	82.5%	90.2%	77.2%	87.8%	93%
Low achievement n(exp)=96; n(narr)=32	88%	12.5%	41.7%	43.8%	43.8%	25%
Recall	.22	.82	.90	.77	.88	.93
Precision	.43	.63	.40	.71	.40	.69
F1-measure	.29	.71	.55	.74	.55	.79

The third method yields similar results to the second method in all measures. In contrast, the first method yields a higher level of agreement between human scoring and automatic scoring when assessing summaries as of low quality (88%). However, it yields a very low value in F1. In the case of narrative texts, the third method shows the highest level of agreement between automatic scoring and hand scoring when judging summaries as of high quality (93%). The third method also yields the highest F1 value. However, all three automatic methods show a high level of agreement with evaluation by human judges. Finally, the second method shows the highest level of agreement between both forms of scoring when judging summaries as of low quality (43.8%).

Conclusion

When assessing summaries in Spanish, the LSA method of automatic scoring which most resembles evaluation by human judges is Method 2. This result is consistent with the idea that the overall semantic content of the source text should be considered when assessing a summary. Methods 2 and 3 yield similar results when the source text for the summary is expository. On the other hand, Method 3 yields the highest values when the source text is narrative. However, from the point of view of natural language processing (NLP) it is not an efficient method as it requires the participation of human experts to write the summaries. On the other hand, Method 2 may be easily implemented using a computer program that determines semantic similarity.

References

Landauer, T., McNamara, D., Dennos, S., and Kintsch, W. (eds). 2007. *Handbook of Latent Semantic Analysis.* N. J.: Erlbaum.

León, J., Olmos, R., Escudero, I., Cañas, J., and Salmerón, L. 2005. Assessing short summaries with human judgments procedure and Latent Semantic Analysis in narrative and expository texts. *Behavior Research Methods* 38(4): 616-627.

Olmos, R., León, J., Escudero, I., and Botana 2009. Efectos sobre el tamaño y especificidad de los corpus en la evaluación de resúmenes mediante el LSA y jueces expertos. *Revista Signos* 41(69): 71-81.

Parodi, G. 2007. *Lingüística de corpus y discursos especializados: puntos de mira*. Valparaíso: EUVSA.

Pérez, D., Alfonseca, E., Rodríguez, P., Gliozzo. A., Strapparava, C. and Magnini, B. 2005. About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment. *Revista Signos* 38(59), 325-343.

Peronard, M., Gómez, L., Parodi, G., and Núñez, P. 1998. *Comprensión de textos escritos: De la teoría a la sala de clases.* Santiago Andrés Bello.

Pisa 2007. PISA 2006. *Science Competencies for Tomorrow's World*. Paris: OECD Publishing.

Van Dijk, T. (1983). *La ciencia del texto. Un enfoque interdisciplinario*. Buenos Aires: Paidos.

Venegas, R. 2007. Using Latent Semantic Analysis in a Spanish research article corpus. In G. Parodi (ed.). *Working with Spanish corpora* (pp. 195- 216). London: Continuum.