Subjetividad y Procesos Cognitivos, Vol. 14, Nº 2, 2010
Pág. 333-349, ISSN impreso: 1666-244X, ISSN electrónico: 1852-7310

Evaluación de resúmenes de textos narrativos y expositivos utilizando análisis semántico latente

Evaluation of narrative text and presentation text summaries using latent semantic analysis

René Venegas*

Resumen

El objetivo de este trabajo es identificar un método de evaluación automática de resúmenes realizados a partir de textos de tipo narrativo y expositivo en español. Para llevar a cabo esta tarea se correlaciona la evaluación realizada por tres docentes a 373 resúmenes con los resultados entregados por el análisis semántico latente. Los puntajes asignados por el análisis semántico latente se obtienen utilizando tres métodos 1) Comparación de los resúmenes con el texto fuente, 2) Comparación de los resúmenes con un resumen consensuado 3) Comparación de los resúmenes con tres resúmenes construidos por tres evaluadores. Entre los resultados más relevantes se destacan: a) una alta correlación entre la evaluación realizada por los evaluadores ($\overline{x}$ 0,63); b) una alta correlación entre los métodos computacionales utilizados ($\overline{x}$ 0,62) y c) una correlación promedio positiva media-alta entre las evaluaciones realizadas por los docentes y el análisis semántico latente en el segundo y tercer método ($\overline{x}$ 0,53 en ambos casos y tipos de textos). Ambos métodos presentaron mayor correlación promedio con los evaluadores cuando los textos evaluados eran predominantemente narrativos ($\overline{x}$ 0,59 y 0,45 respectivamente).

Palabras clave: análisis semántico latente, evaluación automatizada de resúmenes, evaluación holística, textos narrativos y expositivos, resúmenes breves.

Summary

The objective of this study is to identify a method for the automatic evaluation of the summaries developed from narrative and expository Spanish texts. In order to fulfill this task evaluation of 373 summaries carried out by three teachers is correlated with the results delivered by latent semantic analysis. Scores assigned by the latent semantic analysis are obtained through three methods: 1) Comparison of the summaries with the source text, 2) Comparison of the summaries with a consensuated one, 3) Comparison

* Pontificia Universidad Católica de Valparaíso. Dirección: Av. Brasil 2830, Valparaíso, Chile. E-mail: rene.venegas@ucv.cl

UCES Subjetividad y Procesos Cognitivos, Vol. 14, Nº 2, 2010
Pág. 333-349, ISSN impreso: 1666-244X, ISSN electrónico: 1852-7310

of the summaries with three summaries developed by three evaluators. The most relevant results include: a) a high correlation between assessments by the evaluators (:0.63), b) a high correlation between the computational methods used (:0.62) and c) a positive medium-high average correlation between assessments undertaken by the teachers and the latent semantic analysis in the second and third method (;0.53 in both cases and types of texts). Both methods presented greater average correlation with testers when the texts evaluated were predominantly narratives (;0.59 and 0.45 respectively).

Key words: latent semantic analysis, automated summary evaluation, holistic evaluation, narrative and expository texts, brief summaries.

1. Introducción

Como sabemos, la educación de la lecto-escritura en Latinoamérica no es lo suficientemente efectiva como para que nuestros estudiantes desarrollen las competencias necesarias para comprender y producir textos que les permitan desenvolverse adecuadamente en la sociedad actual (Peronard, 1989; Peronard, Gómez, Parodi, & Núñez, 1998; PISA, 2007). En general los investigadores de la comprensión al evaluar los procesos psicolingüísticos realizados por los individuos, utilizan múltiples opciones para su evaluación, por ejemplo, respuesta a preguntas de tipo literales e inferenciales (tanto locales como globales), desarrollo de mapas conceptuales y/o mentales, construcción de preguntas, parafraseo, escritura de resúmenes, etc.

En nuestro caso nos interesa indagar en el resumen, como técnica de evaluación. El uso de esta técnica, encuentra apoyo teórico en los planteamientos de van Dijk (1978). Desde esta perspectiva psicolingüística, lo que se pretende es formalizar los procesos mediante los cuales los contenidos de los textos son elaborados por el lector para ser ingresados a su memoria. Para ello el autor propone las denominadas macrorreglas (van Dijk, 1978) y luego las macroestrategias (van Dijk & Kintsch, 1983), mediante las cuales intenta explicar el hecho de que después de la lectura de un texto relativamente extenso, lo que la persona recuerda y verbaliza no son todas las ideas expresadas originalmente en él. De este modo, el buen comprendedor aplica las reglas, eliminando las proposiciones que estima poco relevantes y reelaborando otras para construir su propia versión del texto. Sin embargo, desde el punto de vista evaluativo esta técnica presenta algunos problemas, principalmente en relación a variables humanas como la carga cognitiva del evaluador, la atención paralela a elementos formales (ortografía, caligrafía) de la producción escrita, aspectos subjetivos que puedan intervenir, sistematicidad en la aplicación de los criterios de evaluación, consenso entre múltiples evaluadores y la extensa cantidad tiempo requerida para la evaluación.

