

Using Latent Semantic Analysis vs. Human Judgements assessing short summaries in expository texts

Ricardo Olmos Albacete, José A. León, Guillermo Jorge-Botana, Universidad Autónoma de Madrid
E-mail: ricardolmos@inicia.es, joseantonio.leon@uam.es, jorgeybotana@psi.ucm.es

Abstract: In this study we compare four expert graders with LSA to assess short summaries taken from Spanish middle grade students in an expository text. In order to improve the reliability between LSA human graders, we analyze three new algorithms into two holistic methods using in a previous research (León et al. 2006). The new three algorithms were (1), an adaptation of predication algorithm (Kintsch, 2001), (2), a dimension measure that include the 80% best dimensions (Hu et al., 2007), and (3), Euclidean distance measure (Rehder et al., 1998). The results supported more reliability of LSA as a computerized assessment tool in expository text used to develop web-based e-learning platforms and graders.

Introduction

Latent Semantic Analysis (LSA) has been employed for many web sites and user interaction applications developments. One of these applications is to provide a simpler interface such as the University of Colorado built the LSA site (<http://lsa.colorado.edu>). The Web site contains several precomputed semantic spaces and tools to manipulate those spaces in a number of ways (see Dennis, 2007). Other applications were used for BETA versions (<http://www.quintura.com>), for SEO/SEM business-words research with tools as iMetaSearch (<http://www.puffinwarellc.com>), for call-routing applications in IVR environments in TELCOs laboratories such as Lucent Technologies Bell Labs (Chu-Carroll & Carpenter, 1999, Jorge-Botana, Olmos & León, 2008), for to complete confidence levels in voice recognition packages and improve words disambiguation decisions in AVAYA laboratories (Tyson & Matula, 2004), and for modelling and predict user behavior in web navigation (Blackmon, Polson, Kitajima & Lewis, 2002; Juvina, van Oostendorp, Karbor, & Pauw, 2005; Jorge-Botana, 2006). Recently, LSA has been used to develop web-based e-learning platforms and graders, where LSA permits comparison about semantic similarity between different pieces of textual information such as sentence, paragraphs (Foltz, 1996; Landauer & Dumais, 1997, Landauer, 1998; Landauer, Foltz & Laham, 1998), as well as summaries (Foltz, 1996; Kintsch, Steinhart & Stahl, 2000; León, Olmos, Escudero, Cañas & Salmeron, 2006; Kintsch, Caccamise, Franzke & Dooley, 2007). For example, Kintsch et al., (2000) built Summary Street as a robust and practical educational tool that is increasingly used in Colorado classes. It is accessible over a set of specifically assigned Internet pages (<http://www.colit.org>). Nevertheless, there are some problems related to with genre of discourse. Usually research supported LSA as an excellent reliability in narrative text, but only moderate reliability with expository texts.

The study and objectives

In this paper we tested a computer-based procedure for assessing summaries using LSA combined with four expert human judgements in an expository text. This study is an extension of León et al. (2006) in which LSA was used with six standard methods (four holistic and two componential methods) in order to compare very short summaries assessments with four expert graders in narrative and expository texts. The results supported an excellent reliability of LSA in narrative text, but only moderate with expository texts. Only two holistic methods (*Summary-Text and Summary-Expert summaries*) obtained good results. In León et al. study *Summary-Text method* consists of comparing each student's summary with the whole text that was read to derive the LSA cosine (Kintsch et al., 2000). The higher the cosine between the summary and the text is, the better the summary will score. On the other hand, *Summary-expert summaries method* consist of assessing student summaries by comparing them with an expert summary (Landauer, Laham, & Foltz. 1998). For present study, six summaries written by experts were chosen as the standard, and the LSA cosine of each student summary compared with the average LSA cosine of the six expert summaries was computed. Thus, the student summary that was most similar to the expert one was evaluated as the best. In this study we analyse three new algorithms into these holistic methods (*Summary-Text and Summary-Expert summaries*) in order to improve the reliability of LSA and humans graders in expository texts. We describe these three new algorithms.

