Revista Signos
2009, 42(69)
71-81

Análisis del tamaño y especificidad de los corpus en la evaluación de resúmenes mediante el LSA. Un análisis comparativo entre LSA y jueces expertos

Ricardo Olmos
José Antonio León
Inmaculada Escudero
Universidad Autónoma de Madrid
España

Guillermo Jorge-Botana
Universidad Complutense de Madrid
España

Resumen: El Análisis Semántico Latente (LSA) es una sofisticada herramienta computacional de análisis semántico capaz de obtener una representación matemática del significado de las palabras o textos. LSA, entre otras aplicaciones, ha demostrado ser eficiente en la evaluación de textos. Esta herramienta adquiere la representación matemática de los textos analizando previamente un corpus lingüístico compuesto por documentos digitalizados. El principal objetivo de este estudio fue analizar qué propiedades han de tener distintos corpus lingüísticos (general, condensado, diversificado, y corpus de base) para que las evaluaciones de los resúmenes efectuadas por el LSA se parezcan lo máximo posible a las realizadas por 4 jueces humanos. Dichos resúmenes fueron elaborados por 390 estudiantes de Primaria, ESO y universitarios españoles. Los resultados indicaron que el tamaño de los corpus no tiene por qué ser tan generales ni tan grandes como los que se utilizan en Boulder (compuesto por millones de textos y más de un millón de palabras), ni tampoco demasiado específicos (menos de 300 textos y 5000 palabras) para que la evaluación que se desee hacer de ellos resulte satisfactoriamente eficiente.

Palabras Clave: Análisis Semántico Latente (LSA), resúmenes, evaluación del discurso, corpus lingüístico, estudiantes universitarios.

Recibido:
18-VI-2007
Aceptado:
9-V-2008

Correspondencia: Ricardo Olmos (ricardolmos@inicia.es). Departamento de Psicología Básica, Facultad de Psicología, Universidad Autónoma de Madrid. C/Iván Pavlov 6, Carretera de Colmenar, km 15, Campus de Cantoblanco, 28049, Madrid, España.

An analysis of size and specificity of corpora in the assessment of summaries using LSA. A comparative study between LSA and human raters

Abstract: Latent Semantic Analysis (LSA) is an automatic statistical method for representing the meanings of words and text passages. An emerging body of evidence supports the reliability of LSA as a tool for assessing the semantic similarities between units of discourse. LSA has also proved to be comparable to human judgments of similarities in documents. Before analyzing a linguistic corpus composed by digitized documents, this tool acquires the mathematical representation of the texts. The main objective of this study was to analyze what properties (general, condensed, diversified, and base corpus) different linguistic corpora should have so that the assessment of the summaries carried out by the LSA is as similar as possible to the assessment made by four human raters. Three hundred and ninety Spanish middle and high school students (14-16 years old) and undergraduate students read a narrative text and later summarized it. Findings indicate that the size of the corpora need not be as general and as big as those used in Boulder (made up by millions of texts and over one million words), nor do they have to be too specific (fewer than 300 texts and 5000 words) for the assessment to be satisfactorily efficient.

Key Words: Latent Semantic Analysis (LSA), summary, discourse assessment, linguistic corpus, university students.

INTRODUCCIÓN

Actualmente hay un fuerte interés en desarrollar herramientas de evaluación automáticas en textos abiertos y no ya solo en exámenes de tipo test. LSA constituye un gran avance en este campo, mostrando ser uno de los mejores candidatos para constituirse como una pieza fundamental en la evaluación computarizada de textos. Indudablemente contar con un sistema que simule correctamente el comportamiento de expertos al evaluar exámenes escritos puede ser de gran ayuda en la actividad académica, en procesos de selección o en general en cualquier contexto que sea evaluativo. Si bien la mayor parte de estos trabajos se han desarrollado fundamentalmente con corpus en lengua inglesa, existen, afortunadamente, algunos trabajos que se han interesado en desarrollar corpus en nuestra lengua española, con resultados muy interesantes (Pérez, Alfonseca, Rodríguez, Gliozzo, Strapparava & Magnini, 2005; León, Olmos, Escudero, Cañas & Salmerón, 2006; Venegas, 2006).