Tanto estos problemas como el interés de captar automáticamente la información central de los textos han incentivado el estudio del resumen y su evaluación desde

la perspectiva computacional. De este modo, tanto construcción automática de resúmenes como la evaluación de ellos ha sido un problema que ha sido tratado desde mediados de los años sesenta en adelante, aunque aún no se han encontrado sistemas computacionales suficientemente confiables para desarrollar ambas tareas (Batten, 2003). No obstante, los avances realizados a partir de la lingüística computacional, el procesamiento del lenguaje natural y el desarrollo de diversas técnicas de recuperación de información permiten pensar que hoy en día se está más cerca al cumplimiento de la generación de resúmenes automatizados y evaluación de ellos de modo similar a cómo los realizan los seres humanos.

En general, las técnicas para construir y evaluar automáticamente resúmenes de textos se clasifican en dos grandes categorías: lingüísticas y estadísticas. Las lingüísticas usan conocimiento relativo a la sintaxis, la semántica o el uso la lengua, en tanto que las estadísticas operan identificando valores para las palabras y las frases que se encuentran en el texto utilizando técnicas estadísticas como: frecuencias, n-gramas, co-ocurrencias, etc.

Kintsch (1998, 2000, 2001, 2002) presenta la posibilidad de utilizar el Análisis Semántico Latente (LSA, por su sigla en inglés) para extraer desde los textos similitudes léxico-semánticas que permitan acceder a las proposiciones del texto, por medio de una orientación estadístico-matemática (Landauer & Dumais, 1997) con el fin de representar y evaluar el proceso de la comprensión por medio de técnicas computacionales.

Los estudios teóricos y empíricos en torno al LSA se han desarrollado ampliamente en distintas lenguas, aunque preferentemente en inglés (Landauer & Dumais, 1997; Landauer, Laham, Rehder & Schreiner, 1997; Rehder, Schreiner, Wolfe, Laham, Landauer, & Kintsch, 1998; Kintsch, Steinhart, Stahl, LSA Research Group, Matthews, & Lamb, 2000; Landauer & Psotka, 2000; Landauer, Laham, Foltz, 2003, etc). Estos estudios se han orientado fundamentalmente a probar que los modelos vectoriales, en particular el LSA, pueden dar cuenta del procesamiento de los significados que realizan los seres humanos y son efectivos en la evaluación automatizada de ensayos y resúmenes. Sin embargo, en el ámbito de la evaluación de resúmenes en lengua española los estudios que consideran el LSA aún son escasos (Pérez, Alfonseca, Rodríguez, Gliozzo, Strapparava & Magnini, 2005; León, Escudero, Cañas & Salmerón, 2005; Venegas 2007, 2009a, b, en prensa).

El objetivo de esta investigación es, por tanto, identificar un método de evaluación automática de resúmenes de textos de tipo predominantemente narrativo y expositivo en español, utilizando el análisis semántico latente, que se correlacione adecuadamente con la evaluación realizada por evaluadores humanos.

2. El resumen como técnica de evaluación de la comprensión del discurso escrito

Van Dijk (1978, 1983) reconoce en un texto dos tipos de estructuras semánticas: las microestructuras y las macroestructuras. Ambas son de tipo proposicional. En la primera, las relaciones se establecen entre proposiciones, en cambio en la segunda, las relaciones se establecen entre proposiciones que expresan la síntesis de un conjunto de proposiciones de nivel microestructural. En términos más simples, estas macroproposiciones corresponden a las ideas principales del texto y que, una vez integradas, constituyen la macroestructura, las que en su textualización pueden corresponder al resumen, a frases tópicas, a un titular o a un listado de ideas.

El resumen no debe resultar una mera transposición de fragmentos del texto original, sino una reelaboración de lo que este propone globalmente. En este sentido, el resumen es la textualización de un significado que representa de modo abstracto el significado total del contenido del artículo, también llamado "macrosemantización" (van Dijk, 1983; Venegas, 2005).

El resumen, en comparación con otras técnicas, resulta ser relativamente fácil de aplicar y permite distinguir a los sujetos que comprendieron la esencia de un texto de aquellos que no lo hicieron (Peronard, 1997). Sin embargo, en su utilización como técnica evaluativa pueden presentarse algunos problemas debido a factores relacionados con los evaluadores, tales como: la falta de sistematicidad en la aplicación de los criterios de evaluación, la valoración excesiva de aspectos formales como la ortografía y la caligrafía, la mayor atención al proceso de producción del resumen que al de la comprensión en sí misma, el tiempo que requiere su evaluación, entre otros.

3. Análisis semántico latente

El LSA es una técnica matemático-estadística que sirve para la extracción y representación de relaciones de significado entre palabras y párrafos, lo que se realiza a partir de una gran cantidad de textos (Landauer, McNamara, Dennos & Kintsch, 2007).

El LSA extrae sus representaciones de significado de palabras y párrafos, exclusivamente a partir del análisis matemático-estadístico del texto. Nada de su conocimiento viene desde la información perceptual sobre el mundo físico, del instinto, o de la experiencia generada por funciones corporales, sentimientos y/o intenciones. Así, su representación del significado es parcial y limitada, puesto que no hace uso de relaciones sintácticas ni lógicas ni morfológicas. A pesar de lo anterior, Landauer (2002) explica, al menos para la lengua inglesa, que el 80% de la información potencial en el lenguaje está en la elección de palabras, sin tener en cuenta el orden en el que ellas aparecen. Junto con esta idea de representación, sin sintaxis, se asume que en estas grandes cantidades de corpora existen interrelaciones semánticas débiles entre palabras que son potenciadas por el método de reducción de dimensiones denominado Descomposición en Valores

Singulares (SVD, por su sigla en inglés) (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). Por medio del SVD se construye un espacio semántico representativo de la información necesaria para uno o varios dominios de conocimiento a partir del corpus textual.

