Ensamble de Modelos Sintácticos y Semánticos para la Evaluación Automática de Ensayos

Profesor: John Atkinson
Alumno: Diego Andrés Palma Sánchez

Mayo de 2016

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 [22].

En la actualidad, es un tema ampliamente debatido es la capacidad de redacción y comprensión que debiesen tener las personas que egresan del sistema escolar [23]. Esto aborda temas tales como la carencia en el manejo del lenguaje escrito que evidencian los estudiantes en todos los niveles educacionales y estratos socioculturales [24].

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; en síntesis, ideas que podrían ser bastante buenas e innovadoras podrían llegar a verse opacadas o, peor aún, rechazadas por el receptor al ser comunicadas de manera defectuosa.

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. Existen dos propiedades que los buenos textos deben tener, las cuales son *coherencia* y *cohesión* [33].

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. Se debe tener en cuenta la diferencia entre corrección y evaluación [19].

- **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 realizar 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 otro modo que evalúe 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. Sumado a lo anterior, la cantidad de estudiantes ha crecido con el paso del tiempo, por lo que los costos de corregir y revisar textos se vuelven abrumadores [11].

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 *Automatic Essay Scoring* [10]. Los primeros enfoques para esta tarea evaluan características relacionadas a la calidad del texto tales como la dicción, 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 método tiene algunas debilidades [25][26]:

- No se evalúa la extructura 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 se considera la coherencia textual, pues las características con las que el método evalúa un texto no representan su contenido a nivel de las representaciones mentales que tendría un lector.

- No se 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, como por ejemplo la repetición de la palabra estudio y tampoco queda claro que la segunda aparición de la palabra se refiera a lo mismo que se refiere la primera. El pronombre *él* aparece innecesariamente y es redundante luego de mencionar a *Manfred Walzl*. Estos problemas pasan desapercibidos si sólo se considera la frecuencia de términos.

- No se considera la estructura sintáctica del texto. Por ejemplo, se consideraría *"Resfriado me habría la lluvia mojado con me si hubiera"* equivalente a *"Si me hubiera mojado con la lluvia me habría resfriado"*.

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 automática de textos, entre otros[34][35][36].

Existen diferentes enfoques para evaluar coherencia textual:

- Teoría de centrado [37]: intenta caracterizar 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 estos métodos están relacionados con la ambigüedad que presentan algunos textos, por ejemplo cuando se habla sobre múltiples entidades de discurso está el problema de a cuál se hace referencia (*coreference resolution*).

- *Rethorical Structure Theory* [37]: caracteriza la coherencia mediante relaciones existentes entre una entidad principal de un texto y el texto e información que hace referencia a dicha entidad. Se define un conjunto de relaciones, las cuales pueden ser detectadas mediante los marcadores de discurso utilizados en el texto a analizar (como *porque*, *por lo tanto*, etc.). Sin embargo, existen marcadores de discurso que tienen más de un propósito, o mapean a más de una relación, dependiendo de lo que se está expresando en el texto, esto tiene el problema de que en algunos casos se detectan relaciones erróneas.
- Modelos semánticos: representan los textos mediante modelos matemáticos, como por ejemplo un un *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 exista entre las distintas ideas, oraciones y párrafos del texto.

Consecuentemente, se han realizado estudios que comparan el rendimiento de los distintos modelos para evaluar la coherencia textual (generalmente utilizando métricas como correlación con humanos), 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 lo que podrían combinarse para evaluar la coherencia textual [34].

1.1. Hipótesis

Un método de evaluación de textos que considera características sintácticas y semánticas para evaluar coherencia textual es más efectivo para la tarea de evaluación automática de ensayos en comparación a modelos que utilicen medidas superficiales estadísticas (como conteo de palabras, largo de las oraciones, etc.).