El Análisis Semántico Latente (LSA) es una herramienta informatizada que representa las palabras en vectores de *k* elementos, siendo *k* el número de dimensiones de un nuevo espacio semántico en que se ubican las palabras. LSA comienza analizando una matriz de palabras (filas) por documentos (columnas), en cuyas celdas está representado el número de veces que cada palabra ocurre en cada documento. Esta matriz contiene las relaciones 'brutas' que hay entre las palabras. Para llegar a una representación más profunda de las relaciones entre las palabras

se aplica una técnica matemática conocida como 'descomposición de valores singulares', técnica que reduce la matriz original en otras de dimensiones más reducidas pero preservando las relaciones latentes entre las filas y las columnas de la matriz original. Esta técnica matemática construye un nuevo espacio semántico *k* dimensional reducido en el que las palabras quedan representadas de forma que se preservan las relaciones semánticas profundas entre ellas, un espacio del que se han eliminado las connotaciones espurias de las palabras y donde emerge una estructura más abstracta o latente. Para analizar la relación semántica entre dos palabras se utiliza el coseno del ángulo entre los vectores que las representan. Cuanto más próximo a uno es el coseno mayor es la relación semántica entre las palabras y cuanto más próximo a cero es el coseno menor es la relación entre las palabras. Pero las relaciones semánticas se pueden examinar además entre una palabra y un texto, o entre dos textos de cualquier longitud. Al fin y al cabo el modelo asume que un texto es el vector suma de los vectores de las palabras que lo componen[1], por lo que el significado de un texto queda representado por la suma de los significados de sus palabras. La Figura 1 representa vectorialmente a tres textos.

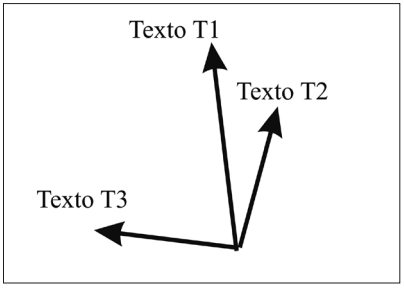

Figura 1. Tres textos en un nuevo espacio semántico.

Cuando se compara la similitud semántica entre los tres textos dentro del nuevo espacio semántico definido por LSA tenemos que los textos 1 y 2 son parecidos porque forman un ángulo cerrado y por lo tanto su coseno es próximo a 1. La relación semántica de los textos 1 y 2 con el tercero es casi nula. De esta manera, dos textos o dos palabras quedan representados en un nuevo espacio semántico que permite describir las relaciones de significado. Se ha comprobado reiteradamente que las representaciones semánticas de LSA son muy parecidas a las que tenemos los humanos (Landauer & Dumais, 1997; Landauer, Foltz & Laham, 1998).

La representación que LSA tiene de las palabras depende de qué textos ha analizado, es decir, depende de los 'corpus lingüísticos' que se hayan utilizado para conformar la matriz original de palabras. Los corpus lingüísticos van desde unos cientos de textos de carácter muy específico

(corpus de dominio específico) hasta otros formados por varios millones de textos cuya temática puede ser cualquiera (corpus de dominio generalista). En este estudio se verá precisamente el impacto que tiene la composición de los corpus en la evaluación de resúmenes. Para ello utilizaremos cuatro corpus diferenciados y así ayudar en el proceso de creación de corpus.

1. Método

1.1. Objetivos

El objetivo de este trabajo fue doble. Por un lado, analizar qué propiedades han de tener distintos corpus lingüísticos (general, condensado, diversificado y corpus base) para una mejor implementación en español de LSA. Por otro, confrontar las evaluaciones de los resúmenes efectuadas por el LSA con las realizadas por 4 jueces humanos.

1.2. Muestra

En este estudio se contó con 390 resúmenes de estudiantes de primaria, de estudiantes de la educación secundaria obligatoria (ESO) y de estudiantes universitarios elaborados a partir de un texto que cuenta cómo sobreviven unos árboles en su lucha por captar la luz del sol.

