Universidad de Concepción
Dirección de Postgrado
Facultad de Ingeniería
Programa de Magíster en Ciencias de la Computación

# ENSAMBLE DE MODELOS SINTÁCTICOS Y SEMÁNTICOS PARA LA EVALUACIÓN AUTOMÁTICA DE ENSAYOS

por

**DIEGO ANDRÉS PALMA SÁNCHEZ**

**Patrocinante:** John Anthony Atkinson Abutridy

Tesis para optar al grado de

MAGÍSTER EN CIENCIAS DE LA COMPUTACIÓN

Enero, 2017
Concepción, Chile

## Agradecimientos

Me gustaría agradecer a mi familia, en especial a mi madre que me dio su apoyo cuando decidí cambiar de rumbo y comenzar estudios de postgrado y a Gabriela por todo su apoyo incondicional y por darse el tiempo de leer y ayudarme a redactar esta tesis.

También agradezco a mi profesor guía John Atkinson, que fue un mentor en toda esta etapa de postgrado y me transfirió bastante conocimiento, no sólo en el tema teórico, si no que también en otros ámbitos que me ayudaron a crecer en este período de postgrado. Por otro lado, sin su ayuda esta tesis no sería posible.

Agradezco a la gente del Departamento de Ingeniería Informática y Ciencias de la Computación, y a la dirección de Postgrado por su disposición. También me gustaría agradecer a la gente que conocí en el laboratorio de Inteligencia Artificial, ya que las discusiones que surgieron me ayudaron a aclarar mi cabeza en momentos de duda.

Finalmente, agradezco a CONICYT por financiar mis estudios de postgrado a través de la beca de Magíster nacional.

## Resumen

La evaluación automática de ensayos (descriptivos y argumentativos) es el problema de determinar de forma automática la calificación que un humano le daría a dicho ensayo bajo una rúbrica dada. Ello se logra mediante la detección de elementos específicos del texto. Una evaluación con tales características debe ser capaz de considerar las propiedades fundamentales de los textos que son la coherencia y la cohesión. Hasta ahora se han desarrollado métodos que permiten hacer esto de forma parcial, principalmente basados en propiedades estadísticas de los textos, siendo la más común el conteo de palabras. En este trabajo, se propone un nuevo método de evaluación automática que considere las distintas componentes de la coherencia textual. La propuesta incorpora métodos basados en teoría lingüística para el análisis de discursos. Trabajos anteriores mostraron que integrando componentes de la semántica de un texto se pueden obtener mejores resultados que sólo considerando propiedades estadísticas, sin embargo, no se han utilizado combinaciones de métodos basados en teoría de discurso para esta tarea. Los resultados experimentales mostraron que el método propuesto supera en términos de correlación con humanos a un método que utilice puramente características superficiales. Además, el método propuesto es competitivo con sistemas actuales de evaluación de ensayos, superando a la mayoría de los enfoques tradicionales.

# Tabla de Contenidos

# Índice de tablas

# Índice de figuras

# Capítulo 1

## Introducción

La escritura es una habilidad que se adquiere a temprana edad, pues se nos enseñan las letras, las palabras, las oraciones, etc. Sin embargo, esta habilidad no se desarrolla por completo, pues lo que se enseña no es suficiente para expresar claramente lo que se piensa, y como consecuencia nace la necesidad de saber redactar y/o de exponer de manera coherente y precisa las ideas. En la actualidad, la capacidad de redacción y comprensión que debiesen tener las personas que egresan del sistema escolar es un tema ampliamente debatido, tales como la carencia en el manejo del lenguaje escrito en todos los niveles educacionales y estratos socioculturales. Una mala capacidad de redacción tiene consecuencias relevantes como por ejemplo reprobar un examen porque las ideas expresadas no están claras. Por otro lado, una persona podría perder una oportunidad laboral debido a una mala redacción de un texto. En síntesis, ideas que podrían ser bastante buenas e innovadoras podrían llegar a verse opacadas o, peor aún, rechazadas por el receptor al no ser comunicadas adecuadamente.

Un texto se produce en función de un lector, con el objetivo de lograr comprensión sobre un tema que se busca comunicar. Por otra parte, debe haber relaciones entre las ideas planteadas dentro del texto, para lograr asegurar un significado claro del mismo y legibilidad. Para esto, existen dos características que los buenos textos deben tener: *coherencia* y *cohesión* [13]. La *coherencia textual* es una propiedad del texto que define las conexiones semánticas entre unidades de información y está relacionada con la representación mental que el lector tenga del mismo. Esta conexión se da tanto localmente a nivel de oraciones adyacentes, como globalmente (texto completo). Por otro lado, la *cohesión* constituye un conjunto de recursos léxicos y gramaticales que enlazan una parte del texto con otra, y por esto, es uno de los factores fundamentales para determinar si un texto puede ser considerado como tal, y no una sucesión de oraciones inconexas.

Una forma de mejorar las capacidades para formular adecuadamente las ideas en un texto es "practicar", realizando producciones textuales para que sean evaluadas y corregidas por un especialista humano y, a través de sucesivas repeticiones perfeccionar la calidad del texto producido. En este sentido, se debe tener en cuenta la diferencia entre corrección y evaluación:

- **Corrección**: Ayuda a que un estudiante mejore sus habilidades de escritura mediante la revisión de sus textos. El objetivo es corregir errores y avanzar en el manejo de estructuras y recursos lingüísticos necesarios para elaborar textos de mejor calidad y que expresen mejor las ideas.

- **Evaluación:** Busca determinar el nivel de competencias que tiene un estudiante para producir un texto, según un marco de evaluación definido.

Debe tenerse en cuenta que la evaluación de textos es una tarea costosa en términos de tiempo y personal requerido. Además, no existe otra forma de evaluar mejor el aprendizaje de un estudiante que no sea mediante la expresión de sus ideas a través de un escrito, por lo que se debe repetir el ejercicio constantemente en el tiempo.

Para reducir los costos del personal requerido para revisar evaluaciones textuales a gran escala, se han propuesto métodos para evaluar textos de manera automática, tarea conocida como *Evaluación Automática de Ensayos* [18]. Esta tarea se entiende como un problema de regresión sobre el dominio de valores posibles para las calificaciones. Los enfoques tradicionales para esta tarea intentan evaluar características relacionadas a la calidad del texto tales como el uso de vocabulario, coherencia, entre otros aspectos. Para ello, se utilizan propiedades superficiales del texto como por ejemplo: conteo de palabras, conteo de signos de puntuación, largo promedio de las palabras del texto, entre otros. Sin embargo, este tipo de enfoque tiene algunas debilidades [10]:

- No evalúa la estructura sintáctica del texto, pues no considera el orden de las palabras. Por ejemplo, la oración *"El árbol está seco."* sería equivalente a *"seco el está árbol."* Un evaluador humano consideraría estas oraciones como diferentes.

- No considera la coherencia textual, pues las características superficiales utilizadas no consideran el contenido ni el orden en que las ideas se plantean dentro del texto. Como consecuencia, malos textos que cumplan ciertas regularidades podrían ser bien evaluados (por ejemplo un ensayo con muchas palabras).

- No considera la cohesión de un texto en términos del correcto uso de recursos lingüísticos para expresar las ideas. Por ejemplo el texto: "*Los beneficios de la siesta son bien conocidos, aunque parece que quedan algunas cosas por aclarar. Manfred Walzl, neurólogo austriaco, pone en marcha un estudio; con un estudio él pretende demostrar que la siesta aumenta la productividad laboral*", tiene problemas de cohesión. Por ejemplo, se repite la palabra *estudio* cuando podría omitirse, esta repetición dificulta la lectura del texto. Por otro lado, no queda claro que la segunda aparición de dicha palabra se refiera a lo mismo que se refiere la primera. El pronombre *él* aparece innecesariamente y es redundante. Estos problemas pasan desapercibidos si sólo se considera características relacionadas a la frecuencia de términos.

Por lo tanto un problema en los métodos de evaluación automática de ensayos tradicionales es que no logran evaluar la coherencia textual de forma directa. La *evaluación automática de coherencia textual* es un problema de investigación que aún se encuentra abierto y tiene múltiples aplicaciones, como por ejemplo la generación automática de resúmenes, traducción automática, generación de textos, entre otros [26][1][34].

La teoría que intenta modelar los textos en base a cierta estructura se conoce como *teoría de discursos* e intenta encontrar un modelo que logre describir la coherencia textual. En la literatura existen tres modelos que se han utilizado para abordar esto: *Teoría del Centrado* [13], *Rethorical Structure Theory* [21], y *modelos semánticos basados en contenido textual* [8][35].

La *Teoría de centrado* caracteriza textos que puedan considerarse coherentes basándose en la forma en que se introducen y discuten *entidades de discurso*, que generalmente incluyen: nombres (por ejemplo: Juan), descripciones (por ejemplo: "El hombre barbudo"), pronombres (él, ella). Algunos problemas que tienen los métodos que utilizan esta teoría están relacionados con la ambigüedad que presentan algunos textos, por

ejemplo: *Daniela invitó a Susana a su casa, y ella le preparó un rico almuerzo*. En este caso, el pronombre *ella* podría hacer referencia a *Daniela* o a *Susana*.

*Rethorical Structure Theory* caracteriza la coherencia mediante relaciones existentes entre una entidad principal de un texto y el texto que hace referencia a dicha entidad. El modelo define un conjunto de relaciones las cuales pueden ser detectadas mediante *marcadores de discurso*. Estos marcadores de discurso son palabras o frases utilizadas para conectar y organizar segmentos de un texto (algunos ejemplos son: *porque*, *por lo tanto*, etc.). Estos marcadores de discurso establecen relaciones entre dos unidades textuales, por ejemplo: *Pedro estaba triste, porque perdió su juguete favorito*. En este caso, el marcador *porque* establece una relación de causa y consecuencia entre la *tristeza de Pedro* y el hecho de que *perdió su juguete*. Algunos problemas prácticos que existen en este modelo es que hay marcadores de discurso que tienen más de un propósito o se mapean a más de una relación, dependiendo de lo que se está expresando en el texto, por lo que se podrían detectar relaciones erróneas.

Los *modelos semánticos basados en contenido textual* representan los textos a través de sus características léxicas, generalmente utilizando palabras que representen el contenido del texto (*modelo de espacio vectorial*). En dicha representación, se define una medida de similitud entre fragmentos de un texto. La coherencia se mide como el grado de similitud que existe entre partes consecutivas del texto, considerando el supuesto que textos coherentes tienden a expresar ideas similares entre oraciones.

Se han realizado estudios que comparan el rendimiento de los distintos modelos para evaluar la coherencia textual correlacionándolos con evaluaciones humanas, y se ha concluido que no existe modelo que evalúe todos los aspectos relacionados a la coherencia. Sin embargo, los métodos evalúan propiedades complementarias de coherencia (por ejemplo los modelos semánticos evalúan relaciones entre secuencias de palabras, y los modelos basados en discurso evalúan estructura del texto), por lo que podrían combinarse para evaluar la coherencia textual [26].

Tomando en consideración lo anterior, en esta tesis se propone un método de evaluación automática de ensayos que considere características de la coherencia textual basadas en teoría de discurso. Estas características reflejarán qué tan coherente es un texto considerando su estructura y en cómo se relacionan y se expresan las ideas

dentro del mismo. El método también considerará características semánticas, es decir, las relaciones existentes entre los conceptos utilizados dentro del texto.

## 1.1. Hipótesis

Este trabajo explora la hipótesis que un método de evaluación de textos que considere características basadas en la teoría de discursos en combinación con medidas semánticas, será más efectivo para la tarea de evaluación automática de ensayos en comparación a métodos que utilicen medidas superficiales estadísticas (como conteo de palabras, largo de las oraciones, etc.), considerando la correlación con humanos.

## 1.2. Objetivos

### 1.2.1. Objetivo General

Desarrollar un método computacional que permita evaluar automáticamente textos en forma de ensayos considerando aspectos de coherencia textual.

### 1.2.2. Objetivos Específicos

- Analizar estrategias de evaluación de ensayos en forma de texto, basados tanto en modelos de estadísticos, como en teoría de discurso.
- Desarrollar un método de evaluación automática de ensayos que considere coherencia a nivel de contenido y sintaxis.
- Crear un prototipo computacional.
- Evaluar el método propuesto.

# Capítulo 2

# Trabajo Relacionado

La evaluación automática de ensayos se ha estudiado durante muchos años, y ha avanzado con el desarrollo de nuevas técnicas de procesamiento del Lenguaje Natural (NLP). Lograr realizar esta tarea es relevante, pues tiene potenciales aplicaciones con alto impacto económico y social.

Esta tesis está enfocada en el problema de evaluar automáticamente ensayos textuales escritos por estudiantes, pues lograr realizar esta tarea podría generar un impacto en el ámbito educativo y serían posibles aplicaciones tales como sistemas tutoriales inteligentes para ayudar a mejorar la habilidad de los estudiantes al producir composiciones textuales.