(1) *The Kintsch's predication algorithm.* LSA is not capable to distinguish multiple senses of a word. It is well known as LSA's polysemy problem. Kintsch (2001) proposes a network algorithm to adjust the sense of a word as it is applied to different contexts. The algorithm is applied to sentences with the structure *Argument* –

Predicate. The algorithm extracts the most semantic related neighbours of the predicate of a sentence and then links the most related neighbours to the argument. The essence of the algorithm is strengthening features of the predicate that adjust well with the argument. This algorithm extracts a context dependent meaning. For example, the LSA's representation of the sentence, "the river bank", could be much related to finance terms, unless the algorithm links uniquely the bank neighbours which at the same time are semantically related with river (e.g. area or shore). The general idea of this algorithm is provide additional semantic information to a textual piece. Thus, if we have a summary, instead of representing the vector with the sum of its words, we add to the summary those words which are the most related with it. Now, the summary it is composed with its own words and the others semantically related with it. Psychologically, the algorithm means that when we express something in a piece of language, the meaning conveyed has more than it is expressed explicitly. Therefore, this algorithm in our study means that the final vector represented in the LSA space consists in the words of the summary and the most semantically related concepts.

We make an adaptation of this algorithm in our study. First, we add to the student summary the n most semantically related but we only chose de p most related terms with the expository summarized text (where $p < n$). The number of related terms n is completely arbitrary (Kintsch, 2001) and we chose 50 and 20 for p . Thus, 20 terms are added to the vector summary, those 20 terms that at the same time are semantically related with de summary and the expository text. The predication algorithm was used firstly in sentences with a predicate and an argument. We extend this use taking the summary as a predicate of the text, and the text as the argument. The psychological idea is that the vector representing the summary conveys more information than the words it has. So we use these vectors to assess the summaries.

(2) *The best dimension algorithm.* Instead of using all the semantic space to represent each summary, Hu et al. (2007) only use dimensions that best contribute to improve the LSA assessment's verbal protocols. This algorithm supposes an intelligent and discriminative use of the semantic space. In our case, we applied it to summaries selecting randomly 30 out of 192 summaries. The four graders rate them and we take the average rate in each summary. Therefore we have 30 ratings that were assessed previously by graders. Now, LSA rates these 30 summaries as follows. Firstly, we removed the dimension than made the worst Pearson correlation between LSA-average human grader, obtained the best $p - 1$ dimension semantic space in terms of LSA-human graders reliability. Secondly, we removed the worst dimension in this reduced space that made poorer the LSA-human grader reliability. The algorithm continues until the 20% worst dimensions were removed. The algorithm gives us the semantic space than more contribute to the LSA-human graders reliability. Thus, each summary vector has the 80% of the original information. Then LSA use only the most relevant features of the semantic space as well as human graders only consider the most relevant features to assess summaries.

(3) *The Euclidian distance.* Instead using LSA the cosine measure to evaluate the texts where semantic similarity is overestimated, we use this algorithm (Rehder et al., 1998). The Euclidian distance incorporates at the same time vector length and the cosine of the angle, conveying more information about the summary contents and improving the capable of assessing of LSA.

Procedure and design

Material:

The corpus used in this study contains 372 documents related with de summarized text obtained in websites and include 5995 lemmatized words. The semantic space was set at 75 dimensions which cover 40% of the total variance (see Wild, Stahl, Stermsek, Neumann, 2005). The summaries used for this evaluation were taken from León & RLRG (2004). The summarized expository text was "Los Árboles Estranguladores" (The Strangler Trees). This expository text was extracted from a general encyclopedia adapted to the general reading skill of all participants. It contained 500 words and also required prior general knowledge. The summaries were obtained from 192 students attending middle/high school (14 - 16 years old) and six from experts (PhD students). The summaries had a maximum length of 50 words. The 192 summaries were rated by four graders in a 0-10 point scale. The LSA ratings, as we referred to earlier, was conducted with three new methods (Kintsch adapted algorithm, the best dimensions and the Euclidian distance). All three methods derive distinct vector for each summary. To rate each summary we compare each vector with the text vector or with six expert summaries (the six comparisons were averaged in one score).