El LSA permite calcular las similitudes semánticas existentes entre palabras y párrafos de textos, estableciendo mediciones de su representación vectorial a través del cálculo de coseno de sus ángulos en un espacio multivectorial (Landauer, Foltz & Laham 1998; Landauer, McNamara, Dennos & Kintsch, 2007). Los valores de coseno van de 1 para vectores con la misma dirección (esto significa que lo medido es igual) a 0 para aquellos vectores ortogonales (perpendiculares en el espacio multivectorial, es decir, que lo medido es completamente distinto). Los valores deben ser normalizados, para hacer más efectiva la comparación entre ellos, ya que vectores más largos (correspondiente a documentos más largos) podrían tener una ventaja injusta respecto de los vectores más cortos. Además, la normalización de los valores de coseno permite que estos sean calculados como un producto simple (multiplicación de los vectores) (Deerwester et al., 1990; Landauer et al, 1998, Manning & Schütze, 2003).

4. Trabajos relacionados

Los distintos estudios de evaluación automatizada de resúmenes han mostrado, en general, una buena correlación con los juicios realizados por humanos. A continuación mencionaremos algunos ejemplos en los que se utilizan técnicas estadísticas.

Lin y Hovy (2003) y Lin (2004) desarrollan un programa de evaluación automática de resúmenes llamado *"Recall-Oriented Understudy for Gisting Evaluation"* (ROUGE). Este programa utiliza 5 medidas para, de forma automática, determinar la calidad de un resumen por comparación con otros resúmenes creados por seres humanos. Estas medidas se basan en el conteo de la correspondencia de unidades tales como n-gramas y grupos (y pares) de palabras entre el resumen generado por el programa y los resúmenes ideales creados por los seres humanos. Así por ejemplo, en el caso de los valores asignados por ROUGE (y sus distintas medidas) a los resúmenes y los asignados por los seres humanos en el caso de resúmenes de no más de 100 palabras la correlación con tres evaluadores humanos es ($\overline{x}$ 0,82 pearson). En el caso de resúmenes breves (alrededor de 10 palabras) la correlación con cuatro evaluadores alcanza un promedio de 0,76 (pearson). Finalmente, en el caso de resúmenes de multidocumentos el promedio con 2 a 4 evaluadores es de 0,63 (pearson). Cabe señalar que estas correlaciones consideran el cálculo de la aplicación de las técnicas considerando la extracción de *"stopwords"* (palabras que no aportan gran contenido semántico y de mucha frecuencia en un corpus de textos). Finalmente, cabe señalar que en este tipo de evaluación lo que se considera es la correspondencia entre las palabras de ambos documentos (resumen automático y resumen humano), así los algoritmos dan valor cero en los casos en que se utilicen palabras distintas con sentido similar, lo cual es

potencialmente un problema en la evaluación automatizada de resúmenes humanos (Lin, 2004).

Para los casos en los que se ha usado LSA diversos estudios reportan una alta correlación entre la evaluación automatizada y los juicios de evaluadores humanos. Así, por ejemplo, Kintsch et al. (2000) llevó a cabo dos comparaciones diferentes entre las puntuaciones dadas por el LSA y las evaluaciones realizadas por humanos a resúmenes escritos (longitud media de 250-350 palabras) por alumnos de 5º grado. En la primera comparación, se calcula el coseno entre los resúmenes de los estudiantes y el texto fuente leído por los estudiantes. En este estudio la correlación entre el evaluador y el del LSA fue de 0,64. En la segunda comparación, investigaron si el LSA podía identificar una oración dada en el texto fuente con una oración del resumen tal como lo haría dos evaluadores humanos. Las calificaciones otorgadas por el LSA se asemejaron en un 84,9% con las puntuaciones de los evaluadores de primer grado y en un 83,2% con los del segundo grado. Kintsch et al. (2000) concluyó que las puntuaciones LSA eran bastante comparable a las evaluaciones que podían realizar los maestros con experiencia a estos resúmenes y que el LSA funcionaría casi tan bien como los seres humanos para determinar la fuente de conocimiento para una oración dada.

Uno de los programas más conocidos que utiliza LSA en la evaluación y retroalimentación automatizada de ensayos es IEA *(Intelligent Essay Asesor)* desarrollado por Landauer, Laham y Foltz (2003). En este programa, a diferencia de otras perspectivas estadísticas, los autores plantean que el LSA se focaliza en el contenido semántico de los textos más que aspectos mecánicos de la producción escrita como la gramática, la ortografía literal y puntual, permitiendo así obtener una evaluación más realista de la compresión del discurso escrito de los sujetos. Landauer, Foltz y Laham, (1998). Lo anterior no significa que el IEA no proporcione retroalimentación sobre los aspectos formales en la evaluación del ensayo, de hecho el programa evalúa también aspectos como el estilo, la ortografía y la coherencia sintáctica. El LSA en el AIE ha sido utilizado para evaluar la calidad y cantidad del conocimiento que se transmite en un ensayo, aplicando tres métodos distintos en los que el ensayo a evaluar se compara con: a) ensayos de otros estudiantes evaluados anteriormente; b) ensayos modelos de expertos y publicaciones anteriores del tema; y c) comparación interna de un conjunto de ensayos no evaluados. Estas mediciones indican el grado en que cierto ensayo tiene un significado igual al de los textos comparados. Esto, según los autores, puede considerarse como una medición de calidad desde una perspectiva semántica. A modo de prueba de lo anterior, Landauer, Laham y Foltz (2003) reportan correlaciones promedio de 0,73 a 0,77 entre el AIE y los evaluadores (al menos 2) en diversos experimentos (por ejemplo, producción de ensayos breves, pruebas estandarizadas y pruebas tomadas en sala de clases sobre diversos temas) considerando los aspectos de cantidad y cualidad de la información incluida en los ensayos evaluados. Esta investigación concuerda con los hallazgos de Kintsch et al. (2000) y permiten pensar