1.2. Objetivos

- Objetivo General
  - Desarrollar un método computacional que permita evaluar automáticamente textos en forma de ensayos considerando aspectos de coherencia textual.
- Objetivos Específicos
  - Analizar estrategias de evaluación de ensayos en forma de texto, basados tanto en modelos de estadísticos, como en teoría de discurso.
  - Desarrollar una estrategia que considere coherencia a nivel de contenido y sintaxis.
  - Crear un prototipo para realizar las pruebas.
  - Evaluar el modelo propuesto.

2. Trabajo Relacionado

La primeras técnicas de evaluación automática de textos [1] modelaban los textos como una combinación lineal de sus características intrínsicas (dicción, contenido, fluidez, etc.), las cuales se estiman a través de características superficiales (*proxes*) tales como: la cantidad de palabras, largo del ensayo, cantidad de signos de puntuación utilizados, largo de las oraciones, etc. La evaluación se realiza en dos etapas: entrenamiento y evaluación. En la etapa de entrenamiento se utilizan textos que ya tienen un puntaje asignado por humanos, que representa qué tan correcto es el

texto bajo la evaluación de un experto. Luego se ajusta un modelo lineal mediante una regresión (multi-variable), de manera de ajustar las ponderaciones (pesos) de cada *proxe*. Finalmente, el puntaje de un ensayo se calcula como se muestra en la ecuación 1.

$$Puntaje = \beta_0 + \sum_{i=1}^{n} \beta_i P_i \tag{1}$$

Donde β_i representa la ponderación correspondiente al *proxe* P_i, y representa el impacto que tiene el *proxe* en el puntaje del ensayo. Los mejores resultados experimentales mostraron una correlación de 0.87 entre los puntajes asignados por PEG y los asignados por humanos [1]. Sin embargo, el método utiliza medidas indirectas de la calidad del texto a evaluar lo que lo hace vulnerable a engaños [10], ya que un estudiante podría mejorar su puntaje mediante trucos (por ejemplo escribir un texto más largo). Por otro lado, el uso de estas medidas no captura características importantes tales como contenido, organización, y coherencia. La razón de ello es que el método se fundamenta únicamente en la frecuencia de aparición de cada *proxe* y no en el significado de lo que está escrito ni en las relaciones que existen entre las ideas expresadas en el texto.

Para evaluar el contenido de un texto, se han utilizado técnicas de *Recuperación de Información* (IR) [27] y *Procesamiento del Lenguaje Natural* (NLP) [28]. Para representar el contenido de un texto se utiliza un modelo de IR conocido como *modelo de espacio vectorial* [27], el cual considera un vocabulario previamente definido (términos relevantes que encapsulan el contenido del texto) que se utiliza para representar un documento a partir de la frecuencia de aparición de los términos que lo componen, como se muestra en la ecuación 2.

$$d = (w_1, w_2, ..., w_n) \tag{2}$$

Donde cada componente del vector d representa la frecuencia en que el término w_i aparece en el documento. Con esta representación vectorial se puede establecer una medida de similitud entre dos textos, como por ejemplo similitud coseno (ecuación 3).

$$cos(d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\|\|d_j\|} \tag{3}$$

Donde d_i y d_j son vectores que representan el contenido del texto. Luego, teniendo a disposición un corpus (conjunto de documentos), se puede utilizar esta técnica para encontrar las similitudes ente un documento nuevo y un documento del corpus.

En el contexto de evaluación automática de ensayos, se tiene un conjunto de documentos previamente evaluados por expertos humanos. Luego, para evaluar un texto nuevo, se calcula la similitud entre el documento nuevo y los documentos del corpus, y se asigna el puntaje del más similar, o una suma ponderada de los puntajes más similares.

La aplicación del método ha logrado una correlación con la evaluación realizada por humanos de 0.76 [6]. Sin embargo, la evaluación depende fuertemente de la co-ocurrencia de términos, por lo que un ensayo que sea sólo un conjunto de palabras sin una conexión clara, podría ser bien evaluado [10]. Por otro lado, la técnica está puramente basada en representaciones de palabras, las cuales podrían no aparecer explícitamente en los ensayos (por ejemplo sinónimos).