To perform the data analysis we applied a two way-ANOVA test where the dependent variable was the reliability LSA-human grade. First factor was Algorithm (four algorithms): 1a: Kintsch adapted algorithm, 1b: best dimensions algorithm, 1c: Euclidian distance algorithm and 1d: standard algorithm that we use as base line. Second factor was Method (the two holistic methods): 2a: *Summary-Text* and 2b: *Summary-expert summaries*.

Results

We did not find method effect, $F(3, 24) = 0.01$, $MSE = 0.02$, $p = 0.94$. We found differences in the magnitude of reliabilities depending on the algorithm $F(3, 24) = 11.19$, $MSE = 0.02$, $p < 0.05$. The *post hoc* tests showed three means groups: Euclidian distance and base line algorithms had the lowest reliability means, base line and Kintsch adapted algorithms were in the middle and Kintsch adapted and Best dimensions had the highest reliabilities. However, these results are modulated by the interaction effect $F(3, 24) = 6.32$, $MSE = 0.02$, $p < 0.05$ (see Figure 1).

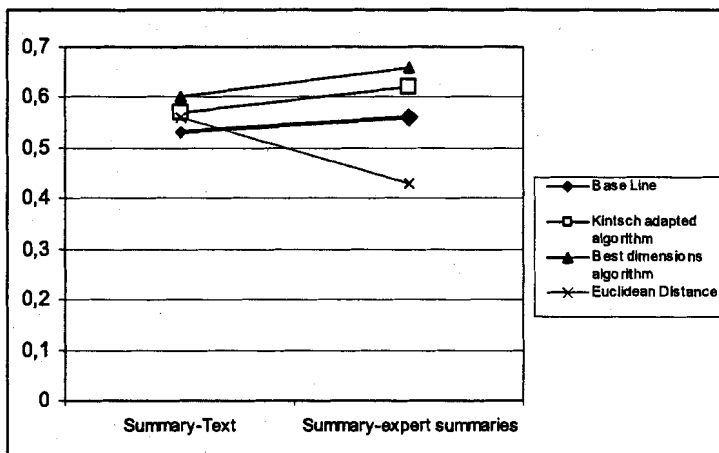


Figure 1. Interaction effect between Algorithm (lines) and Method (horizontal axis).

We found that best dimensions algorithm had more reliability than base line in *Summary-Text* method and in *Summary-expert summaries* method ($p < 0.05$). Euclidean distance reliability was as the rest of the algorithms only in *Summary-Text* method, but was the worst in *Summary-expert summaries* method. Kintsch adapted algorithm and base line algorithm did not significantly differ, in spite of the fact that Kintsch adapted algorithm has higher means in both methods.

Descriptive results show that best dimension algorithm overcomes 0.6 average reliability in both methods (*Summary-Text* and *Summary-expert summaries*), while Kintsch adapted prediction algorithm also overcomes it in *Summary-expert summaries* method.

Discussion

It is well known that LSA has problems to deal with too short texts (e.g., at the level of sentence the LSA results are poorer than at the level of paragraph; Rehder et al., 1998; Wade-Stein & Kintsch, 2004; Wiemer-Hastings, Wiemer-Hastings & Graesser, 1999). In our previous study (León et al., 2006) we obtained hopeful results, but worse whether comparing with other studies due to the length of the summaries (Foltz, Laham & Landauer, 1999; Kintsch et al., 2000). This limitation is marked in expository texts (León et al., 2006). Thus, the relevancy of the present study is that LSA required adding new algorithms in order to improve its assessment quality. Of the three algorithms used in this study the best dimension algorithm obtained the best LSA-human graders reliability. The prediction adapted algorithm had hopefully results. Euclidean distance has not obtained enough LSA-human graders reliability in *Summary-expert summaries* method, but since it incorporates vector length it would be a good measure in certain tasks. The future research will have to incorporate new ideas to confront the new task exigencies, ideas that link LSA with psychological models like Kintsch (2001), or ideas that improve its mathematical possibilities as proposed Hu et al. (2007). LSA provides a basis for a computational theory of meaning that permits to build computational models for many semantic problems. It is an interesting job of providing useful and individualized feedback that helps to guide students through a complex writing task such as e-learning. A good example can observe in Pearson knowledge technologies' products and services Web site (<http://www.pearsonkt.com/products.shtml>) that improve both writing skills and reading comprehension as well as subject area knowledge. ;