en la conveniencia de utilizar el LSA en la evaluación automatizada de resúmenes en otras lenguas.

León et al. (2005) proponen evaluar resúmenes muy concisos (50 palabras de largo) de dos tipos de texto (narrativo y expositivo) mediante el análisis semántico latente (LSA) y comparar estos resultados con las evaluaciones de cuatro expertos humanos. Los autores utilizan LSA para estimar la similitud semántica entre los resúmenes mediante seis métodos diferentes: cuatro de tipo holísticos y dos componencial. La evaluación la realizan respecto de los resúmenes escritos de 390 estudiantes entre 14 y 16 años y los resúmenes escritos por seis humanos expertos. Los resultados obtenidos demuestran la viabilidad de desarrollar una herramienta computacional para la evaluación automatizada utilizando juicios humanos y LSA. Los resultados obtenidos son en general muy similares a los obtenidos por Kintsch et al. (2000). Obteniéndose para los resúmenes de textos narrativos una correlación promedio en los métodos LSA y los expertos de 0,58 (pearson) y para los expositivos una correlación de de 0,35 promedio. En general, el ANOVA realizado en esta investigación confirma que los métodos holísticos con LSA se correlacionan mejor con los expertos, en particular cuando se trata de resúmenes breves provenientes de textos narrativos.

Pérez et al. (2005) propone para la evaluación automatizada de textos el sistema de evaluación de respuestas en texto libre llamado Atenea (Alfonseca & Pérez, 2004) en el cual combinan una versión modificada del algoritmo BLUE *(BiLingual Evaluation Understudy algorithm)* de Papineni, Roukos, Ward y Zhu (2001) con LSA. En esta combinación entonces se considera la búsqueda de coincidencias de n-gramas entre la respuesta del estudiante y los documentos de referencias y además se calcula la similitud semántica. Este sistema es capaz de realizar preguntas, escogidas aleatoriamente o bien conforme al perfil del estudiante, y asignarles una calificación numérica. Los resultados de los experimentos han demostrado para todos los conjuntos de datos en los que las técnicas de PLN se han combinado con LSA la correlación de Pearson entre las notas dadas por Atenea y las notas dadas por los profesores para el mismo conjunto de preguntas alcanza a un valor de correlación promedio de 0,55, mejorando la evaluaciones anteriores que solo consideraban el uso de n-gramas y otras técnicas de NLP. Cabe señalar que en el uso del LSA en Atenea se ha utilizado un corpus en español traducido automáticamente al inglés utilizando Altavista Babelfish, por lo que en estricto las pruebas se realizaron en lengua inglesa (Perez et al. 2005).

Otras indagaciones sobre evaluación de resúmenes utilizando el LSA en español han sido las llevadas a cabo por Venegas (2006b, 2009a, b). En Venegas (2006b) se comparó el puntaje otorgado por el LSA con la evaluación realizada por tres docentes de lengua castellana, quienes utilizaron una pauta de 30 puntos basada en la presencia y ausencia de las ideas principales. En esta investigación los resúmenes evaluados fueron realizados por estudiantes de educación técnico-profesional (14-16 años), quienes