1.3. Corpus

El material consistió en 390 resúmenes elaborados por alumnos de primaria (30,5%), ESO (49,2%) y universitarios (20,3%) y cuatro corpus lingüísticos diferentes sometidos a examen, que describimos a continuación: A) El 'corpus base' compuesto por 275 textos de contenidos similares al texto de prueba (sobre el que se realizaba el resumen) y 3047 palabras diferentes. Le hemos denominado 'corpus base' porque representa la línea base para el análisis de los corpus, el corpus mínimo necesario para poder hacer funcionar al LSA con un mínimo de garantía. B) El 'corpus condensado' que además de contar con el 'corpus base', se le añadieron 63 textos con un contenido muy parecido al texto que resumieron los alumnos. Contó finalmente con 338 textos y 5408 palabras diferentes. En este corpus se restringió el uso de las palabras clave del texto a contextos lo más parecido posible al texto de evaluación. C) El 'corpus diversificado' constituido por el base más 97 textos con usos más diversos de las palabras clave del texto a resumir. Con esto se introdujo una diversidad mayor en los contextos donde aparecen las palabras más importantes del texto (por ejemplo, árbol, raíz, luz, planta, selva, sol, adaptación y supervivencia). D) Finalmente, el 'corpus generalista', contaba con un volumen de más de dos millones de textos y más de un millón de términos diferentes. Representa un corpus vastísimo en contenidos y diversidad de textos, de carácter muy generalista.

Los cuatro corpus se crearon utilizando los mismos procedimientos. Se filtraron en todos ellos las palabras funcionales (*stop words)* y se lematizaron los cuatro corpus. La lematización consistió en agrupar todas las formas verbales bajo el mismo verbo, así como para sustantivos y adjetivos singulares y plurales, masculinos y femeninos, quedasen también marcados como una sola forma. Esta lematización ya se ha empleado en otros estudios (Denhière, Lemaire, Bellissens & Jhean-Larose, 2007).

Para examinar la utilidad de los corpus se estableció como criterio el índice de acuerdo que existe entre los jueces y LSA (fiabilidad), obtenido a través de la correlación de Pearson entre las calificaciones del LSA y los jueces. Como se sabe una correlación próxima a 1 informa de un acuerdo perfecto y próximo a 0, un acuerdo nulo. Como línea base se tomó la correlación media entre los propios jueces.

1.4. Procedimiento y evaluación

Los resúmenes contaban con un máximo de 50 palabras y se evaluaron por 4 jueces expertos, que fueron entrenados específicamente para la tarea. Las evaluaciones se dieron en una escala de 0 a 10 puntos. Por su parte, LSA evaluó los resúmenes con uno de los procedimientos utilizados habitualmente (Foltz, 1996; León et al., 2006), esto es, comparar con el coseno el vector del resumen con los vectores de 6 resúmenes expertos[2]. Los 6 vectores representan 6 resúmenes que elaboran expertos y que por lo tanto representan una muestra de resúmenes idóneos. La calificación que LSA da a cualquier resumen es el coseno medio que el resumen del estudiante tenga con los 6 resúmenes expertos. Se parte de que cuanto más alto sea el coseno medio, mejor es el resumen o mejor están representados los contenidos en el resumen; y viceversa, cuanto más pequeño sea el coseno promedio, peor se considera el resumen. Supongamos dos resúmenes comparados con los 6 resúmenes expertos, como se indica en la Tabla 1:

Tabla 1. Evaluación ficticia de dos resúmenes.