References

- Blackmon, M.H., Polson, P.G., Kitajima, M. & Lewis, C. (2002). Cognitive Walkthrough for the Web. *In CHI 2002: Proceedings of the conference on Human Factors in Computing Systems*, 463-470.
- Chu-Carroll, J., & Carpenter, B. (1999). Vector-based natural language calls routing. *Computational Linguistics*, 25 (3), 361-388.
- Denis, D. (2007). How to use the LSA Web site. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *The handbook of Latent Semantic Analysis* (pp. 57-70). Mahwah, NJ: Erlbaum.
- Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28(2), 197-202.
- 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. 143-167). Mahwah, NJ: Erlbaum.
- Jorge-Botana, G, Olmos, R, León J.A. (2008). Análisis de la Semántica Latente (LSA) y estimación automática de las intenciones del usuario en diálogos de telefonía (call routing). *Revista FAZ*. http://www.revistafaz.org/numero1/call_routing.pdf
- Jorge-Botana, G., Olmos, R., León, J. A. & Molinero, P., (submitted). Variantes a la extracción automática de vecinos semánticos con LSA y con el algoritmo de predicación. *Spanish Journal of Psychology*.
- Juvina, I., Oostendorp, H. van, Karbor, P. & Pauw, B. (2005). Toward Modeling Contextual Information in Web Navigation. *Cognitive Science*, 1078-1083.
- Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., and Dooley, S. (2007) Summary Street: LSA-Based Software for Comprehension and Writing. In Mc Namara, D., Landauer, T., Kintsch, W., and Dennis, S. (Eds.), *Handbook of Latent Semantic Analysis*, Erlbaum.
- Kintsch, E., Steinhart, D., Stahl, G. y LSA research group (2000) Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, 8(2), 87-109.
- Kintsch, W. (2001) Predication. *Cognitive Science*, 25, 173-202.
- Kintsch, W. (in press). Symbols systems and perceptual representations. In M. de Vega, A. Glenberg & A. Graesser (Eds.), *Symbols, Embodiment, and Meaning*. Oxford: University Press.
- Landauer, T. K. (1998). Learning and representing verbal meaning: The Latent Semantic Analysis theory. *Current Directions in Psychological Science*, 7, 161-164.
- 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.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.
- Landauer, T. K., Laham, D., & Foltz, P. W. (1998). *Computer-based grading of the conceptual content of essays*. Unpublished manuscript.
- León, J. A., & the Reading Literacy Research Group (2004). *La competencia lectora y los procesos de comprensión: Un proyecto de investigación basado en la evaluación de los tipos de comprensión* [Reading literacy and reading processes: A research project on assessment of types of comprehension]. Unpublished manuscript.
- 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.
- Rehder, B., Schreiner, M. E., Wolfe, B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25, 337-354.
- Tyson, N. Matula, V. (2004) *Improved LSI-Based Natural Language Call Routing Using Speech Recognition Confidence Scores*, EMNLP '04.
- Wade-Stein, D., & Kintsch, E. (2004). Summary street: Interactive computer support for writing. *Cognition and Instruction*, 22(3), 333-362.
- Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). How latent is Latent Semantic Analysis? In *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence* (pp. 932-937). San Francisco. Morgan Kaufmann.
- Wild, F, Stahl, C , Stermsek, G., & Neumann, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings of the 9th International Computer Assisted Assessment Conference (CAA)*, (pp. 485-494), Loughborough, UK.