produjeron resúmenes provenientes de textos expositivos (divididos en resúmenes de alta y baja densidad informacional) y narrativos. El análisis con LSA fue realizado en términos del promedio de coseno obtenido entre cada resumen de los alumnos y los párrafos de los textos fuentes. Los resultados obtenidos mostraron una correlación positiva significativa entre los puntajes asignados por los evaluadores y el LSA ($\bar{x}$ 0,3 pearson). Sin embargo, los *tests* de comparaciones múltiples presentaron diferencias significativas entre la evaluación realizada por el LSA y los evaluadores en todos los tipos de textos considerados. En Venegas (2009a) se realiza algunas variaciones con respecto al trabajo anterior, entre ellas: 1) Se dividieron los resúmenes entre aquellos que alcanzaron un nivel alto y un bajo nivel de logro en la evaluación realizada por un equipo de investigadores en comprensión (Fondecyt 1020786). Para la segmentación se utilizó un umbral del 60% de logro según los puntajes asignados a partir de una pauta de 30 puntos que consideró la presencia y calidad de las ideas contenidas en los textos fuente. A diferencia de la investigación anterior, los puntajes asignados fueron consensuados por los investigadores. 2) Solo se considero la segmentación entre resúmenes provenientes de textos narrativos y expositivos (no se consideró la variable densidad lingüística al interior de ellos). 3) Se utilizaron tres métodos de asignación de puntaje con LSA: a) resumen-texto fuentes segmentado en párrafos; b) resumen-texto completo (sin segmentación en párrafos) y c) resumen-resumen consensuado por los investigadores. Así como se segmentaron los resúmenes según los puntajes asignados por los evaluadores, para los valores obtenidos con cada método se estableció también un umbral (60%) en relación con el valor más alto de coseno, según el cual los resúmenes se dividieron entre resúmenes con alto nivel de logro (1) y bajo nivel de logro (0). Luego esta clasificación se comparó con los resultados entregados por los investigadores. En este caso se calculó la precisión, la exhaustividad y los valores de F1 para la comparación entre los resultados obtenidos por los métodos y los obtenidos por los profesores. Los resultados generales obtenidos permiten establecer que el método que calcula la similitud semántica entre el resumen de cada alumno y el texto completo es el que mejor representa la evaluación realizada por los investigadores a los resúmenes (F1=0,64). Este método, además, es el que más asemeja la evaluación humana cuando los resúmenes provienen de textos expositivos (F1=0,55). En el caso de los resúmenes provenientes de textos narrativos, el método que considera el resumen consensuado es el que más se asemeja a la evaluación humana (F1=0,79), aunque sin mayor diferencia con el segundo método (F1=0,74). En suma, en esta investigación, se pudo establecer que el método que calcula la similitud semántica entre el resumen de los estudiantes con el texto completo presenta valores muy similares a los entregados por la evaluación humana, en particular cuando estos son considerados con un alto nivel de logro por los humanos. Lo anterior confirma los resultados obtenidos por León et al. (2000) en cuanto a la utilidad del método holístico y en cuanto a la mayor correspondencia entre la evaluación realizada con LSA y por los evaluadores humanos a resúmenes provenientes de textos narrativos.

5. Procedimientos

5.1. Los resúmenes

En esta nueva investigación se utilizan 373 resúmenes, realizados por alumnos de entre 15 y 16 años, pertenecientes a Liceos de educación secundaria técnico-profesional de la ciudad de Valparaíso, Chile. Tal como en Venegas (2006b y 2009a) los resúmenes fueron realizados a partir de textos de tipo expositivo (en tres áreas técnico-profesionales) y de tipo narrativo (uno común a todas las áreas). En total se utilizaron 7 textos fuente, con un largo promedio de 1024 palabras. Cabe mencionar que los resúmenes fueron digitalizados respetando la sintaxis y las ideas de los alumnos, sin embargo, los errores ortográficos fueron corregidos con el fin de que el programa LSA pueda calcular automáticamente la similitud léxico-semántica entre cada resumen y los textos fuente.

Tabla 1. Números de resúmenes por tipo de texto y área

Tipo de texto	Resúmenes
Expositivo	225
Narrativo	148
Total	373

En cuanto al tamaño de los resúmenes, en términos de palabras, los valores descriptivos corresponden a los siguientes: Mínimo=8, Máximo= 141, Mediana= 67, Promedio= 65,43. La compresión promedio alcanza a un 6,39%.

Estos resúmenes se encuentran disponibles en www.elgrial.cl y corresponden al denominado corpus DETP-2004 (Discurso Escrito Técnico- Profesional).

5.2. El espacio semántico y cálculo de similitud léxico semántica

En la construcción del espacio semántico se consideró un corpus de textos del español diversificado denominado COTEGE (Corpus del Español General) constituido por 10.242.384 palabras. Este corpus se conformó a partir de cinco corpora multiregistro: 1) **Corpus PUCV-2003:** Corpus recolectado por el equipo de investigación FONDECYT 1020786. La conformación general del Corpus PUCV-2003 se desglosa en 90 textos que equivalen a 1.466.744 palabras. Este corpus general está dividido, a su vez, en tres grandes registros o subcorpora (Corpus Técnico-Científico -CTC-, Corpus de Literatura Latinoamericana Escrita -CLL-, y Corpus de Entrevistas Orales -CEO-). Este corpus está disponible en www.elgrial.cl. 2) **Corpus Oral del Castellano:** Corpus de textos recolectados y transcritos por la Universidad Autónoma de Madrid. Consta de 1.099.400 palabras e incluye 12 géneros orales. Este corpus está disponible en www.lllf.uam.es/corpus/corpus.html. 3) **Corpus de Referencia del Español Contemporáneo**: Corpus elaborado por la Universidad Autónoma de Madrid y está compuesto por el Corpus Lingüístico de Referencia de la Lengua Española en Chile y el Corpus de Referencia de la Lengua Española en la Argentina. En este corpus se recogen un total de 3.156.491 palabras de 10 tipos de géneros textuales

diferentes. Este corpus está disponible en ftp://ftp.uba.ar/pub/misc/corpus/. 4) **Corpus de Narraciones Escritas:** Corpus construido en esta investigación con objeto de realizar las primeras pruebas de los sistemas computacionales. Este corpus cuenta con 86.3981 palabras e incluye tres obras escritas narrativas (la Biblia, Alicia en el País de las Maravillas e Historia de dos Ciudades). Está disponible en www.elgrial.cl. **5) Corpus ARTICO**: Corpus de Artículos de Investigación Científica Originales está constituido por 678 artículos publicados en revistas indexadas ScIELO entre el año 2000 y 2003. Tiene un total de 3.655.768 de palabras (disponible en www.elgrial.cl). Cabe mencionar que esta conformación del corpus para su posterior transformación en espacio semántico se justifica en relación a los hallazgos de Olmos, León, Escudero y Botana (2009) en los que se establece que la dependencia de domino temático afecta la eficiencia de la evaluación automatizada utilizando LSA. ES-COTEGE luego de la aplicación del SVD quedó constituido por 297 dimensiones y 99.966 palabras únicas, con la asignación de sus correspondientes valores y está disponible para ser utilizado en ww.elgrial.cl. (sección comparación de textos o directamente en: http://158.251.61.111/webapps/compareFiles/).