Se discutirán los dos grandes enfoques para la evaluación automática de ensayos que recoge la literatura, los cuales son: *enfoque estadístico* y *enfoque basado en contenido textual*. Posteriormente se discutirán modelos utilizados para evaluar la coherencia textual, entre los que destacan el modelo de *entity grids* y el análisis de *cadenas léxicas*. Finalmente, se mencionarán sistemas actuales de evaluación de ensayos y sus principios de funcionamiento.

## 2.1. Enfoque Estadístico para Evaluación de Ensayos

Este enfoque utiliza como supuesto que existen ciertas características superficiales de los textos que se aproximan a características intrínsecas que los buenos textos debieran poseer (como coherencia y cohesión). En este caso, el problema de determinar qué nota debiese tener un ensayo es visto como un problema de regresión, es decir, cada característica del texto tendrá una cierta ponderación que servirá para determinar la nota final del ensayo. Para lograr esto, se determinan medidas estadísticas tales como: Conteo de palabras, conteo de oraciones, conteo de verbos, índices de

legibilidad, conteo de errores gramaticales, entre otras. Cada documento se representa como un conjunto de características $\{x_1, x_2, x_3, \ldots, x_n\}$, que sirven para predecir qué calificación le daría un humano al texto. Con esta representación, y utilizando un conjunto de ensayos evaluados previamente por humanos, se busca determinar un modelo matemático que logre predecir la nota del ensayo, por ejemplo el mostrado en la ecuación 2.1. En este ejemplo de modelo, cada característica $x_i$ del ensayo tiene una ponderación $w_i$, y en conjunto determinarán la nota del ensayo.

$$Nota = \sum_{i=1}^{n} w_i \cdot x_i \tag{2.1}$$

En la literatura se han estudiado diversas características que permiten lograr una buena predicción en la nota de un ensayo, si se compara con un evaluador humano. Los tipos de características consideradas son *características superficiales*, *legibilidad* [39], *diversidad léxica* [31], *características gramaticales*.

- Características Superficiales: Dentro de esta categoría, se encuentran características tales como el conteo de caracteres, largo del texto, conteo de palabras, largo promedio de las palabras, largo promedio de las oraciones. Se ha demostrado que este tipo de características son buenos predictores de la nota de un ensayo.

- Medidas de Legibilidad: Se define legibilidad de un texto, como la dificultad que tiene un humano para leer un texto. En la literatura se han propuesto medidas de legibilidad automática, las cuales se representan con un valor numérico denominado *índice de legibilidad* que determina el nivel educacional que una persona necesitaría para leer un texto. Algunas interpretaciones de estos índices de legibilidad es que valores altos indicarían que el texto es difícil de leer y podría influir en la nota que un evaluador humano le dará al ensayo.

- Medidas de Diversidad Léxica: Estas medidas determinan qué tan diverso es el vocabulario utilizado en un texto. La medida más común es el *Type Token Ratio*, que representa la razón entre palabras únicas de un texto (*tokens*) y la cantidad total de palabras del texto. Esta medida está sesgada al largo del texto pues en

textos más largos es más probable que aparezcan términos únicos, por otro lado en textos demasiado largos, la medida sólo se verá reflejada en cuántas palabras tiene el texto. Por ello se han propuesto medidas más sofisticadas tales como *Guiraud's Index*, *Yule's K*, *D estimate*, *hapax legomena* (cantidad de palabras que aparecen una sola vez en el texto). Se ha demostrado que estas medidas son un buen predictor en la nota de un ensayo, y aumentan el rendimiento de un sistema evaluador automático de ensayos [31].

- Medidas Gramaticales: Los buenos textos deben reflejar un buen uso de recursos léxicos y gramaticales, tales como uso de verbos y sustantivos para expresar las ideas. Estas medidas están relacionadas a la frecuencia de aparición de etiquetas gramaticales (verbos, sustantivos, adjetivos, etc.), y también a características tales como errores ortográficos y gramaticales, los cuales pueden ser detectados de forma automática.

## 2.2. Enfoque Basado en Contenido para Evaluación de Ensayos

Este enfoque intenta evaluar el ensayo a partir del contenido, con el fin de tener una aproximación de la semántica del texto. El supuesto utilizado es que ensayos con contenido similar, debiesen tener notas similares. Para lograr determinar la similitud de un texto se realiza una representación de los textos en función de las palabras que lo componen, y el modelo comúnmente utilizado es el *modelo de espacio vectorial* [35]. Este modelo tiene algunas desventajas que se discutirán en esta sección y se han propuesto algunas mejoras tales como el *Análisis Semántico Latente* y el *Análisis Semántico Latente Probabilístico*. En esta sección se describirán los modelos mencionados.

### 2.2.1. Modelo de Espacio Vectorial

El *modelo de espacio vectorial* [35] representa documentos en un espacio altamente dimensional, donde cada dimensión está asociada a términos relevantes del texto. Las palabras o términos que determinan las dimensiones deben ser previamente escogidas y este conjunto escogido se define como *vocabulario*. Luego, un documento se puede

representar como un vector donde cada componente representa la frecuencia en que un término $t_i$ aparece en dicho documento. A partir de esta representación vectorial se puede establecer una medida de similitud entre dos documentos, utilizando el coseno:

$$cos(d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\|\|d_j\|} \tag{2.2}$$

donde $d_i$ y $d_j$ son vectores que representan dos documentos.

En la figura 2.1 se muestra como ejemplo un espacio vectorial de 3 dimensiones y documentos representados en dicho espacio.

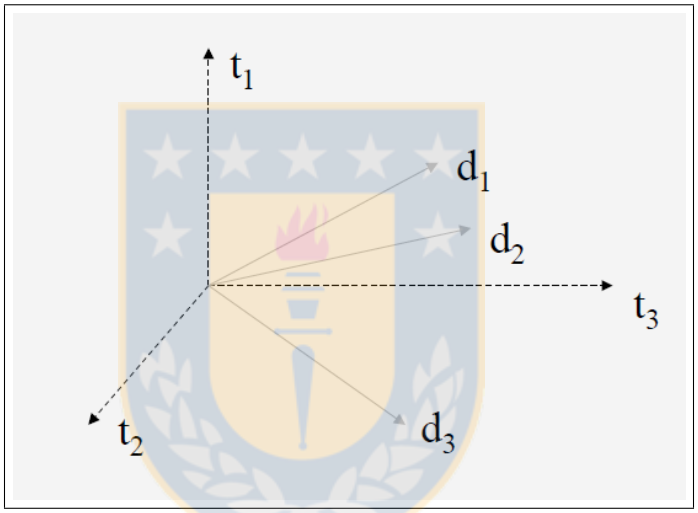


Figura 2.1: Representación de documentos en un espacio vectorial[1].

Utilizando este modelo, el problema qué nota le corresponde a un ensayo a evaluar se determina utilizando valores de similitud con otros ensayos evaluados previamente con humanos. Esto se puede realizar utilizando diferentes criterios, como por ejemplo:

1. La nota se calcula como la suma ponderada de los $k$ ensayos más similares.
2. Se pone una nota proporcional a la similitud con un ensayo *gold standard*, es decir, el ensayo ideal.
3. Se pueden utilizar técnicas estadísticas de regresión con los vectores generados para cada ensayo.

[1]https://www.coursera.org/learn/natural-language-processing

Los métodos de evaluación de ensayos basados en el modelo de espacio vectorial han logrado una correlación con humanos de hasta 0.5, en un conjunto de ensayos en que la correlación humano-humano fue de 0.6 [35].

Sin embargo, este modelo de representación posee varios problemas:

- *Polisemia* ($sim(d_i, d_j) < cos(d_i, d_j)$): Si en el vocabulario definido para representar los vectores existen palabras que tengan múltiples significados, es posible que la similitud calculada sobre-estime la similitud real entre los documentos (debido a que los términos se consideran equivalentes pues solo importa su frecuencia de aparición).

- *Sinónimos* ($sim(d_i, d_j) > cos(d_i, d_j)$): Dos documentos que sean similares en contenido pero que utilizan diferentes términos (sinónimos), se considerarán poco similares. En este caso, el coseno del angulo dependerá de qué tan cercanos se encuentran en el espacio vectorial, es decir, se subestima la similaridad real que existe entre dos documentos.

- *Relaciones entre palabras de los documentos*: Los humanos son capaces de establecer relaciones entre palabras, debido a que un lector tiene una representación mental clara de ellas. Por ejemplo, para un humano es evidente la relación existente entre las palabras *doctor/paciente/enfermera/tratamiento*. El modelo de espacio vectorial no logra detectar estas relaciones en forma clara, debido a la alta dimensionalidad de los espacios generados.

Un método que permite resolver parcialmente estos problemas es el *Análisis Semántico Latente* (LSA) [9], y ha sido utilizado en diferentes sistemas de evaluación automática de ensayos.

### 2.2.2. Análisis Semántico Latente

La motivación de LSA es que existen relaciones ocultas entre las palabras de un texto que podrían descubrirse al reducir la dimensionalidad de los textos reresentados originalmente. El método considera un modelo de *bolsas de palabras* (BOW), a partir de la cual se obtiene una matriz de palabras y documentos del corpus. Luego, se aplica

un método de *descomposición de valores singulares* (SVD), para obtener un espacio reducido o *espacio semántico*. En este espacio semántico se obtiene una representación vectorial de las palabras la cual se ha utilizado para tareas como evaluación de ensayos [18] y para medir coherencia textual [24] [19].

La evaluación de ensayos utilizando LSA se realiza en dos etapas:

1. Se genera un espacio semántico a partir de un corpus de documentos.

2. Compara ensayos pre-evaluados por humanos con ensayos nuevos utilizando algún criterio de similitud (por ejemplo coseno). Esta comparación utiliza espacio semántico previamente generado [24].

Luego, se puede asignar un puntaje (o nota) como la suma ponderada de los $k$ ensayos más similares. También se puede asignar una nota que sea proporcional al grado de similitud con un ensayo ideal. Otro método utilizado es establecer umbrales de similitud entre el ensayo a evaluar y documentos relacionados con el tópico a evaluar.

Un problema con LSA es que al utilizar un modelo de bolsa de palabras no considera el orden de las mismas, por lo que se considerarían equivalentes frases como "literatura fantástica" con "fantástica literatura". Para abordar este el problema, se ha propuesto una variación de LSA denominada *Generalized Latent Semantic Analysis* (GLSA) [22]. Este método utiliza frecuencia la de n-gramas (secuencias de $n$ palabras) en lugar de palabras, lo que permite distinguir segmentos de texto (como por ejemplo "dióxido de carbono" con "Carbono de dióxido"). Experimentalmente, la evaluación de ensayos realizada utilizando GLSA ha obtenido una correlación de hasta 0.8 con la evaluación realizada por humanos [22]. Sin embargo, existen algunas dificultades:

- Alta dimensionalidad del espacio vectorial debido a los n-gramas. En este caso, se consideran combinaciones de palabras, por lo que la dimensionalidad crece exponencialmente con el vocabulario considerado en la representación de los documentos.

- Alto grado de dispersión, pues la co-ocurrencia de n-gramas es menos probable que la de palabras.

- La descomposición de valores singulares es costosa computacionalmente y el costo crece exponencialmente dependiendo de los n-gramas a utilizar (Esto ocurre también en LSA convencional).

Otro problema importante con éste tipo de métodos es que no tienen una base estadística sólida, es decir, no se obtiene un modelo generativo de los datos. Esto trae como consecuencia una dificultad al interpretar el espacio semántico obtenido, lo cual es un punto clave si se quisiera implementar un sistema que entregue retroalimentación a los estudiantes. Existe una variación de LSA que resuelve este aspecto y que ha sido utilizada para evaluar ensayos, y se conoce como *Análisis Semántico Latente Probabilístico* (PLSA) [20].

### 2.2.3. Análisis Semántico Latente Probabilístico (PLSA)

Este método se fundamenta en un modelo probabilístico denominado *Aspect Model*. Este modelo probabilístico asocia *variables ocultas* con variables observadas. Las variables ocultas (o *latentes*) son variables que no se pueden observar a partir de los datos, pero que se pueden inferir mediante un modelo matemático a partir de las variables observadas (variables que se pueden medir). En el contexto de PLSA, las variables medibles son los documentos y las palabras los componen, y las variables ocultas son relaciones que se pueden descubrir (por ejemplo tópicos).

PLSA representa los datos en función de 3 variables:

- Documentos: Los documentos del *corpus* a utilizar, donde el corpus corresponde a un conjunto de documentos.
- Palabras: El vocabulario a considerar en el corpus.
- Tópicos: Variables ocultas (o latentes).

En la figura 2.2 se muestra una representación gráfica del modelo, en la que se describe el proceso generativo de $N$ documentos en el corpus. $N_w$ es la cantidad de palabras en un documento $d$. Cada palabra $w$ tiene asociada un tópico $z$ (variable latente) que la genera.