Para abordar el problema de la co-ocurrencia de palabras, se ha aplicado una técnica de reducción dimensional conocida como *Análisis Semántico Latente* (LSA) [29], la cual intenta extraer y representar el significados de las palabras, lo que permite encontrar similitudes entre documentos aunque no haya co-ocurrencia de términos. El método toma como supuesto que existen relaciones que subyacen de forma latente en la estructura semántica de los datos, y que se encuentran parcialmente ocultas. Este método requiere un corpus, y que se defina un

vocabulario con las palabras más relevantes del corpus para representar los documentos en un espacio vectorial. El método considera un modelo de *bolsas de palabras* (BOW), es decir, asume que el orden de las palabras no importa. Posteriormente, se obtiene una matriz de palabras y documentos del corpus. Luego, se aplica una técnica matemática conocida como *descomposición de valores singulares* (SVD), para obtener una matriz reducida que en este contexto se denomina *espacio semántico*. Las dimensiones de este espacio semántico son una combinación lineal de documentos, las que se interpretan como *conceptos*, y posteriormente se puede definir una medida de similitud entre ellos. Luego, se puede medir la coherencia textual [9] [11] [21] [30], y evaluar ensayos [10].

La evaluación de ensayos en LSA se realiza en dos etapas. En la primera etapa se requiere un corpus que contenga documentos relacionados al tópico a evaluar. Luego, se aplica LSA para obtener un espacio semántico. En la segunda etapa, se consideran ensayos pre-evaluados por humanos, y se comparan con ensayos nuevos utilizando algún criterio de similitud (por ejemplo coseno). Esta comparación se realiza dentro del espacio semántico obtenido en la etapa 1 [11]. Finalmente, al ensayo se le asigna la nota del ensayo más similar dentro del espacio semántico.

Un problema con LSA es que al utilizar un modelo de bolsa de palabras no considera el orden de las mismas. Por ejemplo se consideraría equivalente "literatura fantástica" con "fantástica literatura". El método tampoco considera la forma en que están escritas las oraciones, ya que un texto podría ser bien evaluado si cumple con los criterios de similaridad. Por otro lado, LSA tiene problemas con la polisemia (palabras que tienen distintos significados), y ello afecta al momento de encontrar similitud entre dos documentos, pues dos documentos pueden utilizar una misma palabra pero hablar de temas diferentes. Otro problema ocurre con la elección de la dimensionalidad del espacio semántico, la cual se realiza utilizando heurísticas *ad-hoc*, por ejemplo se dice que se obtienen buenos resultados si se reduce la dimensionalidad a valores entre 100-300 [11].

Para abordar parcialmente el problema relacionado al orden de las palabras, se ha propuesto una variación de LSA denominada *Generalized Latent Semantic Analysis* (GLSA) [14]. Este método utiliza frecuencia de n-gramas en lugar de palabras, lo que permite distinguir segmentos de texto como por ejemplo "dióxido de carbono" con "Carbono de dióxido", lo que LSA convencional consideraría como equivalente. Para la evaluación automática de ensayos, se realiza el mismo procedimiento que LSA, por lo que la única diferencia entre ambos métodos es el espacio semántico obtenido. En LSA se obtienen conceptos como combinación lineal de los vectores que representan las palabras, mientras que en GLSA son combinaciones lineales de vectores de n-gramas. Experimentalmente se ha obtenido una correlación 0.88 entre el puntaje asignado por humanos y el asignado por el método [14]. Algunos problemas de esta técnica incluyen:

- Alta dimensionalidad del espacio vectorial debido a los n-gramas.
- Alto grado de dispersión, pues la co-ocurrencia de n-gramas es menos probable que la de palabras. Esta dispersión puede producir errores numéricos al aplicar descomposición de valores singulares.
- La descomposición de valores singulares es costosa computacionalmente y el costo crece exponencialmente dependiendo de los n-gramas a utilizar.