Con este espacio semántico se calculó el valor de coseno para cada uno de los 373 resúmenes comparándolos a partir de tres métodos. El primer método consideró la comparación entre el resumen y el texto fuente completo. El segundo método consideró la comparación entre el resumen de cada alumno con un resumen consensuado, por parte de un grupo de investigadores, de cada texto fuente. Finalmente, el tercer método consideró la comparación de cada resumen con tres resúmenes escritos por quienes serían los docentes evaluadores de los resúmenes (no se consideró el consenso entre ellos). Como se observa en esta investigación se retoman dos de los métodos utilizados en Venegas (2009a), no considerándose el método de segmentación de texto fuente en párrafos e incluyéndose los resúmenes realizados por los docentes.

5.3. Pauta y entrenamiento de evaluadores

A diferencia del trabajo realizado en Venegas (2009a), en esta investigación se construyó una pauta de evaluación de los resúmenes, según cada uno de los textos de origen. Para la construcción de las pautas de evaluación, acorde a cada texto fuente, se consideraron los siguientes criterios de tipo semántico-cognitivo, acorde con los modelos de Kintsch (1988, 1998): Presencia de las ideas principales, Integración de las ideas, Generalización de las ideas, Secuencia lógica, Macrosemantización (Venegas, 2009b, Órdenes, 2009). La pauta utilizada está construida en base a una escala con un intervalo de seis puntos (McCarthy & McNamara, 2008). Los puntajes fueron interpretados del siguiente modo: 1: Sí cumple absolutamente con el criterio; 2: Sí cumple medianamente con el criterio; 3: Sí cumple suficientemente con el criterio, 4: No cumple suficientemente con el criterio, 5: No cumple mínimamente con el criterio y 6: No cumple en nada con el criterio. La pauta fue validada (85% de acuerdo) y mejorada según la opinión de tres jueces expertos.

Para la evaluación de los resúmenes se seleccionaron y capacitaron tres docentes de Lengua Castellana (todos con formación de postgrado). Esta capacitación se llevó a cabo durante el mes de diciembre 2008 y enero de 2009, con un total de 60 horas. En esta capacitación se revisaron los aspectos teóricos relacionados con la evaluación de la comprensión desde una perspectiva sociocognitiva, se presentaron y explicaron los criterios de evaluación, se aplicó la pauta a una muestra de resúmenes, se establecieron consensos en la aplicación de los criterios y asignación de puntajes, se reevaluó una nueva muestra de resúmenes y se calibraron las asignaciones de puntaje.Todo esto se realizó con el fin de evitar el grado de impacto de ciertas variables, como por ejemplo, la tendencia inherente de cada uno de ellos a evaluar ciertos aspectos en desmedro de otros o evaluar aspectos no considerados en la pauta.

6. Resultados

El primer resultado que se presenta son los valores de correlación obtenidos por los docentes en la evaluación de los resúmenes, provenientes tanto de textos expositivos como narrativos (α de Cronbach= 0,838). La Tabla 2 sintetiza los resultados y muestra la correlación promedio obtenida entre ellos.

Tabla 2. Correlación entre los docentes evaluadores

Tipos de Textos	Correlación entre docentes
Expositivo	0,583**
Narrativo	0,702**
Todos	0,633**

** La correlación es significativa al nivel 0,01 (N=373, pearson, bilateral).

Las correlaciones observadas demuestran una consiste evaluación de los resúmenes de tipo narrativo por parte de los docentes, lo que es congruente con los datos de León et al. (2005). Esto último, confirma que los textos expositivos, además de ser más difíciles de comprender (Otero, León & Graesser, 2002; Parodi, 2005), presentan menor consistencia en la evaluación entre tres evaluadores entrenados.

A modo de comparación entre los métodos se llevó a cabo un estudio correlacional. Los resultados se observan en la Tabla 3.

Tabla 3. Correlación entre los tres métodos LSA

	Método 1	Método 2	Método 3
Método 1	1	,732**	,820**
Método 2		1	,880**
Método 3			1

** La correlación es significativa al nivel 0,01 (N=373, bilateral).

Este primer resultado permite plantear que evaluar la similitud semántica entre los resúmenes de los alumnos utilizando cualquiera de los tres métodos es relativamente similar. En términos computacionales esto es muy relevante, pues el Método 1 (resumen-texto completo) es muy económico en términos de procedimientos y muy fácil de implementarse computacionalmente, pues solo requiere del texto fuente, lo que es congruente la investigación realizada en (Venegas, 2009a). No obstante lo anterior, cabe desatacar que el Método 2 (resumen-resumen consensuado) y el Método 3 (resumen-resúmenes de los docentes) se correlacionan de modo muy fuertemente entre sí, haciendo que los datos aportados por ambos métodos sean muy parecidos.