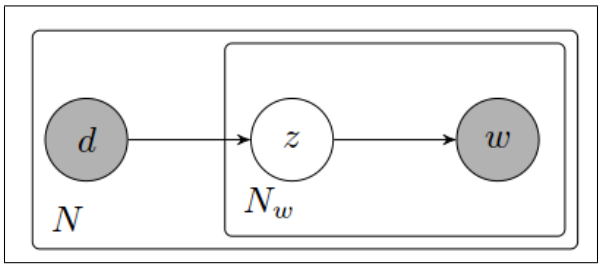


Figura 2.2: Representación gráfica de PLSA [20].

A diferencia de otros modelos, PLSA tiene la siguientes ventajas:

- Base estadística sólida, donde los parámetros están bien definidos y tienen interpretación probabilística clara (variables latentes como tópicos), a diferencia de otros métodos como LSA donde el espacio semántico no tiene una interpretación clara. Esto tiene como consecuencia, que se puede utilizar teoría estadística para la selección del modelo (cantidad de tópicos a considerar, es decir, la dimensionalidad del espacio semántico). Esto es diferente de LSA, donde la selección del modelo (cantidad de dimensiones) se realiza de forma empírica escogiendo entre 100-300 dimensiones.

- Manejo de la polisemia, pues se tienen grupos de palabras que representan un determinado tópico (variable oculta).

## 2.3. Coherencia Textual

Una de las falencias de los métodos de evaluación de ensayos basados en los enfoques anteriores, es que al sólo considerar características superficiales del texto, no consideran el orden en el que se van exponiendo las ideas [26]. Es por ello que no obtienen una buena aproximación a la coherencia y cohesión del texto, por ejemplo, los siguientes textos:

1. *Generalmente los hombres risueños son sanos de corazón. La risa de un niño es como una loca música de la infancia.*

2. *Yerma, la obra teatral de Federico García Lorca, presenta el conflicto sicológico de una mujer estéril. Me gustan las obras teatrales; pero más aún me gustan las poesías.*

Se observa que el texto número uno mantiene una cierta temática entre las oraciones (la *risa*) y existe conexión entre ambos segmentos de texto. Sin embargo, en el texto número dos, se observa un cambio que un lector no entendería, pues la idea de la segunda oración no se relaciona con lo mencionado en la primera. Un sistema evaluador de ensayos que utilice los métodos descritos anteriormente, no detectaría estos errores y podría evaluar incorrectamente un ensayo. Por ello, nace la necesidad de tener medidas de coherencia textual más directas.

Existen diversos métodos para aproximar la coherencia textual de forma automática, y los más aceptados en la literatura están basados en la *teoría del centrado* [1][34][13]. Los métodos más utilizados son *entity grids* y *análisis de cadenas léxicas*, los cuales se discutirán en esta sección.

### 2.3.1. Teoría del Centrado

La teoría del centrado es una teoría que intenta modelar *discursos*. Un discurso es una expresión formal de un acto comunicativo, que se presenta bajo manifestaciones diversas (por ejemplo discurso oral o escrito). Un discurso tiene como finalidad comunicar algo y para ello tiene una estructura dinámica (por ejemplo introducción, desarrollo, cierre).

La teoría del centrado establece que un discurso está conformado por segmentos de texto que están conformados por enunciados (*utterances* $U_i$ - $U_n$). Cada enunciado ($U_i$) tiene asociado un conjunto de *entidades* de discurso o *centros* del discurso denominado *forward looking centers* $C_f(U_i)$. Estas entidades por lo general son nombres, que tienen un rol gramatical (Sujeto, Objeto, Objeto Indirecto, entre otros). Los elementos de $C_f$ se ordenan de acuerdo a su importancia dentro del discurso, esta importancia está relacionada al rol gramatical que cumple dicha entidad en el enunciado $U_i$. El centro de $C_f$ que tiene la mayor importancia se denomina $C_p$ o el centro preferido. También se define un centro que relaciona el enunciado actual con el anterior, y se denomina *backward looking center* $C_b$. El centro $C_b$ es el centro preferido del enunciado previo

$U_{i-1}$ y que aparece en el enunciado actual $U_i$. $C_b$ es un tipo especial de centro, pues se relaciona con la entidad discutida en $U_i$ y conecta ambos enunciados.

Con el fin de modelar la centralidad de los textos, y con ello la coherencia textual, la teoría define cuatro tipos de transiciones de entidades entre enunciados, las cuales se muestran en la tabla 2.1.

Tabla 2.1: Tabla de transiciones en la Teoría del Centrado [13].

| | $C_b(U_i) = C_b(U_{i-1})$- | $C_b(U_i) \neq C_b(U_{i-1})$ |
|---|---|---|
| $C_b(U_i) = C_p(U_i)$ | *Continue* | *Smooth Shift* |
| $C_b(U_i) \neq C_p(U_i)$ | *Retain* | *Rough Shift* |

Estas transiciones se utilizan para evaluar qué tan coherente es un texto, y tienen un orden de preferencia. Una transición tipo *continue* se prefiere respecto a una tipo *retain* la cual se prefiere respecto a una tipo *smooth shift* la cual se prefiere respecto a una tipo *rough shift*. La idea fundamental es que discursos coherentes tienden a mantener un centro entre enunciados (*utterances*) mientras que textos poco coherentes tienden a hacer cambios bruscos de centros (entidades de discurso). La interpretación de las transiciones es la siguiente:

- *Continue*: Los centros de $U_i$ y $U_{i-1}$ hacen referencia a la misma entidad.
- *Retain*: Se mantiene la entidad de mayor importancia discutida en $U_{i-2}$, pero agregando información en $U_{i-1}$.
- *Smooth Shift*: Se realiza un cambio de centro en $U_i$ respecto a $U_{i-2}$, pero este cambio es suave, pues el centro de $U_{i-1}$ se mantiene.
- *Rough Shift*: Existe un cambio brusco en los centros del discurso.

Para entender la idea, asuma los siguientes enunciados de un texto:

1. Juan tenía un terrible dolor de cabeza.
   ($C_b =? \; C_f = Juan > DolordeCabeza \; ,transicion = ninguna$)
2. Cuando la reunión terminó
   ($C_b = ninguno \; C_f = Reunion \; ,transicion =$*Rough Shift*)

3. él corrió a la farmacia
   ($C_b = ninguno\ C_f = Juan$ ,$transicion =$*Rough Shift*)

Se puede observar que es difícil establecer la relación entre el primer y segundo enunciado, debido a que el foco cambia bruscamente en el segundo encunciado: primero se hablaba de *Juan* y su *Dolor de Cabeza* y luego se habla sobre una *reunión*. Se propone que al cambiar el orden de los enunciados, el texto hace más legible:

1. Juan tenía un terrible dolor de cabeza.
   ($C_b =?\ C_f = Juan > DolordeCabeza$ ,$transicion = ninguna$)

2. él corrió a la farmacia
   ($C_b = Juan\ C_f = Juan > Farmacia$ ,$transicion =$*Continue*)

3. cuando la reunión terminó
   ($C_b = ninguno\ C_f = Reunion$ ,$transicion =$*Rough Shift*)

Esta teoría ha sido utilizada en el contexto de evaluación automática de ensayos, determinando que existe una correlación entre la proporción de transiciones *Rough Shift* y el puntaje asignado por humanos. Se ha encontrado que mientras mayor sea esta proporción, los textos tienden a ser menos coherentes [33]. La principal ventaja, es que se considera el orden de los enunciados en un texto. Sin embargo, aplicar la teoría del centrado de forma automática depende fuertemente de la precisión con la que se capturan las entidades de discurso e identificar los conjuntos $C_f$. En la literatura se han utilizado textos etiquetados por humanos para la extracción de centros de discurso y enunciados, debido a que métodos automáticos tienen problemas al momento de identificar las coreferencias dentro del texto [33]. Esto es poco práctico pues en un sistema automático para evaluar ensayos se desearía que todo el proceso se realizara sin intervención humana.

### 2.3.2. Coherencia Textual basada en *Entity Grids*

El modelo de *Entity Grids* está basado en la teoría del centrado, y se utiliza para medir la coherencia textual de forma automática. Este modelo toma como supuesto

que existen patrones que son más probables que aparezcan en discursos coherentes. Para detectar estos patrones se realiza el siguiente procedimiento:

1. Se divide el texto en oraciones.

2. En cada oración se detectan *frases nominales* (*Noun Phrases*). Una frase nominal es una frase que contiene un nombre (por lo general un sustantivo, *Noun*) y puede tener modificadores que lo distingan (ej: *el perro de mi tía*, *El jefe de operaciones*, *Pedro*, etc.). Las frases nominales se conocen como *entidades* en este modelo.

3. Se detecta el *rol gramatical* de todas las entidades en las oraciones. Un rol gramatical hace referencia a la relación existente entre constituyentes en una cláusula. Por ejemplo en la siguiente cláusula *Pedro le entregó el libro a Susana*, existen tres entidades, *Pedro*, *el libro* y *Susana*. En este caso, la relación está dada por el verbo *entregó*, el sujeto de dicha acción es *Pedro*, el objeto de la acción es *el libro* y *Susana* tiene el rol de objeto indirecto, pues recibe el libro que es el objeto en la cláusula. En el modelo de entity grid se consideran los siguientes roles gramaticales *sujeto* (`s`), *objeto* (`o`) u *otro* (`x`).

4. Finalmente se obtiene la grilla de entidades, que será un arreglo bi-dimensional, donde las filas representan las oraciones del texto, las columnas representarán las entidades y en cada celda estará el rol gramatical de la entidad $j$ en la oración $i$.

Un ejemplo de texto y su grilla de entidades se muestran en las figuras 2.3 y 2.4.

1. [Former Chilean dictator Augusto Pinochet]$_{\mathbf{O}}$, was arrested in [London]$_{\mathbf{X}}$ on [14 October]$_{\mathbf{X}}$ 1998.
2. [Pinochet]$_{\mathbf{S}}$, 82, was recovering from [surgery]$_{\mathbf{X}}$.
3. [The arrest]$_{\mathbf{S}}$ was in [response]$_{\mathbf{X}}$ to [an extradition warrant]$_{\mathbf{X}}$ served by [a Spanish judge]$_{\mathbf{S}}$.
4. [Pinochet]$_{\mathbf{O}}$ was charged with murdering [thousands]$_{\mathbf{O}}$, including many [Spaniards]$_{\mathbf{O}}$.
5. [Pinochet]$_{\mathbf{S}}$ is awaiting [a hearing]$_{\mathbf{O}}$, [his fate]$_{\mathbf{X}}$ in [the balance]$_{\mathbf{X}}$.
6. [American scholars]$_{\mathbf{S}}$ applauded the [arrest]$_{\mathbf{O}}$.

Figura 2.3: Texto con anotaciones sintácticas para el cálculo de matriz de entidades [3].

| | Dictator | Augusto | Pinochet | London | October | Surgery | Arrest | Extradition | Warrant | Judge | Thousands | Spaniards | Hearing | Fate | Balance | Scholars | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | O | O | O | X | X | – | – | – | – | – | – | – | – | – | – | – | 1 |
| 2 | – | – | S | – | – | X | – | – | – | – | – | – | – | – | – | – | 2 |
| 3 | – | – | – | – | – | – | S | X | X | S | – | – | – | – | – | – | 3 |
| 4 | – | – | O | – | – | – | – | – | – | – | O | O | – | – | – | – | 4 |
| 5 | – | – | S | – | – | – | – | – | – | – | – | – | O | X | X | – | 5 |
| 6 | – | – | – | – | – | – | O | – | – | – | – | – | – | – | – | S | 6 |

Figura 2.4: Una grilla de entidades [3].

Las columnas de la matriz representan la presencia o ausencia de una entidad en una secuencia de oraciones $(S_1, ..., S_n)$. En particular, cada celda de la matriz representa el rol $r_{ij}$ de la entidad $e_j$ en la oración $S_i$. Estos roles definen si una entidad es un sujeto, objeto, ninguno o simplemente se encuentra ausente. Por ejemplo, en la figura 2, si se considera la entidad *arrest* es un sujeto, en la oración 3, es un objeto en la oración 6, pero se encuentra ausente en el resto de las oraciones. Con esta representación se pueden obtener patrones de transición entre las oraciones de un texto. Por ejemplo, la probabilidad de que ocurra una transición (o-) en la matriz anterior es 0.625, pues las transición (o-) ocurre 5 veces y el total posible de transiciones es 80 (es decir, la cantidad de transiciones de entre dos oraciones contiguas es 5 multiplicado por la cantidad de entidades que es 16).

Así, la coherencia de un texto $T(S_1, ..., S_n)$ con entidades $e_1, ..., e_m$, se puede ver como una distribución de probabilidad conjunta que describe cómo las entidades están distribuidas a través de las oraciones del texto $T$, esta probabilidad queda representada por los patrones de transición descritos previamente tal y como se muestra en la ecuación 2.3.

$$P_{coherence}(T) \approx \frac{1}{m \cdot n} \sum_{j=1}^{m} \sum_{i=1}^{n} log(P(r_{ij}|r_{(i-h)j}, \dots r_{(i-1)j})) \tag{2.3}$$

Para generar este modelo de distribución, se requiere un conjunto de textos que los humanos consideren coherentes [13]. Luego, se puede predecir $P_{coherence}(T)$ para un texto $T$, y este valor de probabilidad será mayor para textos que se consideren más coherentes que los que tengan un $P_{coherence}(T)$ menor. El método ha sido probado para evaluar coherencia de resúmenes generados automáticamente y ha obtenido una alta correlación con humanos comparado con otros métodos [26].