El problema de la polisemia no es totalmente resuelto por GLSA, ya que depende de la distribución de n-gramas utilizados para capturar el contexto en que la palabra fue utilizada. Por otro lado, GLSA tampoco tiene una base estadística sólida (no define un modelo generativo de los datos) lo que dificulta interpretar el espacio semántico obtenido. Es por ello que se ha propuesto un método probabilístico denominado *Probabilistic Latent Semantic Analysis* [42], el cual considera que los datos pueden expresarse en términos de 3 tipos de variables:

- Documentos: Los documentos del corpus a utilizar.
- Palabras: El vocabulario a considerar en el corpus.
- Tópicos: Variables ocultas.

PLSA se fundamenta en un modelo probabilístico denominado *Aspect Model*, el cual expresa que las variables ocultas (tópicos) están asociadas a las variables observadas (documentos y palabras). En la figura 1 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 el documento d. Cada palabra w tiene asociada un tópico z (variable latente) que la genera.

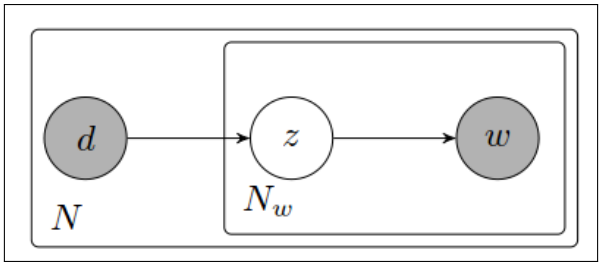

Figura 1: Representación gráfica de PLSA.

Las ventajas que posee PLSA respecto a LSA son incluyen:

- Modelo con base estadística sólida, donde los parámetros están bien definidos y tienen interpretación probabilística clara (variables latentes como tópicos), a diferencia de LSA donde el espacio semántico no tiene una interpretación clara.
- Al estar basado en un enfoque probabilístico se puede utilizar teoría estadística para la selección del modelo (cantidad de tópicos a considerar, es decir, la dimensionalidad del espacio semántico), a diferencia de LSA donde se hace con heurísticas *ad-hoc*.
- Logra solventar el problema de la polisemia, pues se tienen grupos de palabras que representan un determinado tópico (variable oculta).

Este método se ha utilizado para la evaluación automática de ensayos, los resultados obtenidos son similares a los de LSA en conjuntos de datos pequeños, y mejores que LSA en corpus con más documentos.

En general, los métodos basados en modelos de espacio vectorial tienen el problema de que no son capaces de detectar errores semánticos, como por ejemplo hechos contradictorios dentro de un texto, o hechos inconsistentes dentro del dominio en el que se quiere evaluar el texto. Debido a esto, se ha propuesto el uso de ontologías para la detección de errores semánticos [44]. Una ontología es una definición formal de tipos, propiedades, y relaciones entre entidades con respecto a un dominio específico. Básicamente, contiene axiomas que representan el dominio y sus características. Para evaluar la coherencia textual en un dominio específico, lo que se hace es crear una base de conocimiento que tenga axiomas sobre el dominio a considerar. Luego, para evaluar un texto, se genera una representación lógica del contenido del mismo y se evalúa la consistencia de los hechos mencionados en el texto con respecto a la base de conocimiento del dominio. Cualquier hecho que sea incosistente se considerará como un error semántico. Se ha encontrado que este método puede detectar errores semánticos que pasan desapercibidos en

métodos basados en un modelo de espacio vectorial como LSA [45]. Un problema con este método es que la generación de la ontología y los axiomas se hace manualmente, por lo que se vuelve poco práctico si se tiene un corpus con miles de documentos. En los experimentos reportados se utilizan conjuntos de datos del orden de los 30 - 50 documentos.