E1	E2	E3	E4	E5	E6	**Med**
0,70	0,64	0,55	0,83	0,78	0,65	**0,69**
0,25	0,45	0,26	0,45	0,33	0,32	**0,34**

La primera fila (0,70; 0,64, etc.) representa los cosenos del primer resumen con cada uno de los resúmenes expertos (E1, E2, etc.). La segunda fila lo mismo pero para el segundo resumen. La puntuación final es el coseno medio representado en la última columna. El primer resumen obtiene una calificación de 0,69 y el segundo de 0,34. El primer resumen es mejor porque se parece más a los resúmenes expertos. Una de las ideas fundamentales del LSA, como se ha se-

ñalado antes, es que el espacio semántico del LSA recoge la estructura latente y profunda que tienen las palabras, así que no es necesario que el vocabulario recogido en el resumen se solape con el de ningún resumen experto para que el coseno entre sus vectores sea próximo a 1, es decir, para que el resumen sea considerado bueno. Este hecho es realmente lo que le confiere ventaja al LSA sobre otras herramientas, la sutilidad con la que representa el vocabulario en un espacio semántico latente.

2. Resultados

Un primer análisis de los datos requiere evaluar la fiabilidad de las evaluaciones realizadas entre los cuatro jueces. La correlación media (Pearson) entre los cuatro jueces fue de 0,85, lo que evidencia fiabilidad alta en la evaluación de los resúmenes realizada por estos. Por otro lado, también se realizó una correlación de Pearson para evaluar la fiabilidad entre la herramienta y los cuatro jueces. Los resultados se expresan en la Tabla 2. Como puede observarse en ella, las correlaciones entre las evaluaciones de los jueces y las obtenidas por el LSA en sus respectivos corpus fueron, en todos los casos, significativas. En otras palabras, todas las correlaciones entre los jueces y el LSA fueron positivas y estadísticamente significativas. La más alta fue de 0,66, la obtenida entre el corpus generalista y el juez 2 y la más baja la correspondiente al corpus condensado con el juez 4 (0,46). La Figura 2 muestra las correlaciones medias entre el LSA y los jueces. En dicha Figura puede observarse estas mismas correlaciones, en la que la altura de las barras representa la correlación media entre los jueces y el LSA.

Tabla 2. Correlaciones entre el LSA y los cuatro jueces.

	Base	Cond	Div	Gener
Juez1	0,52**	0,52**	0,57**	0,64**
Juez2	0,58**	0,52**	0,63**	0,66**
Juez3	0,53**	0,49**	0,57**	0,62**
Juez4	0,51**	0,46**	0,56**	0,60**

Base = Corpus base, Cond = Corpus condensado
Div = Corpus diversificado, Gener = Corpus generalista
** Correlación estadísticamente significativa al 1%.

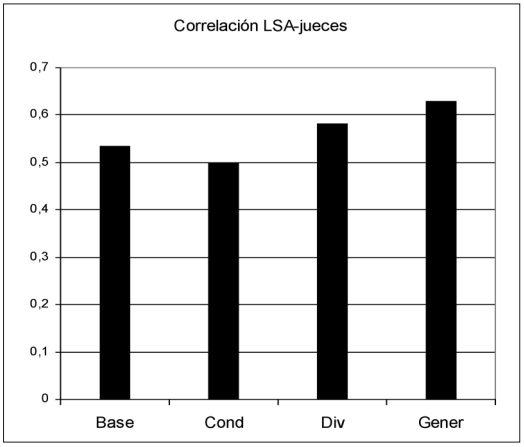

Figura 2. Correlación LSA-jueces.
Base = Corpus base, Cond = Corpus condensado
Div = Corpus diversificado, Gener = Corpus generalista

Para establecer qué propiedades pueden estar tras estos resultados se analizaron los corpus más exhaustivamente. Por un lado se comprobó la relación semántica media entre las palabras de cada corpus (exceptuando el generalista) analizando las correlaciones existentes entre pares de palabras. No se examinaron todos los posibles pares (para un corpus con tres mil palabras distintas existen más de cuatro millones de comparaciones posibles), sino que se extrajeron dos mil comparaciones aleatorias para alcanzar una muestra representativa de las relaciones entre las palabras.

Tabla 3. Correlaciones medias entre las palabras para los tres corpus analizados (excepto el generalista).