Puede observarse que un método de evaluación de coherencia textual basado en el modelo de entity grids considera patrones sintácticos, sin embargo, no considera la relación semántica existente entre oraciones. Como se ha discutido, la coherencia textual también está relacionada con el significado de un texto entre sus oraciones. Existen métodos para calcular coherencia textual considerando la semántica, y por lo general están basados en *análisis de cadenas léxicas*.

### 2.3.3. Análisis de Cadenas Léxicas

Las cadenas léxicas son secuencias de palabras relacionadas que abarcan una unidad textual. El supuesto en el análisis de cadenas léxicas es que los textos coherentes tienen una alta cantidad de palabras relacionadas semánticamente. Así, es posible medir la coherencia de un texto considerando el grado de similitud de cadenas léxicas entre oraciones adyacentes a lo largo de todo el texto (ecuación 2.4)

$$coherencia(T) = \frac{\sum_{i=1}^{n-1} sim(S_i, S_{i+1})}{n-1} \tag{2.4}$$

Donde $sim(S_i, S_{i+1})$ es una medida de similaridad entre las oraciones $S_i$ y $S_{i+1}$.

Según varios experimentos realizados en la literatura, este método tiene una alta

correlación con humanos en la tarea de evaluar qué tan coherente es un texto [26].

## 2.4. Sistemas Actuales para la Evaluación de Ensayos

El propósito de esta sección, es discutir los enfoques tradicionales para la evaluación de ensayos, si bien ha habido una diversidad de sistemas para esta tarea, la literatura reciente recoge los principales sistemas [45], los cuales son:

- SAGE [46]: Es un sistema evaluador de ensayos que combina medidas de coherencia basadas en un modelo de espacio vectorial, con medidas superficiales de los textos. El sistema utiliza medidas de distancia entre las partes de un texto (introducción, desarrollo, conclusiones) e intenta abordar la coherencia textual utilizando estadísticas sobre estas medidas y coeficientes de dispersión basados en análisis de clústers.

- PEG [38]: Inicialmente utilizaba características superficiales para aproximar la coherencia de los textos (conteo de palabras, y medidas basadas en este tipo de característicsa). Actualmente utiliza características lingüísticas y modelos no lineales para determinar la calificación de un ensayo. Los modelos utilizados y las características no han sido publicadas y se mantienen confidenciales.

- e-rater [6]: Utiliza técnicas de NLP y modelos de espacio vectorial para extraer características relacionadas al contenido de los textos. Los atributos que utiliza el sistema se pueden incluyen: Errores gramaticales, errores en el uso de palabras, errores en escritura (por ejemplo, errores ortográficos), presencia de elementos discursivos de un ensayo (por ejemplo introducción, desarrollo, conclusión), entre otros. El sistema utiliza modelos de regresión para ajustar la calificación de un ensayo.

- Intellimetric [12]: El sistema analiza diversos atributos semánticos y superficiales para estimar el *significado* de un texto. El sistema utiliza múltiples predictores basados en regresión líneal, modelos estadísticos y LSA, los que finalmente se combinan para determinar la nota final de un ensayo.

- CRASE [28]: Utiliza características superficiales para aproximar fluidez entre oraciones, contenido del ensayo y estilo. Para determinar la nota de un ensayo utiliza técnicas de aprendizaje automático y combina diferentes modelos de regresión.

- LightSide [30]: Este sistema utiliza diversos atributos basados en modelos de espacio vectorial y medidas superficiales de los textos. El sistema utiliza múltiples algoritmos para determinar la nota final de un texto (regresión lineal, modelos probabilísticos, etc.).

- AutoScore: Utiliza medidas superficiales en combinación con aproximaciones a la coherencia de un texto, basadas en similitudes con ensayos bien evaluados y mal evaluados. Los detalles internos del sistema nunca se publicaron, pero fue evaluado en [17].

- Intelligent Essay Assessor (IEA) [25]: El sistema utiliza medidas basadas en LSA en combinación con características superficiales. Las medidas que utiliza incluyen: similitud de ensayos en un espacio vectorial, largo de los vectores que representan los documentos en un espacio semántico (como aproximación a medidas del contenido de un texto), y medidas superficiales.

- Bookette [37]: Utiliza atributos basados en NLP (como por ejemplo medidas basadas en frecuencia de etiquetas léxicas en un texto). Además, utiliza medidas relacionadas a la gramática de un texto, y a la organización del mismo. Para obtener un modelo de evaluación de ensayos el sistema utiliza redes neuronales.

- Lexile [41]: El sistema utiliza utiliza medidas especiales basadas en características superficiales, las cuales intentan estimar la habilidad de escritura de un estudiante basada en factores relacionados a la complejidad semántica (la complejidad de las palabras utilizadas) y en características sintácticas relacionadas a cómo las palabras fueron utlizadas dentro del texto.

# Capítulo 3

## Método de Evaluación de Ensayos basado en Patrones de Discurso y Medidas de Coherencia

En esta tesis se desarrolló un método de evaluación de ensayos basado en patrones de discurso y medidas semánticas de coherencia. El método combina características del modelo de *entity grids* con el *modelo de espacio vectorial* y medidas de coherencia basadas en LSA. El enfoque propuesto permite abordar en forma directa la coherencia textual. Por otro lado, si se quisiera desarrollar un sistema tutorial inteligente para la enseñanza de la escritura, el método propuesto tiene la ventaja de que permitiría dar retroalimentación no sólo basada en medidas superficiales, sino que también en características estructurales del texto evaluado.

Nuestro método funciona como un predictor de la nota de un ensayo. Para lograr esto, los parámetros del método, que se estiman a partir un corpus de ensayos (textos evaluados por humanos), corresponden a características que podrían ser posibles predictores de "calidad" de un ensayo: *características superficiales* y *características de coherencia*. Las características superficiales están relacionadas a aspectos extrínsecos del texto (tales como número de palabras, sílabas, etc.). Por otro lado, las características de coherencia están relacionadas a aspectos intrínsecos del texto tales como la conexión semántica entre oraciones y la centralidad del texto.

La hipótesis principal de este trabajo es que un método que considere características sintácticas y semánticas para evaluar la coherencia textual será más efectivo para la tarea de evaluación automática de ensayos en comparación a métodos que utilicen medidas superficiales estadísticas. Uno de los problemas de los métodos convencionales es que no logran medir la coherencia de manera directa, por lo que las contribuciones de este trabajo son las siguientes:

- Abordar el problema de evaluación de coherencia en ensayos mediante el uso

características basadas en patrones de discurso y medidas semánticas de coherencia.

- Considerar métricas de "calidad de un ensayo" basadas en el modelo de *entity grids*.

Para esto, este capítulo se divide en tres secciones: la primera, describe el esquema general del método propuesto y las etapas que se llevan a cabo para evaluar un ensayo. Luego, se describen las características superficiales consideradas y cómo se computan. Finalmente, se describen las características de coherencia basadas en patrones de discurso y medidas semánticas.

## 3.1. Esquema de evaluación de ensayos propuesto

La evaluación automática de ensayos es el proceso de calificar de manera automática una composición textual, que se espera esté correlacionada con la de un evaluador experto humano.

Para esto, el método propuesto realiza el siguiente procedimiento, a partir de un conjunto de ensayos:

1. Evaluar ensayos, en base a alguna rúbrica dada. Luego determinar un puntaje final del ensayo, en base a las calificaciones dadas por humanos.

2. Pre-procesar los ensayos y extraer características superficiales y de coherencia.

3. Entrenar un modelo de regresión utilizando técnicas de *aprendizaje automático*, a partir de los ensayos codificados como un conjunto de características.

4. Finalmente, para evaluar un ensayo nuevo, se pre-procesa, se extraen características y luego se aplica el clasificador para asignar una nota.

## 3.2. Características Estadísticas

Este tipo de característica se considera un buen predictor de la calidad de un ensayo [31]. Por ejemplo, considerar el largo de un ensayo podría evaluar la calidad

de un texto, pues los buenos ensayos tienden a explicar de forma precisa las ideas. Otro ejemplo de medida estadística es la cantidad de palabras diferentes utilizadas dentro del texto, lo que podría ayudar en medir qué tan diverso es el texto.

Las características consideradas se dividieron en tres categorías *medidas de legibilidad*, *medidas de diversidad léxica*, y *medidas gramaticales* (ver tabla 3.1). La esencia de utilizar medidas de legibilidad es determinar qué tan legible es un texto, pues un evaluador humano considera la dificultad al leer un ensayo para determinar la nota que le asignará. Por otro lado, utilizar medidas de diversidad léxica tiene el propósito de determinar qué tan variado es el vocabulario que utiliza un estudiante al realizar una composición textual, lo cual es considerado por un evaluador humano al momento de asignar un puntaje. Así mismo, en tareas de composición textual es importante el uso de recursos gramaticales, y es por ello que se utilizan medidas tales como la frecuencia de las categorías léxicas.

Tabla 3.1: Características utilizadas por el método base (Fuente: Elaboración propia).

| | **Características Superficiales** | | **Características Gramaticales** |
|---|---|---|---|
| 1. | Cantidad de Caracteres | | **Conteo de Etiquetas Léxicas** |
| 2. | Cantidad de Palabras | 28. | Adverbios (RB) |
| 3. | Cantidad de Palabras Cortas | 29. | Conjunciones Coordinadas (CC) |
| 4. | Cantidad de Palabras Largas | 30. | Numerales (CD) |
| 5. | Largo más frecuente de palabras | 31. | Determinantes (DT) |
| 6. | Largo promedio de palabras | 32. | Preposiciones (IN) |
| 7. | Cantidad de Oraciones | 33. | Adjetivos (JJ) |
| 8. | Cantidad de Oraciones Largas | 34. | Adjetivos comparativos (JJR) |
| 9. | Cantidad de Oraciones Cortas | 35. | Adjetivos superlativos (JJS) |
| 10. | Largo más frecuente de oraciones | 36. | Auxiliares modales (MD) |
| 11. | Largo Promedio de Oraciones | 37. | Sustantivos comunes en plural (NNS) |
| 12. | Cantidad de palabras diferentes | 38. | Sustantivos propios en plural (NNPS) |
| 13. | Cantidad de *stopwords* | 39. | Sustantivos propios en singular (NNP) |
| | **Medidas de Legibilidad** | 40. | Pronombre personal (PRP) |
| 14. | Gunning Fog Index | 41. | Pronombre posesivo (PRP$) |
| 15. | Flesch Reading Ease | 42. | Adverbio comparativo (RBR) |
| 16. | Flesch Kincaid Grade Level | 43. | Adverbio superlativo (RBS) |
| 17. | Dale-Chall Readability Formula | 44. | Verbo *to* como preposición (TO) |
| 18. | Automated Readability Index | 45. | Verbos en su forma base (VB) |
| 19. | Simple Measure of Gobbledygook | 46. | Verbos en pasado (VBD) |
| 20. | LIX | 47. | Verbos en gerundio (VBG) |
| 21. | Word Variation Index | 48. | Verbos en pasado participio (VBN) |
| 22. | Nominal Ratio | 49. | Wh-determiner (WDT) |
| | **Diversidad Léxica** | 50. | Wh-pronoun (WP) |
| 23. | Type Token Ratio (TTR) | 51. | Possesive Wh-pronoun (WP$) |
| 24. | Guiraud's Index | 52. | Wh-adverb (WRB) |
| 25. | Yule's K | 53. | Sustantivos comunes en singular (NN) |
| 26. | The D Estimate | 54. | verb, sing. present, non-3d (VBP) |
| 27. | Hapax Legomena | 55. | verb, 3rd person sing. present (VBZ) |

| | **Mecánicas** |
|---|---|
| 56. | Errores Ortográficos |
| 57. | Capitalization Errors |

### 3.2.1. Medidas de Legibilidad

Estas medidas proporcionan una noción de qué tan legible es un texto. La razón de utilizar medidas de legibilidad es que un evaluador humano considera la dificultad de leer un ensayo al momento de evaluarlo. Las medidas que utiliza el método propuesto son las siguientes:

1. *Gunning Fog Index* (GFI): Estima la cantidad de años de educación formal que se requeriría para entender el texto en una primera lectura. El índice se calcula como [11]:

$$GFI = 0.4(average\ \ sentence\ \ length + HW) \tag{3.1}$$

   donde $HW$ es la cantidad de palabras que tienen más de dos sílabas.