A partir de los datos obtenidos de la evaluación realizada por los docentes y la asignación de puntaje otorgado a cada resumen se ha llevado a cabo un análisis correlacional entre los evaluadores y entre el promedio otorgado por los evaluadores y cada método.

La Tabla 4 muestra los resultados correlacionales obtenidos entre los evaluadores y entre los evaluadores y cada uno de los métodos, de acuerdo con el tipo de texto de origen.

Tabla 4. Comparación entre docentes y los tres métodos LSA

Tipo de texto	Evs.	Evs. y Método 1	Evs. y Método 2	Evs. y Método 3
Expositivo (n=225)	0,583**	0,394**	0,429**	0,471**
Narrativo (n=148)	0,702**	0,470**	0,585**	0,604**

**La correlación es significativa al nivel 0,01 (N=373, bilateral)

Como se observa, todas las correlaciones son significativas y positivas. Las más altas se dan entre los docentes evaluadores y el Método 3 (Resumen-Resúmenes de los docentes). Además, nuevamente se halla diferencia a favor de los resúmenes provenientes de textos narrativos(r=0,604). Entre los docentes y el Método 2 (resumen-resumen consensuado) se presentan valores de correlación con magnitudes algo menores que las del Método 3. Las magnitudes de correlación entre los docentes evaluadores y el Método 1 son menores que en los otros métodos. En todos los casos se presentaron correlaciones más bajas entre los docentes evaluadores y los tres métodos cuando los resúmenes provienen de textos predominantemente positivos.

Los datos de correlación obtenidos entre los evaluadores y los métodos muestran en todos casos correlaciones significativas y positivas con magnitudes variables mayores a 0,39 (pearson). Esto en principio permite confirmar que los métodos evaluarían de modo

similar a los docentes evaluadores, siendo el Método 1 el menos similar y el Método 3 el más similar. Sin embargo, dado que los valores de correlación entre los docentes son más altos debemos confirmar si existe o no diferencia estadísticamente significativa entre los valores de correlación de estos con los obtenidos con cada uno de los métodos de LSA. La Tabla 5 muestra los valores de probabilidad obtenidos en la comparación de los coeficientes de correlación, según cada método y tipo de texto.

Tabla 5. Comparación de los coeficientes de correlación obtenidos entre los docentes y los obtenidos entre los docentes y cada método.

Valores de P en Fisher Z-Transformation	Evs. y Método 1	Evs. y Método 2	Evs. y Método 3
Expositivo (n=225)	0,0083**	0,0281	0,1010
Narrativo	0,0021**	0,0866	0,1434
(n=148)			

** La diferencia es significativa a nivel de 0,01.

Los resultados de la comparación de los coeficientes de correlación permiten establecer que el Método 1 (resumen-texto completo) se diferencia significativamente de la correlación obtenida por los docentes evaluadores en la evaluación de cada uno de los tipos de resúmenes. Por otra parte, el Método 2 (resumen-resumen consensuado) no se diferencia significativamente de la correlación obtenida por los docentes entre sí cuando se evalúan resúmenes provenientes tanto de textos expositivos como narrativos. Finalmente, los valores de correlación entre los docentes y el Método 3 tampoco se diferencian significativamente de los valores obtenidos por los docentes entre sí.

7. Conclusión

Los resultados obtenidos nos permiten concluir que la evaluación automatizada de resúmenes utilizando LSA es muy similar a la realizada por tres docentes entrenados en la evaluación de resúmenes provenientes tanto de textos, predominantemente, expositivos como narrativos en español. Esto cuando se considera tanto la comparación entre el resumen a evaluar y un resumen consensuado (Método 2) o los resúmenes escritos por los docentes que luego harán de evaluadores (Método 3). Cabe aclarar que los docentes escribieron sus resúmenes antes de iniciar el proceso de capacitación y evaluación.

Por otra parte, es muy relevante el hecho de que los resúmenes provenientes de textos narrativos tanto en las evaluaciones realizadas por los docentes como las realizadas por los métodos con LSA se presente mayor consistencia en las evaluaciones que en los provenientes de textos expositivos, relevándose no solo la mayor dificultad en la comprensión de estos últimos sino que también en su evaluación. Tal situación confirma los planteado en Kintsch et al. (2000), León et al. (2005) y Venegas (2009a).

En cuanto al Método 1, si bien este es computacionalmente menos costoso y se correlaciona positiva y significativamente con la evaluación de los docentes, presenta diferencias significativas en los valores de correlación, lo que significa que la evaluación con este método no es similar a los resultados de la evaluación de los docentes entrenados. Este resultado, un tanto inesperado, se aleja de lo observado en (Venegas 2009a) y se explicaría por el hecho de que tanto la pauta de evaluación como la capacitación de los evaluadores hace que las correlaciones entre los evaluadores sea muy fuerte y se diferencia de la correlación entre los evaluadores el este método holístico.

Finalmente, podemos establecer que estos resultados permiten confirmar que al menos dos de los tres métodos holísticos pueden ser utilizados en la evaluación de resúmenes relativamente breves, considerando su contenido semántico. Esto permite proyectar un sistema computacional que no solo considere la asignación de puntajes en base al uso de LSA sino que también provea de retroalimentación a los estudiantes. Lo anterior implica además considerar otros aspectos de la evaluación automatizada de resúmenes (ortografía, estilo, cohesión) que en los que se deberá considerar otras técnicas de NLP.