Por otro lado, se han propuesto nuevas medidas de coherencia para evaluar textos automáticamente [41]. Estas medidas están basadas en un modelo de espacio vectorial. Se divide un texto en segmentos fundamentales (Introducción, desarrollo, y conclusión), luego estos segmentos son llevados a un espacio vectorial, y se representan como puntos en dicho espacio. En esta representación, se definen características tales como distancia entre oraciones más distantes, distancia promedio entre oraciones, distancia entre párrafos, y otras medidas estadísticas relacionadas a la distribución de los datos, como por ejemplo distancia al centroide de cada párrafo. Estas características se interpretan como una medida de la coherencia local y global del texto a evaluar. Se encontró que agregar estas características a un método convencional de evaluación automática de ensayos, mejora la precisión y el grado de acuerdo con los puntajes asignados por humanos y se pueden obtener correlaciones de hasta 0.9, lo que supera a los métodos en el estado del arte [41].

Existen otros métodos para medir coherencia textual que consideran la sintaxis de un texto [34] [35] [36] y se fundamentan en la *teoría de centrado* [37]. Esta establece que un discurso está conformado por segmentos de texto y que cada segmento de texto está conformado por expresiones o enunciados (*utterances* U_i - U_n). Cada expresión (U_i) tiene asociada un conjunto de entidades de discurso o centros del discurso (los cuales pueden ser nombres, pronombres, objetos, etc.), el cual se denomina *forward looking centers* $C_f(U_i)$. Los elementos de C_f se rankean de acuerdo a su importancia dentro del discurso. El centro de C_f que tiene el mayor rango se denomina C_p o el centro preferido. También se define un centro que relaciona la expresión actual con la anterior, y se denomina *backward looking center* C_b, y se define como el centro con mayor rango en la expresión previa U_{i-1} que es mencionado en la exprsesión U_i. C_b es un tipo especial de centro ya representa el tema que está tratando U_i.

La teoría define cuatro tipos de transiciones, las cuales se muestran en la tabla 1.

Tabla 1: Tabla de transiciones en la Teoría del Centrado

	$C_b(U_i) = C_b(U_{i-1})$_	$C_b(U_i) \neq C_b(U_{i-1})$
$C_b(U_i) = C_p(U_i)$	*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, el cual se define como: 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 expresiones (*utterance*) mientras que textos poco coherentes tienden a hacer cambios bruscos de centros (entidades de discurso). Para entender la idea, se propone el siguiente ejemplo:

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

b) Cuando la reunión terminó
($C_b = ninguno$ $C_f = Reunion$,$transicion$ =*Rough Shift*)

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

Observamos que el fragmento de discurso es complicado de leer, debido a que el foco en la segunda expresión cambia bruscamente, primero se hablaba de *Juan* y su *Dolor de Cabeza* y luego se habla sobre una *reunión*. Se observa que si se cambia el orden de las oraciones, se torna en un texto más coherente:

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

b) él corrió a la farmacia
($C_b = Juan$ $C_f = Juan > Farmacia$,$transicion$ =$Continue$)

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

En este caso es claro que *Juan corrió* a la *farmacia* una vez terminada la *reunión*.

Esta teoría ha sido utilizada para determinar la correlación entre la proporción de transiciones *Rough Shift* que ocurren en un discurso y el puntaje asignado por humanos a un texto. Se ha encontrado que mientras mayor sea esta proporción, los textos tienden a ser menos coherentes [46]. La principal ventaja, es que se considera la sintaxis del texto (el orden de las expresiones es relevante). La desventaja que tiene es que depende fuertemente de la precisión con la que se capturan las entidades de discurso y se identifican los conjuntos C_f. En [46] se utilizaron textos etiquetados por humanos para la extracción de centros de discurso, debido a que métodos automáticos tienen problemas al momento de identificar las coreferencias dentro del texto.

Otro método de evaluación basado en esta teoría, toma como supuesto que existen patrones que es más probable que aparezcan en discursos coherentes. Para detectar estos patrones, un texto se representa mediante una matriz de entidades. Un ejemplo de texto y su matriz de entidades se muestran en las figuras 1 y 2.

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: Texto con anotaciones sintácticas para el cálculo de matriz de entidades.

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. Los roles gramaticales reflejan 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.