2. *Flesch Reading Ease* (FRE): Es una medida que estima qué tan fácil es leer un texto, y estima el nivel educacional requerido para su comprensión. Esta medida es complementaria al *Gunning Fog Index* y se calcula como sigue:

$$FRE = 206.835 - (1.015ASL) - (8.46ASW) \tag{3.2}$$

   donde el puntaje varía entre 0 (difícil) y 100 (fácil de leer). $ASL$ es el largo promedio de las oraciones dentro del texto y $ASW$ es la cantidad promedio de sílabas por palabra [11]. La interpretación del puntaje entregado por la fórmula se muestra en la tabla 3.2:

Tabla 3.2: Descripción del índice *Flesch Reading Ease* [11].

| Puntaje | Descripción | Grado Estimado |
|---|---|---|
| 0 - 30 | Muy Difícil | Universidad, Post-Grado |
| 30 - 50 | Difícil | $13^{th}$ - $16^{th}$ |
| 50 - 60 | Poco Difícil | $10^{th}$ - $12^{th}$ |
| 60 - 70 | Promedio | $8^{th}$ - $9^{th}$ |
| 70 - 80 | Poco Fácil | $7^{th}$ |
| 80 - 90 | Fácil | $6^{th}$ |
| 90 - 100 | Muy Fácil | $5^{th}$ |

3. *Flesch Kincaid Grade Level* (FKGL): Esta medida es una mejora del índice *Flesch Reading Ease*. La razón de utilizar ambas medidas es que en algunos ensayos la primera podría ser un mejor predictor de la nota y viceversa. Esta medida se calcula como:

$$FKGL = (0.39ASL) + (11.8ASW) - 15.9 \tag{3.3}$$

donde $FKGL$ representa el nivel de educación formal necesario para entender el texto. $ASL$ es el largo promedio de las oraciones del texto, y $ASW$ es la cantidad de sílabas promedio por palabra.

4. *Dale-Chall Readability* (DCR): Esta medida estima la capacidad de una persona para responder preguntas de selección múltiple sobre un texto dado [11], cuánto del texto una persona logrará aprender y retener. El método considera una lista de 3000 palabras "fáciles" que una persona promedio debiese conocer (la cual se conoce como *lista de Dale* [11]). Para calcular este índice se deben seguir los siguientes pasos:

   i. Escoger una muestra de $N$ palabras del texto.

   ii. Calcular el largo promedio de las oraciones.

   iii. Calcular el porcentaje de palabras que no aparecen en la *lista de Dale*.

   iv. Calcular el índice como:

$$DCL = 0.1579PDW + 0.0496ASL + 3.6365 \tag{3.4}$$

donde *Score* es la nota que obtendría un lector que pudiese responder la mitad de preguntas que podría tener el texto, $PDW$ es el porcentaje de "palabras difíciles", y $ASL$ es el largo promedio de las oraciones.

5. *Automated Readibility Index* (ARI): Es una medida diseñada para estimar qué tan entendible es un texto. A diferencia de las medidas anteriores, considera el

número de caractéres en lugar de sílabas. El índice entrega una aproximación del nivel de educación requerido para entender un texto [11]:

$$ARI = 0.50(WPS) + 4.71(CPW) - 21.43 \tag{3.5}$$

donde $ARI$ es el nivel de educación a calcular, $WPS$ es la cantidad promedio de palabras por oración, y $CPW$ es la cantidad promedio de caracteres por palabra.

6. *Simple Measure of Gobbledygook* (SMOG): Es una variación de las medidas de legibilidad convencionales y similar a la medida *Gunning Fog*, siendo en algunos casos mejor predictor de la legibilidad de un texto. Para calcular esta medida, se cuenta la cantidad de palabras que tengan más de dos sílabas ($PSC$, *polysyllable count*) y calcula el siguiente puntaje [11]:

$$SMOG = 3 + \sqrt{PSC} \tag{3.6}$$

7. *Word Variation Index* (WVI): Es la proporción entre palabras diferentes en un texto, visto como una forma de *densidad de ideas*, y se calcula como [39]:

$$WVI = \frac{log(n(w))}{log\left(2 - \frac{log(n(uw))}{log(n(w))}\right)} \tag{3.7}$$

donde $n(w)$ es la cantidad de palabras en el texto, y $n(uw)$ es la cantidad de palabras diferentes en el texto.

*Nominal Ratio* ($NR$): Mide qué tan bien escrito está un texto en relación a recursos léxicos y gramaticales utilizados y se calcula como:

$$NR = \frac{n(noun) + n(prep) + n(part)}{n(pro) + n(adv) + n(v)} \tag{3.8}$$

donde $n(noun)$ es el número de sustantivos del texto, $n(prep)$ es el número de preposiciones, $n(part)$ es el número de verbos en participio, $n(pro)$ es el número de pronombres, $n(adv)$ es el número de adverbios y $n(v)$ es el número de

verbos. Un valor alto del NR indica un texto desarrollado y de estilo profesional, mientras que valores pequeños de esta medida indican textos más simples e informales.

### 3.2.2. Medidas de Diversidad Léxica

Estas medidas permiten determinar la diversidad del vocabulario utilizado en un texto: un ensayo que utilice un vocabulario variado es mejor que aquel que utilice un vocabulario pobre [31].

1. *Type Token Ratio* ($TTR$): Mide la proporción de palabras diferentes del texto y la cantidad total de palabras:

$$TTR = \frac{v}{N} \tag{3.9}$$

donde $v$ es la cantidad de tipos de palabras y $N$ es la cantidad total de palabras.

*Guiraud's Index* (GI): Es una medida de diversidad léxica, que a diferencia del $TTR$ se ve penalizada con el largo del ensayo. Esta medida se utiliza porque en conjuntos de ensayos que posean una elevada variabilidad de su longitud, podría ser un mejor predictor:

$$GI = \frac{v}{\sqrt{N}} \tag{3.10}$$

esta medida aumenta con el largo del ensayo hasta aproximadamente 100 palabras y luego decrece de forma estable con el largo del ensayo, lo que permite hacer comparar la diversidad léxica de dos ensayos indepnediente del largo [29].

2. *Yule's K*: Mide la diversidad léxica, y a diferencia de las medidas anteriores, considera un promedio ponderado del $TTR$ considerando frecuencia de ocurrencia de cada palabra:

$$K = 10^{-4}\frac{\sum r^2 V_r - N}{N^2} \quad r = 1, 2, \ldots \tag{3.11}$$

donde $V_r$ es el número de palabras que ocurren $r$ veces en un texto que contenga $N$ *palabras*. Esta medida se utiliza en conjunto con las demás, debido a que en algunos conjuntos de ensayos es un mejor predictor de la nota final, pero en otros casos no.

3. *D estimate* (D): Mide la diversidad léxica utilizando un modelo no lineal [29]. Para calcular esta medida se realiza el siguiente procedimiento:

   i. Tomar una muestra aleatoria de $N$ palabras del texto.

   ii. Calcular el $TTR$.

   iii. Encontrar el valor de $D$ que mejor se ajuste a la siguiente ecuación:

$$TTR = \frac{D}{N}\left[\sqrt{\left(1 + 2\frac{N}{D}\right)} - 1\right] \tag{3.12}$$

   finalmente el valor de $D$ encontrado es una estimación de la diversidad léxica del texto.

4. *Hapax Legomena*: Esta medida calcula el número de palabras que ocurren una sola vez en el texto, ya que es un buen discriminador entre un texto escrito por alguien que esté aprendiendo un lenguaje y un texto escrito por un nativo de dicho lenguaje [39].

### 3.2.3. Medidas Gramaticales

Estas medidas están relacionadas con a la frecuencia de aparición de palabras con una determinada categoría léxica (por ejemplo, verbos, sustantivos, adjetivos, etc). Cada palabra de un texto es etiquetada con su correspondiente categoría léxica a partir de la cual se calcula la frecuencia de aparición.

## 3.3. Características Basadas en Patrones de Discurso y Medidas de Coherencia

Este tipo de características permite estimar la coherencia de los ensayos y utilizar tal estimación como predictor de la calidad. El conjunto de características propuesto

se dividió en dos categorías: características basadas en modelo de *entity grids*, y características semánticas (ver tabla 3.3).

Tabla 3.3: Conjunto de características basados en patrones de discurso y medidas semánticas de coherencia (Fuente: Elaboración propia).

| | **Características Sintácticas** | | **Características Semánticas** |
|---|---|---|---|
| | **Patrones de Discurso** | 16. | Nota del ensayo |
| 1. | Transiciones {SS} | | más similar utilizando coseno. |
| 2. | Transiciones {SO} | 17. | Similitud Coseno con ensayos |
| 3. | Transiciones {SX} | | de nota máxima. |
| 4. | Transiciones {S-} | 18. | *Pattern Cosine* |
| 5. | Transiciones {OS} | 19. | *Weighted Sum of* |
| 5. | Transiciones {OO} | | *all cosine correlation values* |
| 6. | Transiciones {OX} | | **Medidas basadas en LSA** |
| 7. | Transiciones {O-} | 20-23. | Similitud Coseno entre oraciones |
| 8. | Transiciones {XS} | | (Prom., Min., Máx., Desv.) |
| 9. | Transiciones {XO} | | |
| 10. | Transiciones {XX} | | |
| 11. | Transiciones {X-} | | |
| 12. | Transiciones {-S} | | |
| 13. | Transiciones {-O} | | |
| 14. | Transiciones {-X} | | |
| 15. | Transiciones {--} | | |
| | **Similitud Estructural** | | |
| 24-27. | Similitud Coseno representación vectorial de los textos con los mejores ensayos (Prom., Min., Máx., Desv.) | | |
| 28. | Puntaje de los ensayos más similares en la representación | | |

### 3.3.1. Características basadas en *Entity Grids*

Este tipo de características está basado en el modelo de coherencia local mediante *entity grids*. Cada ensayo se representa como una *entity grid*, a partir de la cual se pueden extraer patrones locales de transiciones de entidades. Una *transición local* es una secuencia $\{s, o, x-\}^n$ que representa la ocurrencia de una entidad y su rol gramatical en oraciones adyacentes. Utilizando la *entity grid* se puede calcular la probabilidad de que ocurra una determinada transición dentro de un texto, lo que permite representar dicho texto como un conjunto fijo de secuencias de transición. Cada *entity grid* $j$ de un documento $d_i$ corresponde al vector de características $\Phi(x_{ij}) = \{p_1(x_{ij}), p_2(x_{ij}), \ldots, p_m(x_{ij})\}$, donde $m$ es la cantidad predefinida de transiciones, y $p_t(x_{ij})$ es la probabilidad de la transición $t$ en la matriz $x_{ij}$. En esta tesis se consideró una ventana de dos oraciones, es decir, $n = 2$, debido a que utilizar este valor ha obtenido buenos resultados [3]. Por lo tanto, los patrones de transición considerados son del tipo $\{ss\}$, $\{so\}$, etc. Un supuesto fundamental en el uso de estas características es que la distribución de entidades en ensayos *coherentes* exhiben ciertas regularidades que se reflejan en la topología de la *entity grid*, y por tanto, ensayos que presenten estas regularidades serán mejor evaluados que ensayos que no las presenten (por ejemplo, un texto coherente tendrá una alta concentración de transiciones $\{--\}$ [26]).

Por otro lado, utilizando el vector de características con las probabilidades de transición, se definió una medida de *similitud estructural*, dados dos supuestos:

1. Si un evaluador considera características de coherencia local al evaluar un determinado ensayo, entonces ensayos similares en estructura, es decir, que posean una distribución de entidades similar, se calificarán con notas similares.

2. Ensayos mejor evaluados deberán una mayor similitud con ensayos que se evaluaron con buena nota, que con ensayos mal evaluados.

Se consideran como predictores de la nota de un ensayo las siguientes medidas: el promedio, la desviación estándar, la máxima y la mínima similitud. Por otro lado, se dividen los ensayos en $k$ conjuntos, donde $k$ es el número de notas diferentes que pueden tener los ensayos. Luego, los vectores con las probabilidades de las transiciones,

se calcula la similitud promedio entre un ensayo nuevo y los ensayos de cada grupo. Se considera como predictor la nota de los ensayos con la mayor similitud.

### 3.3.2. Características Semánticas

Este tipo de características se divide en dos grupos: *medidas basadas en contenidos* [33] y medidas de coherencia basadas en *Análisis Semántico Latente* [23]. Las *medidas de contenido* están basadas en el *modelo de espacio vectorial* e intentan capturar el contenido de un ensayo en base a la frecuencia de aparición de palabras relevantes del tópico de un ensayo. Por otro lado, las medidas de coherencia basadas en LSA, estiman la coherencia de un ensayo en considerando las relaciones semánticas existentes entre sus oraciones.