Referencias

Alfonseca, E. & Pérez, D. (2004). "Automatic assessment of short questions with a Bleu-inspired algorithm and shallow NLP". Ponencia presentada en 4th International Conference, EsTAL 2004, Alicante, España.

Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K. & Harshman, R. (1990). „Indexing by latent semantic analysis". *Journal of the American Society for Information Science, 41*(6), 391-407.

Farr, R. & Carey, R. (1986). *Reading. What can be measured*. Newark, Delawere: IRA.

Kintsch, E.; Steinhart, D.; Stahl, G.; LSA Research Group; Matthews, C. & Lamb, R. (2000). "Developing summarization skills through the use of LSA-based feedback". *Interactive learning enviroments, 8*(2), 87-109.

Kintsch, E.; Steinhart, D.; Stahl, G.; LSA Research Group; Matthews, C. & Lamb, R. (2000). "Developing summarization skills through the use of LSA-based feedback". *Interactive learning enviroments*, *8*(2), 87-109.

Kintsch, W. (1988). "The role of knowledge in discourse comprehension construction-integration model". *Psychological Review, 95,* 163-182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W. (2000). "Metaphor comprehension: A computational Theory". *Psychonomic Bulletin & Review, 7-2,* 257-266.

Kintsch, W. (2001). "Predication". *Cognitive Science*, *25*(2), 173-202.

Kintsch, W. (2002). "On the notions of theme and topic in psychological process models of text comprehension". En Louwerse, M. & van Peer, W. (eds.). *Thematics: Interdisciplinary Studies* (pp. 151-170). Amsterdam: Benjamins.

Landauer, T.K.; Laham, D. & Foltz, P.W. (2003). "Automated Essay Scoring: A Cross Disciplinary Perspective". En Shermis, M. & Burstein, J. (eds.). *Automated Essay Scoring and Annotation of Essays with the Intelligent Essay Assessor.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Landauer, T.K. & Dumais, S.T. (1997). "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge". *Psychological Review*, *104,* 211-240.

Landauer, T.K. & Psotka, J. (2000). "Simulating text understanding for educational applications with latent semantic analysis: Introduction to LSA". *Interactive Learning Environments*, *8*(2), 73-86.

Landauer, T.K.; Laham, D.; Rehder, B. & Schreiner, M.E. (1997). "How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans". En Shafto, M.G. & Langley, P. (eds.), *Actas de The 19th annual meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, N.J.: Erlbaum.

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

Lin, Ch. & Hovy, E. (2003). "Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics". En *Proceedings of 2003 Language Technology Conference*. Edmonton, Canadá.

Lin, Ch. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries". En *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*. Barcelona, Spain.

McCarthy, Ph. & McNamara, D (2008). "The User-Language Paraphrase Challenge". [En line]. Disponible en: https://umdrive.memphis.edu/pmmccrth/public/Phil%27s%20papers.htm?uniq=-xq6brv

Marinkovich, J.; Peronard, M. & Parodi, G. (2006). "Programa de optimización de la competencia estratégica para comprender y producir textos escritos (LECTES)". [En línea]. Disponible en: www.lectes.cl

Olmos, R.; León, J.; Escudero, I. & 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.

Órdenes, A. (2009). "El resumen como instrumento de evaluación de la comprensión del discurso escrito: correlación entre evaluadores expertos". Tesis de pregrado. Pontificia Universidad Católica de Valparaíso.

Otero, J.; León, J. & Graesser, A. (2002*). The psychology of science text comprehension*. Mahwah: Erlbaum.

Parodi, G. (ed.) (2005). *Discurso especializado e instituciones formadoras*. Valparaíso: EUVSA

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

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.

Peronard, M. (1989). "Estrato social y estrategias de comprensión de lectura". *Lenguas Modernas, 16,* 19-32.

Peronard, M.; Gómez, L.; Parodi, G. & Núñez, P. (eds.) (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.

Rehder, B.; Schreiner, M.E.; Wolfe, M.B.; Laham, D.; Landauer, T.K. & Kintsch, W. (1998). "Using latent semantic analysis to assess knowledge: Some technical considerations". *Discourse Processes, 25,* 337-354.

Shermis, M. & Burstein, J. (2003). *Automated essay scoring: a cross-disciplinary perspective*. Mahwah: Erlbaum.

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

Van Dijk, T. (1978). *La ciencia del texto*. Buenos Aires: Paidós.

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

Venegas, R. (2005). "Las relaciones léxico-semánticas en artículos de investigación científica: Una aproximación desde el análisis semántico latente". Tesis doctoral, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

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

Venegas, R. (2006b). "Comparación de la evaluación realizada por docentes y por el análisis semántico latente. Informe de Investigación DI 184.719/2006". Pontificia Universidad Católica de Valparaíso.

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

Venegas, R. (2009a). "Towards a method for assessing summaries in Spanish using LSA". En Chad Lane, H. and Guesgen, Hans W. (eds.), Proceedings of the Twenty-Second International Florida Artificial Intelligence Research Society Conference (pp. 113-115). Washington, DC: AAAI.

Venegas, R. (2009b). Informe de Investigación Final Fondecyt Nº 11070225 "Evaluación de resúmenes en español: correspondencia entre profesores y el análisis semántico latente".

Fecha de recepción: 15/12/09
Fecha de aceptación: 10/05/10