Así, la coherencia de un texto $T(S_1, ..., S_n)$ con entidades $e1, ..., e_m$, se puede ver como una distribución de probabilidad conjunta que describe cómo las entidades están distribuidas a través de las oraciones del texto T:

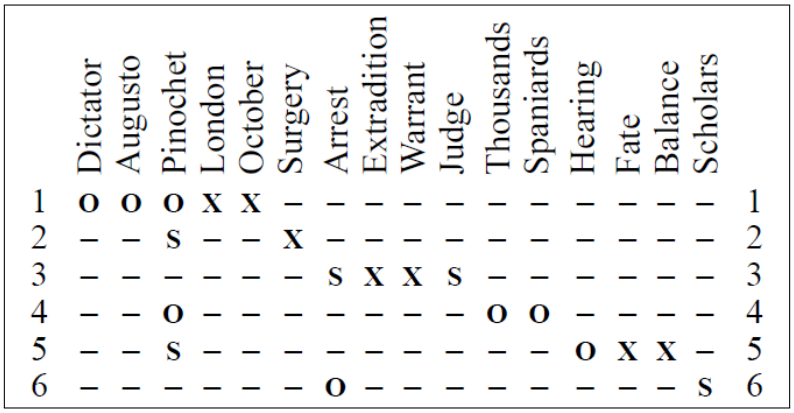

Figura 3: Una matriz de entidades

$$P_{coherence}(T) = P(e_1, ..., e_m; S_1, ..., S_n) \tag{4}$$

Para generar este modelo de distribución, se requiere un conjunto de textos que los humanos consideren coherentes [37]. Luego, se puede predecir $P(e_1, ..., e_m; S_1, ..., S_n)$ en textos en los que se quiera evaluar la coherencia. Luego, $P_{coherence}(T)$ será mayor para textos que se consideren más coherentes que los que tengan un $P_{coherence}(T)$ menor.

El método también considera una componente semántica que modela la forma en que se conectan oraciones en términos de la representación mental que hace un lector al leer el texto. La cohesión léxica se representa a través de un modelo de *cadenas léxicas* [40], es decir, secuencias de palabras relacionadas que abarcan una unidad textual (por ejemplo: oración, párrafo, etc.). De ahí que las unidades textuales coherentes deberán tener una alta concentración de estas cadenas, pues un texto coherente posee muchas palabras relacionadas semánticamente. Esto permite representar un texto sin considerar el orden de las palabras, luego, cada oración es un conjunto de palabras. Así, para medir la coherencia local de un texto se debe cuantificar la relación semántica entre oraciones adyacentes. Es decir, la coherencia de un texto T se puede considerar como el promedio del grado de similaridad entre oraciones:

$$coherencia(T) = \frac{\sum_{i=1}^{n-1} sim(S_i, S_{i+1})}{n-1} \tag{5}$$

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

Este método ha sido comparado con otros enfoques para medir coherencia textual (entre ellos LSA, basados en ontologías, basados en características superficiales) en el ámbito de resumenes generados automáticamente [34]. Los resultados muestran que no existe correlación entre los métodos, por lo que cada uno mide distintas componentes de la coherencia textual. Así, un modelo que considere estas componentes tendrá un mejor desempeño en términos de correlación con humanos que cualquier modelo por sí solo.

Tomando lo anterior, un modelo que considere la sintaxis y semántica de un ensayo podría mejorar el rendimiento de la evaluación debido a que consideraría diferentes componentes de la coherencia textual.

3. Propuesta

Se propone desarrollar un método computacional que permita evaluar automáticamente ensayos considerando aspectos de coherencia textual y que considere la sintaxis y semántica. Para ello, se tomarán como punto de partida los métodos basados en un modelo de espacio vectorial y los métodos que se fundamentan en teoría de discurso.

Se tomará como *baseline* el método descrito en [41], el cual considera las características estadísticas que se muestran en la tabla 2 para luego aplicar una regresión. Por otro lado, se considerarán otros métodos para medir la coherencia textual, entre ellos LSA y los métodos que utilizan teoría de centrado.