#### Medidas Basadas en Contenidos

Este tipo de medidas compara el contenido contenido léxico de un ensayo con el contenido de un grupo de ensayos evaluados para una misma tarea [2][35]. Primero, se define un vocabulario que considera $m$ palabras y cada ensayo se representará como un vector $(F_1, F_2, \ldots F_m)$ cuyas componentes corresponden a la frecuencia de aparición dichas palabras. Luego, se divide el corpus de ensayos en $k$ grupos, donde cada grupo tiene la misma nota, y se obtiene una representación vectorial de cada ensayo. Finalmente, se estima una ponderación de cada palabra en un grupo de ensayos dado, para determinar qué tan relevante es la palabra $i$ en el grupo de ensayos con nota $s$. Cada ponderación se calcula como:

$$W_{is} = \frac{F_{is}}{MaxF_s} log\left(\frac{N}{N_i}\right) \tag{3.13}$$

donde $F_{is}$ es la frecuencia de la palabra $i$ en el grupo de ensayos con puntaje $s$, $MaxF_s$ es la frecuencia de la palabra que más se repite en los ensayos con puntaje $s$, $N$ es la cantidad total de ensayos pre-evaluados a utilizar y $N_i$ es la cantidad de ensayos que tienen la palabra $i$ considerando todos los ensayos. El resultado de este cálculo entrega $k$ vectores de $m$ componentes, donde cada vector es de la forma $W^k = \{W_1, W_2, \ldots, W_m\}$

Luego, para un ensayo nuevo, se obtiene su representación vectorial y las ponderaciones de cada palabra en dicho ensayo:

$$W_i = \frac{F_i}{MaxF} log\left(\frac{N}{N_i}\right) \tag{3.14}$$

donde $F_i$ es la frecuencia de la palabra $i$ del ensayo y $MaxF$ es la frecuencia de la palabra que más se repite en el ensayo. Luego, se obtiene un vector $W$ donde cada componente son las ponderaciones de las palabras en el ensayo.

Utilizando los vectores de ponderaciones, se pueden computar similitudes coseno entre el vector del ensayo nuevo y los vectores de cada grupo de ensayos con puntaje $s$. Luego se definen tres características:

1. *Puntaje basado en similitud coseno*: Esta característica es la nota que tendría el ensayo nuevo considerando la máxima similitud coseno entre el vector de ponderaciones del ensayo y los vectores de ponderaciones de los grupos de ensayos.

2. *Similitud con ensayos de máximo puntaje*: Esta característica indica qué tan similar es el vocabulario del ensayo nuevo con el vocabulario de ensayos con máximo puntaje. Este se calcula como el coseno entre el vector de ponderaciones del ensayo nuevo y el vector de ponderaciones del grupo de ensayos con máxima nota.

3. *Pattern Cosine*: Mide la nota de un ensayo en base a su similitud con los grupos de ensayos. La lógica tras esta característica, es que si la nota del ensayo está positivamente correlacionada con la similitud con los mejores ensayos, también estará correlacionada con los otros grupos de ensayos. Esta medida se calcula como:

$$Pat.\ Cos. = \sum_{i}^{k} S_i R_i \tag{3.15}$$

donde $S_i$ son las calificaciones posibles para un ensayo y $R_i$ la similitud del ensayo nuevo con los grupos de ensayos que tengan nota $S_i$.

### Medidas de Coherencia Basadas en LSA

Las medidas basadas en Análisis Semántico Latente determinan la coherencia global del texto utilizando una representación vectorial de las palabras [23]. A diferencia de un *modelo de espacio vectorial*, LSA no se ve afectado por la variabilidad de las palabras y logra detectar relaciones latentes entre ellas. Para estimar la coherencia de un texto, se utiliza la representación de las palabras en LSA con la cual se calcula la similitud entre oraciones:

$$sim(S_i, S_j) = \frac{S_i \cdot S_j}{\|S_i\|\|S_j\|} \tag{3.16}$$

donde $S_i$ y $S_j$ es una representación vectorial de las oraciones y se calculan como la suma de los vectores de las palabras que las componen [9]. Luego, el predictor del puntaje de un ensayo considera las siguientes características: máxima similitud entre oraciones adyacentes, la mínima similitud entre oraciones adyacentes, promedio de las similitudes entre todas las oraciones (esto es una forma de estimar la *coherencia global* del texto) y la desviación estándar de estas similitudes.

# Capítulo 4

# Experimentos y Resultados

En este capítulo se describen los experimentos realizados con el método de evaluación de ensayos presentado anteriormente. Además, se realizan pruebas que comparan el método propuesto con otros enfoques de evaluación de ensayos del estado del arte.

La hipótesis planteada en este trabajo es que un método de evaluación que considera características de coherencia basadas en teoría de discurso es más efectivo para la tarea de evaluación automática de ensayos que un método que utilice sólo medidas superficiales. Para probar la hipótesis, se comparó el método de evaluación de ensayos propuesto basado en Patrones de discurso y medidas semánticas de Coherencia (EES + DS) con respecto a un método basado sólo en características superficiales (EES), utilizando distintas métricas aceptadas en la literatura.

El capítulo se divide en tres secciones: primero se describe el conjunto de datos utilizado para realizar los experimentos, luego se compara el método propuesto EES + DS con EES. Finalmente, se realiza una comparación del método EES + DS con otros sistemas del estado del arte.

## 4.1. Conjunto de Datos Utilizado

Los datos utilizados en los experimentos son los de la competencia *Automated Essay Scoring* [17]. La competencia consistió en crear un sistema evaluador de ensayos a partir de ensayos previamente evaluados y luego medir su rendimiento utilizando la métrica *Quadratic Weighted Kappa* [7][16][23] (ver tabla 4.3). Esta métrica se utiliza para medir el grado de acuerdo que existe entre dos evaluadores, siendo su valor máximo igual a 1.0 cuando existe un acuerdo total entre ambos evaluadores y puede llegar a tomar valores negativos si los evaluadores están totalmente en desacuerdo. Se utiliza para medir el grado de acuerdo entre la evaluación realizada por humanos y la evaluación automática.

Los ensayos de esta competencia se organizan en ocho grupos, cada uno sobre una temática diferente: *descriptivos*, *narrativos*, *argumentativos* y *respuestas basadas en un texto fuente*. Los ensayos escogidos tienen una longitud promedio entre 150 a 500 palabras y corresponden a alumnos de educación secundaria (octavo a décimo grado). Todos los ensayos fueron evaluados manualmente por humanos. Un caso especial es el conjunto de ensayos número 2, pues este conjunto fue evaluado de dos formas diferentes con una escala de notas diferente.

Tabla 4.1: Descripción de Ensayos de Prueba (Fuente: Elaboración propia).

| ID | Grado | Cantidad | Largo Promedio | Rango de Notas |
|---|---|---|---|---|
| 1 | 8 | 1785 | 350 | 2-12 |
| 2 | 10 | 1800 | 350 | 1-6, 1-4 |
| 3 | 10 | 1726 | 150 | 0-3 |
| 4 | 10 | 1772 | 150 | 0-3 |
| 5 | 8 | 1805 | 150 | 0-4 |
| 6 | 10 | 1800 | 150 | 0-4 |
| 7 | 10 | 1730 | 250 | 0-24 |
| 8 | 10 | 918 | 650 | 10-60 |

## 4.2. Comparación EES + DP con EES

En esta serie de experimentos se comparó el método propuesto basado en patrones de discurso y medidas semánticas de coherencia, con un método que sólo considere características superficiales. Para realizar este experimento se deben llevar a cabo tres tareas: extracción de características, selección de características y generación del modelo predictivo.

### 4.2.1. Extracción de Características

Cada ensayo fue procesado con el fin de computar características superficiales, los índices de legibilidad, los índices de diversidad léxica y características gramaticales tales como conteo de etiquetas léxicas. Primero, se segmentó el texto en oraciones y las oraciones se *tokenizaron*, es decir, se separaron por palabras y se realizó un etiquetado léxico de las palabras. Para realizar este procesamiento se utilizó la biblioteca *python*

*nltk* [4] y para detectar errores ortográficos se utilizó la biblioteca *pyEnchant* [1].

Para extraer características de contenido del texto, se realizó una representación de los ensayos en un modelo de espacio vectorial. Además, se eliminaron las *stopwords* y signos de puntuación, exceptuando apóstrofes. Luego, se realizaron correcciones ortográficas utilizando *pyEnchant* y se dejaron en minúsculas todas las palabras de los ensayos. Posteriormente, se realizó *stemming* (para determinar la raíz de cada palabra) mediante la implementación del método Porter disponible en la biblioteca *nltk*, esto se realizó con el fin de reducir la variabilidad de las palabras en los conjuntos de ensayos. Luego representó cada documento en un modelo de espacio vectorial utilizando el paquete en R cuyo nombre es *lsa* (*lsa* es el nombre de un paquete que cuenta con facilidades para generar una representación vectorial de los textos en base a sus palabras). Luego, se extrajeron las siguientes características: nota del ensayo más similar, similitud coseno con ensayos de nota máxima, y pattern cosine.

Para extraer medidas de coherencia basadas en LSA, se utilizaron los paquetes *lsa* y *LSAfun* disponibles en R [15], que cuentan con funciones para calcular el coseno entre dos textos dado un espacio semántico. Para ello, se segmentó cada ensayo en oraciones y se calculó el coseno entre oraciones adyacentes, donde las oraciones se representaron en un vector que se calcula como la suma de los vectores que representan cada palabra. El espacio semántico utilizado fue obtenido aplicando Análisis Semántico Latente al *corpus TASA*, y considera la representación de más de 100.000 palabras en inglés[2].

Finalmente, para obtener las características basadas en patrones de discurso se representó cada ensayo como una *entity grid*, utilizando la implementación del framework *cohere* [40], la cual utiliza el parser de dependencias de Stanford para extraer los roles gramaticales de las entidades en el texto[3]. Luego, se computaron las probabilidades de transiciones de entidades de cada ensayo. Sólo se consideraron transiciones de largo dos (por ejemplo `ss`, `so`, etc.). Posteriormente, se generó una representación vectorial de los ensayos, donde cada componente del vector corresponde a una probabilidad de transición. Finalmente se computó la similitud coseno con los mejores

[1] http://pythonhosted.org/pyenchant/
[2] http://www.lingexp.uni-tuebingen.de/z2/LSAspaces/
[3] http://nlp.stanford.edu/software/lex-parser.shtml

ensayos, el promedio, máximo, mínimo y la desviación estándar.

### 4.2.2. Selección de Características

Con el fin de capturar las características más relevantes y no redundantes se realizó una selección de características utilizando *stepwise regression* [43], la cual permite escoger predictores de forma automática en base a algún criterio. En este caso, se utilizó el criterio de selección *Akaike Information Criterion* [42], el cual escoge las características en base a su relevancia, es decir, variables que logran predecir con mejor precisión la nota de los ensayos. En la figura 4.1 se muestra el resultado del proceso de selección de características. Un resumen de las características utilizadas en el modelo final y su importancia se muestran en la tabla 4.2.

Tabla 4.2: Resumen de características del modelo final y su relevancia. (Fuente: Elaboración Propia)

| Característica | Relevancia |
|---|---|
| Número de Caracteres | 16.5 |
| Número de Palabras Diferentes | 14.9 |
| Número de Palabras | 14.3 |
| Número de Palabras Largas | 13.8 |
| Número de Stopwords | 12.8 |
| Hapax Legomena | 12.5 |
| Número de etiquetas NN | 12.4 |
| WVI | 12.3 |
| GI | 11.9 |
| Número de etiquetas IN | 11.5 |
| Advanced Guiraud | 11.3 |
| Coseno con Mejores Ensayos | 10.5 |
| Número de Palabras Cortas | 10.1 |
| Número de Oraciones | 9.9 |
| Patrones de Transición -- | 9.8 |
| Número de Etiquetas DT | 9.7 |

| | |
|---|---|
| Número de Etiquetas JJ | 8.3 |
| Número de Oraciones Largas | 7.8 |
| Número Etiquetas NNS | 7.7 |
| D Estimate | 7.5 |
| Maxima Similitud Estructural | 7.3 |
| TTR | 6.8 |
| DCR | 6.6 |
| Patrones de Transición X- | 6.5 |
| ARI | 6.5 |
| Patrones de Transición XO | 6.4 |
| Patrones de Transición XX | 6.4 |
| Patrones de Transición O- | 6.3 |
| Largo Promedio de Palabras | 6.3 |
| Nota del Ensayo más Similar basado en contenido | 6.1 |
| Número de Etiquetas RB | 5.8 |
| Yule K | 5.5 |
| Número de Etiquetas VB | 5.4 |
| Número de Etiquetas VBP | 5.3 |
| Número de Etiquetas CC | 5.2 |
| Desviación Estándar Similitud Estructural | 5.2 |
| Patrones de Transición S- | 5.1 |
| Patrones de Transición XS | 5.1 |
| Largo Promedio de Oraciones | 5.1 |
| FRE | 5.1 |
| Número de Etiquetas TO | 5.1 |
| Desviación Estándar Similitud LSA | 5.0 |
| Número de Etiquetas PRP | 4.9 |
| FKE | 4.9 |
| Número de Etiquetas VBD | 4.8 |
| Pattern Cosine | 4.8 |
| Coherencia Global LSA | 4.8 |

| | |
|---|---|
| GF | 4.7 |
| Patrones de Transición XS | 4.7 |
| Nominal Ratio | 4.6 |
| Patrones de Transición XX | 4.5 |
| Número de Etiquetas VBN | 4.5 |
| Número de Etiquetas VBG | 4.3 |
| Patrones de Transición SS | 4.2 |

Las variables que son los mejores predictores de acuerdo al criterio de selección están marcadas con verde en el gráfico. Las variables que podrían eventualmente ser buenos predictores están en color amarillo y las variables que no aportan información relevante están marcadas en rojo. Se observa que medidas de coherencia basadas en LSA, y los patrones de discurso se encuentran entre los mejores predictores. Esto, es un indicio de que la hipótesis planteada es correcta. Es más, los patrones de discurso basados en *entity grids* se prefieren a la mayoría de las medidas de legibilidad basadas en características superficiales. Sin embargo, como se explicará más adelante, las medidas basadas en el número de caracteres son las que más relevancia tienen en los modelos finales tanto en EES como en EES + DP, esto da un indicio de un sesgo en este tipo de características, pues las calificaciones en el conjunto de datos escogido están fuertemente correlacionadas con el largo del ensayo. Sin embargo, de los valores de importancia de las características escogidas, se puede observar que los patrones de transición son relevantes en la predicción de la calificación del modelo final.