	N	Media	Desv. típ.
Correlaciones palabras corpus base	2000	0,0106	0,127
Correlaciones palabras corpus condensado	2000	0,0087	0,125
Correlaciones palabras corpus diversificado	2000	0,0065	0,128

Las correlaciones mayores se dan en el corpus base y las menores en el corpus diversificado, aunque estas diferencias no son estadísticamente significativas ($p > 0,05$). Las desviaciones típicas tampoco difieren significativamente entre los corpus. Las diferencias entre los corpus

no residen en diferencias en las correlaciones medias. Se analizó si los corpus eran diferentes en densidad. Por densidad podemos pensar en términos de órdenes (Kontostathis & Pottenger, 2003; Lemaire & Denihere, 2006) o co-ocurrencias entre palabras. Dos palabras co-ocurren dos veces por ejemplo si en el conjunto de documentos han aparecido conjuntamente dos veces. Estas ocurrencias son las llamadas de 'primer orden'. Podría darse la situación de que el 'corpus condensado' tiene más densidad que el 'corpus base' y éste a su vez que sea más denso que el 'corpus diversificado'.

Tabla 4. Medias de ocurrencia de primer orden en tres corpus.

Corpus	Media	Desv. típ.	N
Base	0,11	0,07	3047
Condensado	0,25	0,37	3313
Diversificado	0,10	0,06	3646
Total	0,15	0,23	10006

Los resultados muestran diferencias estadísticamente significativas entre los órdenes de los tres corpus ($F(2,10003) = 528,03$, $p < 0,01$) y las comparaciones múltiples señalan al 'corpus condensado' como el más denso o con más órdenes entre sus palabras, luego el 'corpus base' y por último el 'corpus diversificado'. Estos resultados parecen mostrar que efectivamente en el 'corpus condensado' es más probable que las palabras aparezcan asociadas más veces en los mismos documentos que en los otros dos corpus. Este resultado no es más que una comprobación de que los corpus tienen efectivamente una densidad diferente. La media del corpus condensado de 0,25 significa que por término medio cada palabra se asocia con el 25% del resto.

CONCLUSIONES

En este estudio hemos analizado la capacidad del LSA para evaluar resúmenes procedentes de estudiantes comparándola con la evaluación realizada por jueces expertos. La elección de analizar resúmenes está suficientemente reflejada en la literatura (véase, por ejemplo, la importancia del resumen en el modelo de comprensión del discurso de van Dijk & Kintsch, 1983). Según estos autores, la tarea de resumir sería mucho más productiva para la comprensión que releer un texto. Resumir requiere un gran esfuerzo de abstracción y organización de la información que no solo se limita a reproducir lo leído, sino a reconstruir las principales ideas leídas, reforzando la construcción del conocimiento en la memoria. A su vez, los profesores han notado marcados cambios en el aprendizaje de los alumnos cuando además de hacerles leer textos les han hecho resumir.

Los resúmenes siguen siendo una estupenda forma de promover el conocimiento entre los estudiantes, ya que requiere un enorme esfuerzo de abstracción y síntesis sobre el material aprendido. Dentro de este contexto el LSA se ha centrado fundamentalmente en una herramienta automatizada que descargue y libere de trabajo a los profesores y sea capaz de evaluar resúmenes. Ya existen estudios que han utilizado el LSA con corpus en español para evaluar resúmenes de ciencias (Venegas, 2006) o resúmenes realizados a partir de textos narrativos y expositivos (León et al., 2006), con resultados muy esperanzadores.

De manera general y refiriéndonos a este estudio, las correlaciones obtenidas entre los jueces y el LSA son bastante altas. Con ningún corpus se bajó en promedio de 0,50, lo cual permite ser optimista respecto a las posibilidades del LSA en cuanto a su uso en evaluación. Por otra parte, estos resultados son coherentes con otros estudios (Landauer et al., 1998; Kintsch, Steinhart & Stahl, 2000) y hay que tener en cuenta que LSA funciona peor cuanto menor longitud tienen los textos. Se forzó la tarea para que los resúmenes no excedieran de 50 palabras y se sabe que LSA obtiene un rendimiento óptimo a partir de 250 palabras (Rehder, Schreiner, Wolfe, Laham, Landauer & Kintsch, 1998), así que lo natural es que el grado de acuerdo entre LSA y los jueces sea más parecido al que hay entre los propios jueces con resúmenes un poco más extensos.