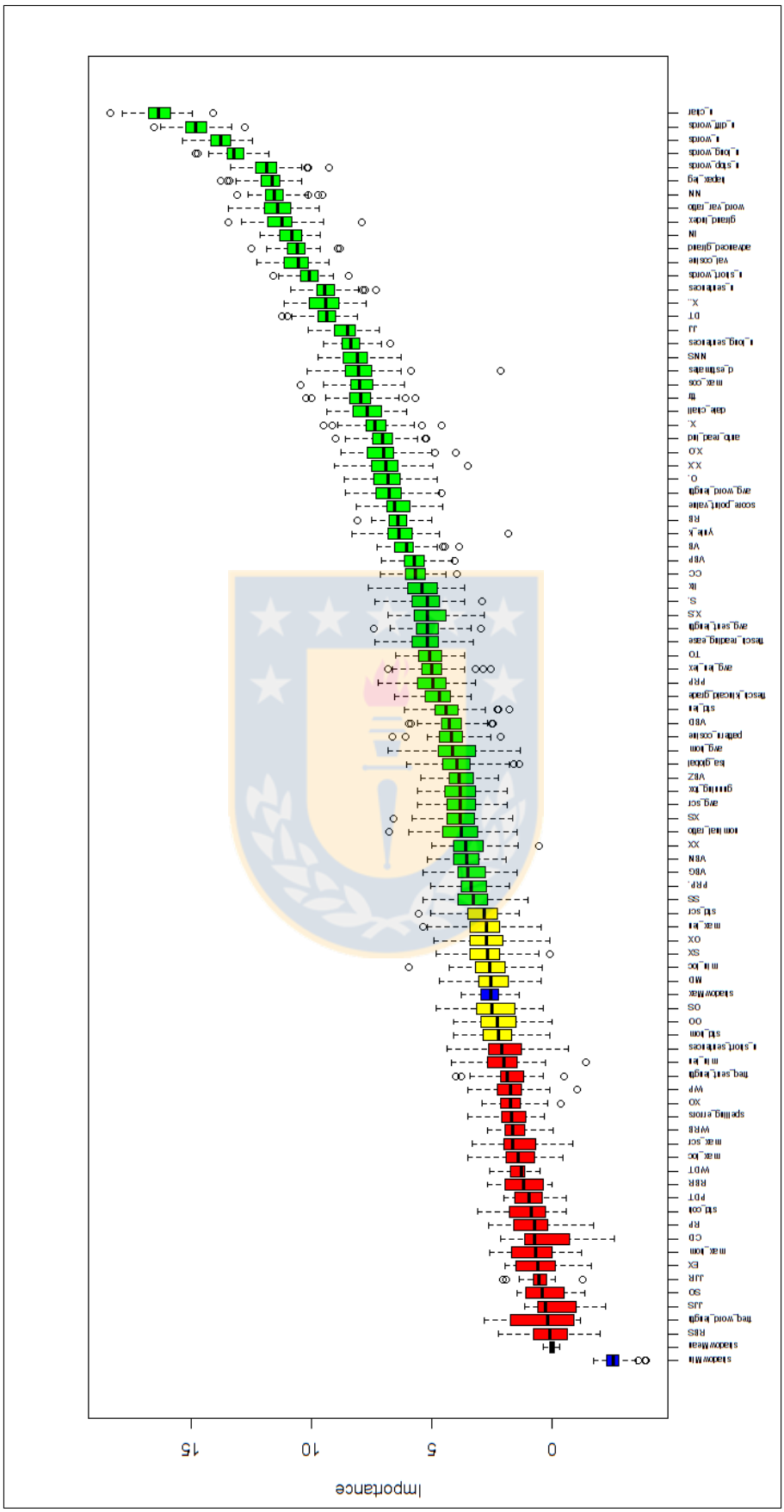


Figura 4.1: Importancia de las variables predictoras de la calificación de un ensayo (Fuente: Elaboración propia).

### 4.2.3. Generación del Modelo Predictivo

Con el fin de crear el modelo predictivo, se entrenó y utilizó el método de regresión *Random Forest* [5], pues es un método robusto frente a sobreajuste, ruido en los datos, y ha sido utilizado en la literatura [44]. Para aplicar este método se divide el conjunto de ensayos en dos, un conjunto de entrenamiento y un conjunto de pruebas. El conjunto de entrenamiento se utiliza para generar el modelo predictivo y el conjunto de pruebas se utiliza para medir el rendimiento del modelo generado. Para realizar una evaluación representativa, se implementó una validación cruzada tipo *k-fold*. Se decidió utilizar un valor $k = 10$ para la cantidad de particiones, obteniéndose así un 90 % de los ensayos para entrenamiento y un 10 % para prueba. Esta clase de validación cruzada forma subconjuntos entrenamiento-prueba y los prueba uno por uno, de forma que cada parte del conjunto de ensayos tiene oportunidad tanto de ser usada para entrenamiento como para probar. El resultado final de la validación cruzada se tomó como el promedio de los sub-resultados de cada *fold*. Se utilizó la implementación de *random forests* disponible en R y se utilizó 100 árboles por bosque (el resto de los parámetros de dejaron por defecto). Finalmente, para determinar si los resultados obtenidos son estadísticamente significativos, se usó el test de *Wilcoxon* [44]. En este test, un valor de $p < 0.05$ indica que hay una diferencia estadística significativa en los datos y por lo tanto la hipótesis planteada se acepta. Por otro lado, si el valor de $p > 0.05$, significa que la diferencia en los datos pudo ocurrir por casualidad.

### 4.2.4. Resultados Experimentales

Para comparar la efectividad de EES + DP y EES se consideraron cuatro métricas:

- *Quadratic Weighted Kappa* [16][23].
- *Correlación de Spearman*: Mide la correlación entre los puntajes asignados por humanos y los entregados por el método [22].
- *Exact Agreement*: mide la proporción de ensayos que fueron evaluados con la misma nota que un evaluador humano.

- *Adjacent Agreement*: mide la proporción de ensayos que fueron evaluados con a lo más una nota de diferencia con respecto al evaluador humano.

Los resultados obtenidos se muestran en las tablas 4.3 4.4 4.5 y 4.6 respectivamente.

Tabla 4.3: *Quadratic Weighted Kappa* para los conjuntos de ensayos (Fuente: Elaboración propia).

| | 1 | 2a | 2b | 3 | 4 | 5 | 6 | 7 | 8 | Promedio |
|---|---|---|---|---|---|---|---|---|---|---|
| EES + DP | 0.83 | $0.70^{x}$ | $0.65^{x}$ | 0.68 | $0.77^{*}$ | $0.81^{*}$ | $0.80^{*}$ | 0.80 | $0.74^{*}$ | $0.76^{*}$ |
| EES | 0.84 | 0.68 | 0.63 | 0.66 | 0.70 | 0.78 | 0.69 | 0.78 | 0.70 | 0.71 |

$^{*}$: $p < 0.05$, por lo tanto el método propuesto es significativamente mejor
$^{x}$: $p \approx 0.2$, es probable que lo propuesto es mejor.

Tabla 4.4: *Correlación de Spearman* para los conjuntos de ensayos (Fuente: Elaboración propia).

| | 1 | 2a | 2b | 3 | 4 | 5 | 6 | 7 | 8 | Promedio |
|---|---|---|---|---|---|---|---|---|---|---|
| EES + DP | 0.83 | $0.73^{x}$ | $0.69^{x}$ | 0.73 | $0.84^{*}$ | $0.86^{x}$ | $0.82^{*}$ | 0.80 | $0.72^{*}$ | $0.78^{*}$ |
| EES | 0.83 | 0.72 | 0.67 | 0.72 | 0.78 | 0.84 | 0.75 | 0.81 | 0.71 | 0.76 |

$^{*}$: $p < 0.05$, por lo tanto el método propuesto es significativamente mejor
$^{x}$: $p \approx 0.2$, es probable que lo propuesto es mejor.

Tabla 4.5: *Exact Agreement* para los conjuntos de ensayos (Fuente: Elaboración propia).

| | 1 | 2a | 2b | 3 | 4 | 5 | 6 | 7 | 8 | Promedio |
|---|---|---|---|---|---|---|---|---|---|---|
| EES + DP | 0.53 | 0.69 | 0.68 | 0.68 | $0.68^{*}$ | $0.69^{*}$ | $0.67^{x}$ | 0.14 | 0.11 | $0.54^{*}$ |
| EES | 0.51 | 0.69 | 0.68 | 0.66 | 0.63 | 0.66 | 0.60 | 0.15 | 0.10 | 0.52 |

$^{*}$: $p < 0.05$, por lo tanto el método propuesto es significativamente mejor

Tabla 4.6: *Adjacent Agreement* para los conjuntos de ensayos (Fuente: Elaboración propia).

| | 1 | 2a | 2b | 3 | 4 | 5 | 6 | 7 | 8 | Promedio |
|---|---|---|---|---|---|---|---|---|---|---|
| EES + DP | 0.94 | 1.0 | 1.0 | 0.98 | 0.99 | 0.99 | $1.0^{*}$ | 0.44 | 0.32 | 0.84 |
| EES | 0.94 | 0.99 | 0.99 | 0.98 | 0.98 | 0.99 | 0.97 | 0.44 | 0.30 | 0.84 |

$^{*}$: $p < 0.05$, por lo tanto el método propuesto es significativamente mejor

Estos experimentos muestran que el método EES + DP obtiene mejores resultados en la métrica *quadratic weighted kappa* que EES, con una diferencia significativa en el

conjunto de ensayos 4, 5, 6 y 8. Esto significa que el grado de acuerdo existente entre la nota asignada por humanos y la asignada de forma automática es mayor y por lo tanto las notas entregadas por EES + DP son más aproximadas a las que asignaría un humano. Esto sugiere que agregar características directas de coherencia textual tales como patrones de discurso, logran una mejora en la evaluación de ensayos.

Las mejoras observadas se dan en ensayos que son de tipo argumentativo y dependientes de un texto fuente, es decir, que el ensayo tiene el objetivo de responder preguntas basadas en información que el estudiante leyó previamente. Esto quiere decir que los ensayos deben tener una cierta estructura que explique lo requerido relacionado al texto y además las respuestas deben estar fundamentadas en el texto fuente. El método propuesto, a diferencia de un enfoque tradicional basado en características superficiales, considera características relacionadas a la centralidad del texto, las cuales son importantes en este tipo de ensayo.

Por otro lado, se observa que en los conjuntos de ensayos 1, 2a y 2b, la diferencia no es significativa, debido a que estos ensayos son de texto libre, y las evaluaciones de los expertos humanos están sesgadas al largo del ensayo [36]. Sin embargo, el valor de $p$ da un indicio de que las características agregadas podrían mejorar el rendimiento del método.

Para las métricas de *exact agreement*, se puede observar una mejora cuando se utilizan medidas de discurso (ver tabla 4.5). Los bajos resultados en los conjuntos de ensayos 7 y 8 se deben a la variabilidad de los puntajes finales que pueden tener los ensayos, es poco probable dar la nota exacta si los puntajes tienen mucha variabilidad (por ejemplo, los del conjunto 8 varían de 10 - 60). Para *adjacent agreement* no se observa una mejora significativa y los dos métodos pueden obtener buenos resultados.

Por lo tanto, agregar características basadas en patrones de discurso mejora la efectividad en cuanto a correlación con humanos y exactitud del puntaje asignado por el método computacional.

## 4.3. Comparación con Sistemas Actuales de Evaluación de Ensayos

En esta prueba se comparó el método propuesto con otros sistemas existentes, utilizando los resultados de la competencia de *Automated Essay Scoring* [17] (ver

tabla 4.7).

Tabla 4.7: Comparación de *Quadratic Weighted Kappas* con otros sistemas en el Estado del Arte [17].

| Sistema | 1 | 2a | 2b | 3 | 4 | 5 | 6 | 7 | 8 | Promedio |
|---|---|---|---|---|---|---|---|---|---|---|
| **EES + DP** | 0.83 | 0.70 | 0.65 | 0.68 | 0.77 | 0.81 | 0.80 | 0.80 | 0.74 | **0.76** |
| SAGE | 0.93 | 0.79 | 0.67 | 0.83 | 0.81 | 0.87 | 0.78 | 0.88 | 0.81 | 0.82 |
| PEG | 0.82 | 0.72 | 0.70 | 0.75 | 0.82 | 0.83 | 0.81 | 0.84 | 0.73 | 0.78 |
| e-rater | 0.82 | 0.74 | 0.69 | 0.72 | 0.80 | 0.81 | 0.75 | 0.81 | 0.70 | 0.76 |
| IntelliMetric | 0.78 | 0.70 | 0.68 | 0.73 | 0.79 | 0.83 | 0.76 | 0.81 | 0.68 | 0.75 |
| CRASE | 0.76 | 0.72 | 0.69 | 0.73 | 0.76 | 0.78 | 0.78 | 0.80 | 0.68 | 0.74* |
| LightSIDE | 0.79 | 0.70 | 0.63 | 0.74 | 0.81 | 0.81 | 0.75 | 0.77 | 0.65 | 0.74* |
| AutoScore | 0.78 | 0.68 | 0.66 | 0.72 | 0.75 | 0.82 | 0.76 | 0.67 | 0.69 | 0.73* |
| IEA | 0.79 | 0.70 | 0.65 | 0.68 | 0.77 | 0.81 | 0.80 | 0.78 | 0.72 | 0.73* |
| Bookete | 0.70 | 0.68 | 0.63 | 0.69 | 0.76 | 0.80 | 0.64 | 0.74 | 0.60 | 0.69* |
| Lexile | 0.66 | 0.62 | 0.55 | 0.65 | 0.67 | 0.64 | 0.65 | 0.58 | 0.63 | 0.63* |

*: $p < 0.05$, por lo tanto el método propuesto es significativamente mejor

Se observa que el rendimiento del método propuesto es significativamente mejor que sistemas principalmente basados en enfoques estadísticos como lo son (Lexile, Bookete, LightSide, AutoScore). Estos sistemas utilizan diferentes características superficiales para aproximar la coherencia textual y el estilo de escritura de los estudiantes, y por lo general combinan características superficiales como medidas de legibilidad para predecir la nota de un ensayo. Es por ello que obtienen resultados similares a los que obtuvo EES.

A diferencia de los métodos anteriores, nuestro enfoque utiliza estimaciones de la coherencia textual mediante el modelo de *entity grids* y representaciones de las palabras utilizando LSA, lo que se traduce en una mejora en términos de correlación con humanos. Esta mejora puede deberse a que los humanos al evaluar textos consideran como están enlazadas las oraciones y como se mantiene la centralidad dentro del ensayo. Otros sistemas como Intellimetric, PEG y e-rater utilizan una combinación de medidas estadísticas con otros atributos lingüísticos y combinan diferentes modelos predictivos para computar la nota de un ensayo (los atributos lingüísticos y modelos son confidenciales). Si el método propuesto utilizara una combinación de modelos predictivos, se podría mejorar la efectividad (por ejemplo combinar distintos modelos de regresión), aunque esto traería como consecuencia dos problemas: incremento de

tiempos de entrenamiento, y reducción de la interpretabilidad de los resultados.

El método EES + DP obtiene peores resultados que SAGE [46]. La razón de ello, es que SAGE utiliza múltiples medidas de coherencia textual en combinación con medidas basadas en bases de conocimiento, utilizando una ontología e inferencia lógica para determinar si ciertas cláusulas son correctas o incorrectas dentro del ensayo. También utiliza otras medidas de coherencia basadas en un *modelo de espacio vectorial*, las cuales intentan determinar la centralidad de un texto mediante análisis de clústers. Sin embargo, las medidas de SAGE no consideran directamente la centralidad de los textos, por lo que no podría decirse que evalúa mejor la coherencia textual, ya que las métricas están basadas puramente en combinación de palabras entre oraciones; el orden de las oraciones no es relevante pues cada texto se representa en un espacio multidimensional.

De los resultados, se puede concluir que el método propuesto es competitivo con sistemas aceptados en el estado del arte, superando a gran parte de dichos sistemas. Por otro lado, las estimaciones de la coherencia textual utilizadas mejoran el acuerdo entre el humano y el método computacional en la evaluación.

# Conclusiones

En este trabajo se propuso un método de evaluación automática de ensayos que combina modelos semánticos y sintácticos, utilizando medidas lingüísticas basadas en características superficiales de los textos, en combinación con medidas utilizadas para análisis de discurso. Este método es capaz de evaluar un ensayo a partir de su estructura sintáctica y léxica, y también considerando características basadas en teoría de discurso, las cuales permiten al sistema distinguir entre dos documentos con las mismas oraciones pero en diferente orden.

Diversos experimentos mostraron que el método propuesto obtiene una mejora significativa respecto a un método convencional que utilice sólo medidas estadísticas superficiales, y esto se probó empíricamente utilizando las métricas de *Quadratic Weighted Kappa* y *correlación de Spearman*. Por otro lado, se comparó el método propuesto con los sistemas del actual estado del arte y se demostró que es competitivo con tales sistemas, superando a la gran mayoría de métodos tradicionales utilizados para la evaluación automática de ensayos.

Como trabajo futuro se propone explorar medidas basadas en otros modelos de representación de textos tales como *word2vec* [32], *paragraph2vec* [27], pues son representaciones alternativas de los textos que podrían servir para medir propiedades complementarias de la coherencia textual. Esto podría mejorar la evaluación automática de ensayos en términos de correlación con humanos. También se propone investigar el impacto que tendría aplicar otros métodos para medir coherencia textual basados en *entity grids*, tales como *entity graphs* [14]. Este modelo intenta abordar la coherencia global del texto y podría significar mejoras al método propuesto para evaluar ensayos. Además, lo realizado en este trabajo se puede complementar con métodos recientes en la literatura, tales como SAGE [46], pues éste último utiliza representaciones semánticas basadas en lógica, las cuales obtienen buenos resultados para la evaluación de ensayos.

Por otro lado, se ampliará el estudio para verificar si el método propuesto puede utilizarse para crear un sistema más complejo que logre proveer de retroalimentación

específica a un usuario final, para ello se requerirá hacer un estudio exhaustivo de las razones lingüísticas por las cuales el sistema logra evaluar mejor ensayos utilizando patrones de discurso, es decir, establecer umbrales con los que se pueda determinar un *feedback* útil para un usuario final. En este contexto, se podría crear una interfaz de usuario, por ejemplo un sitio web, que a partir de un ensayo escrito por usuario, le entregue retroalimentación automática basadas en las características computadas, esto incluiría aspectos tales como: la centralidad del texto, legibilidad del texto y contenido.

# Bibliografía


[1] Laura Alonso i Alemany and Maria Fuentes Fort. Integrating cohesion and coherence for automatic summarization. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 1–8. Association for Computational Linguistics, 2003.

[2] Yigal Attali. A differential word use measure for content analysis in automated essay scoring. *ETS Research Report Series*, 2011(2):i–19, 2011.

[3] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. .ºReilly Media, Inc.", 2009.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] Jill Burstein, Lisa Braden-Harder, Martin Chodorow, Shuyi Hua, Bruce Kaplan, Karen Kukich, Chi Lu, James Nolan, Don Rock, and Susanne Wolff. Computer analysis of essay content for automated score prediction: A prototype automated scoring system for gmat analytical writing assessment essays. *ETS Research Report Series*, 1998(1):i–67, 1998.

[7] Hongbo Chen, Ben He, Tiejian Luo, and Baobin Li. A ranked-based learning approach to automated essay scoring. In *Cloud and Green Computing (CGC), 2012 Second International Conference on*, pages 448–455. IEEE, 2012.

[8] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[9] Simon Dennis, Tom Landauer, Walter Kintsch, and Jose Quesada. Introduction to latent semantic analysis. In *Slides from the tutorial given at the 25th Annual Meeting of the Cognitive Science Society, Boston*, 2003.

[10] Semire Dikli. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1), 2006.

[11] William H DuBay. *Smart Language: Readers, Readability, and the Grading of Text.* ERIC, 2007.

[12] Scott Elliot. Intellimetric: From here to validity. *Automated essay scoring: A cross-disciplinary perspective*, pages 71–86, 2003.

[13] Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225, 1995.

[14] Camille Guinaudeau and Michael Strube. Graph-based local coherence modeling. In *ACL (1)*, pages 93–103, 2013.

[15] Fritz Günther, Carolin Dudschig, and Barbara Kaup. Lsafun-an r package for computations based on latent semantic analysis. *Behavior research methods*, 47(4):930–944, 2015.

[16] Luo Haijiao and Ke Xiaohua. Study of automated essay scoring based on small dataset extraction algorithm. In *Computer Science and Network Technology (ICCSNT), 2015 4th International Conference on*, volume 1, pages 112–116. IEEE, 2015.

[17] Ben Hamner and Mark D Shermis. Contrasting state-of-the-art automated scoring of essays: analysis. *NCME*, 2012.

[18] Marti A Hearst. The debate on automated essay grading. *IEEE Intelligent Systems and their Applications*, 15(5):22–37, 2000.

[19] SERGIO HERNÁNDEZ OSUNA and ANITA FERREIRA CABRERA. Evaluación automática de coherencia textual en noticias policiales utilizando análisis semántico latente. *RLA. Revista de lingüística teórica y aplicada*, 48(2):115–139, 2010.

[20] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.

[21] Ahmed Ibrahim and Tarek Elghazaly. Improve the automatic summarization of arabic text depending on rhetorical structure theory. In *Artificial Intelligence (MICAI), 2013 12th Mexican International Conference on*, pages 223–227. IEEE, 2013.

[22] Md Monjurul Islam and ASM Latiful Hoque. Automated essay scoring using generalized latent semantic analysis. In *Computer and Information Technology (ICCIT), 2010 13th International Conference on*, pages 358–363. IEEE, 2010.

[23] Cancan Jin and Ben He. Utilizing latent semantic word representations for automated essay scoring. In *Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), 2015 IEEE 12th Intl Conf on*, pages 1101–1108. IEEE, 2015.

[24] Tuomo Kakkonen, Niko Myller, Erkki Sutinen, and Jari Timonen. Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society*, 11(3):275–288, 2008.

[25] Thomas K Landauer, Darrell Laham, and Peter W Foltz. Automated scoring and annotation of essays with the intelligent essay assessor. *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112, 2003.

[26] Mirella Lapata and Regina Barzilay. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, volume 5, pages 1085–1090, 2005.

[27] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[28] Susan M Lottridge, E Matthew Schulz, and Howard C Mitzel. 14 using automated scoring to monitor reader performance and detect reader drift in essay scoring. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, page 233, 2013.

[29] David D Malvern, Brian J Richards, Ngoni Chipere, and Pilar Durán. Lexical diversity and language development. *Houndmills, Hampshire, UK: Palgrave Macmillan*, 2004.

[30] Elijah Mayfield and Carolyn Penstein-Rosé. An interactive tool for supporting error analysis for text mining. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 25–28. Association for Computational Linguistics, 2010.

[31] Andrew Mellor. Essay length, lexical diversity and automatic essay scoring. *Memoirs of the Osaka Institute of Technology*, 55(2):1–14, 2011.

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[33] Eleni Miltsakaki and Karen Kukich. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55, 2004.

[34] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4, 2007.

[35] Xingyuan Peng, Dengfeng Ke, Zhenbiao Chen, and Bo Xu. Automated chinese essay scoring using vector space models. In *Universal Communication Symposium (IUCS), 2010 4th International*, pages 149–153. IEEE, 2010.

[36] Les Perelman. When "the state of the art" is counting words. *Assessing Writing*, 21:104–111, 2014.

[37] Changhua S Rich, M Christina Schneider, and Juan M D'Brot. Applications of automated essay evaluation in west virginia. *Handbook of Automated Essay Evaluation: Current Applications and New Directions, MD Shermis and J. Burstein, Eds. New York: Routledge*, pages 99–123, 2013.

[38] Mark D Shermis, Howard R Mzumara, Jennifer Olson, and Susanmarie Harrington. On-line grading of student essays: Peg goes on the world wide web. *Assessment & Evaluation in Higher Education*, 26(3):247–259, 2001.

[39] Christian Smith and Arne Jönsson. Automatic summarization as means of simplifying texts, an evaluation for swedish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010), Riga, Latvia*, 2011.

[40] Karin Sim Smith, Wilker Aziz, and Lucia Specia. Cohere: A toolkit for local coherence. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[41] Malbert Smith III. The reading-writing connection. *Reading*, 400:200L, 2009.

[42] William N Venables and Brian D Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[43] Walter Zucchini. An introduction to model selection. *Journal of mathematical psychology*, 44(1):41–61, 2000.

[44] Kaja Zupanc and Zoran Bosnic. Automated essay evaluation augmented with semantic coherence measures. In *2014 IEEE International Conference on Data Mining*, pages 1133–1138. IEEE, 2014.

[45] Kaja Zupanc and Zoran Bosnic. Advances in the field of automated essay evaluation. *Informatica*, 39(4):383, 2015.

[46] Kaja Zupanc and Zoran Bosnić. Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, 2017.