En cuanto a nuestro objetivo de investigación, los resultados son claros respecto a los corpus y a la evaluación de la calidad de los resúmenes, en el sentido de que una restricción del uso de las palabras a contextos muy limitados no favorece necesariamente la calidad de la evaluación del LSA. Un caso claro que ilustra esta conclusión es el caso del 'corpus condensado', ya que arroja correlaciones LSA-jueces por debajo de los restantes corpus. Este corpus contiene un 20% de los textos con unos contenidos casi idénticos al texto que los estudiantes resumieron. Una posible explicación de este hecho puede deberse a que con este corpus LSA se ha restringido a un espacio semántico cuyo contexto es extremadamente específico. El 'corpus base' funciona un poco mejor, como se aprecia en la Figura 2. Por su parte, el 'corpus diversificado' aún se muestra mejor en las evaluaciones que estos dos. Por último, el 'corpus generalista' es el que mejores resultados ofrece. Este corpus es miles de veces más grande que cualquiera de los otros tres y la diferencia está en que se confecciona con textos que no tienen una temática específica, por lo que el uso de las palabras encuentra muchísimos más contextos que en el resto de los corpus. La diversidad de contextos parece mejorar algo las evaluaciones que con respecto a los demás corpus. El gran esfuerzo que conllevaría abordar un corpus como este y la diferencia que tiene respecto del diversificado hace aconsejable y, mucho más práctico, trabajar con tamaños pequeños con una temática no excesivamente ceñida al texto de evaluación. Consideramos interesante que futuras líneas de investigación estudien más exhaustivamente las características de los corpus para que LSA pueda evaluar con una calidad parecida a la de los jueces.

REFERENCIAS BIBLIOGRÁFICAS

Berry, M., Dumais, S. & O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM: Review*, 37, 573-595.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science* , 41, 391-407.

Denhière, G., Lemaire, B., Bellissens, C. & Jhean-Larose, S. (2007). A semantic space modeling children's semantic memory. En T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.) *The handbook of Latent Semantic Analysis* (pp. 143-167). Mahwah, NJ: Erlbaum.

Foltz, P. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*(2), 197-202.

Kintsch, E., Steinhart, D., Stahl, G. & LSA research group (2000). Developing summarization skills through the use of LSA-Based feedback. *Interactive Learning Environments, 8*(2), 87-109.

Kontostathis, A. & Pottenger, W. (2003). A framework for understanding LSI performance. *Proceedings of ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval-ACMSIGIR MF/IR* [en línea] Disponible en: http://www.cse.lehigh.edu/~billp/pubs/IEEEICDM02WorkshopApril.pdf

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

Landauer, T. & Dumais, S. (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.

Lemaire, B. & Denhière, G. (2006). Effects of high-order co-occurrences on word semantic similarity. *Current Psychology Letters, 1*(18), 628-637.

León, J., Olmos, R., Escudero, I., Cañas, 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.

Pérez, D., Alfonseca, E., Rodríguez, P., Gliozzo, A., Strapparava, C., & 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.

Rehder, B., Schreiner, M., Wolfe, M., Laham, D., Landauer, T. & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes,* 25, 337-354.

Van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Venegas, R. (2006). La similitud léxico-semántica en artículos de investigación científica en español: Una aproximación desde el Análisis Semántico Latente. ***Revista Signos, 39***(60), 75-106.

NOTAS

1 Esta suma se pondera por el inverso del valor singular de cada dimensión (método Holding-in) (Berry, Dumais & O'Brien, 1995; Deerwester, Dumais, Furnas, Landauer & Harshman, 1990).

2 Se ha comprobado que este método funciona muy bien. En el estudio se utilizaron 6 resúmenes de expertos, número que es arbitrario y que creemos suficiente como para tener representados distintos resúmenes de calidad.