Tabla 2: Atributos Lingüísticos y de Contenido

lexical sophistication	**grammar**
number of characters	*Number of different POS tags*
number of words	Number of Each POS tag
number of long words	*adverb*
number of short words	*coordinating conjunction*
most frequent word length	*numeral*
average word length	*determiner*
number of sentences	*existential there*
number of long sentences	*preposition*
number of short sentences	*adjective*
most frequent sentence length	*superlative adjective*
average sentence length	*comparative adjective*
number of different words	*modal auxiliary*
number of stop words	*singular or mass common noun*
Readability measures [47], [48]	*plural common noun*
Gunning Fog Index	*singular proper noun*
Flesch Reading Ease	*preposition participle*
Flesch Kincaid Grade Level	*predeterminer*
Dale Chall Readability formula	*personal pronoun*
automated readability index	*possessive pronoun*
Simple Measure of Gobbledygook	*adverb*
LIX	*genitive marker*
Word Variation Index	*comparative adverb*
Nominal Ratio	*superlative adverb*
Lexical Diversity [49]	*verb - base form*
TTR	*verb - past tense*
Guiraud's index	*verb - gerund/present participle*
Yule's K	*verb - past participle*
hapax legomena	*wh-determiner*
Advanced Giraud	*wh-pronoun*
	possessive wh-pronoun
	wh-adverb

Content
1. cosine similarity with source text
2. score point level for maximum cosine
similarity over all score points
3. cosine similarity with essays that
have highest score point level
4. pattern cosine [50]
5. weighted sum of all cosine correlation values [50]

4. Metodología Experimental

1. Revisión bibliográfica de métodos para evaluación de coherencia a nivel de discurso.
2. Establecer una representación de textos con la que se pueda modelar la sintática y semántica del contenido textual.
3. Recopilación de datos de ensayos evaluados por humanos, los cuales se limpiarán y prepararán para utilizarlos en el método a desarrollar. Para este procesamiento y limpieza se utilizarán herramientas existentes como por ejemplo nltk, pyEnchant, entre otros. Los datos a utilizar serán los proporcionados por Kaggle[1] en la competencia *Automatic Essay Scoring*, el cual contiene datos de ocho categorías diferentes de ensayo, que del total cuatro de ellas consisten en géneros de escritura tradicional (persuasivo, narrativo, etc.) y los otros cuatro están basados en una fuente (es decir, los estudiantes leen un documento fuente y discuten preguntas respecto a dicho documento).
4. Desarrollo de un método de evaluación que considere sintaxis y semántica del ensayo a evaluar. Para ello, se estudiarán modelos de evaluación de coherencia textual que se fundamenten en teoría de discursos, pues estos consideran estas dos componenetes.
5. Implementación de un prototipo computacional de los distintos métodos propuestos en la literatura, con la finalidad de realizar experimentos que validen la hipótesis.
6. Evaluación del método propuesto y comparación con otros métodos existentes en la literatura (basados en semántica (LSA), y basados en características superficiales(PEG)), para ello se utilizarán métricas tales como:
   - *Exact Agreement*: Mide la proporción de ensayos que fueron calificados igualmente por el evaluador humano y el método computacional.
   - *Adjacent Agreement*: Mide la proporción ensayos que fueron evaluados igual por el evaluador humano y el método computacional o que difiere en a lo más 1 punto (de calificación).
   - *Quadratic Weighted Kappa*: Mide el grado de acuerdo entre los puntajes asignados por dos evaluadores. Su valor máximo es 1, si los evaluadores están completamente de acuerdo, si hay desacuerdo, la métrica puede tomar valores negativos.
   - *Correlación de Spearman*: Mide qué tan buena es la relación entre la evaluación humana y la automática.

5. Plan de Trabajo

- Revisión bibliográfica de métodos de evaluación de coherencia textual, basados en teoría de discurso: 1 Enero - 1 Marzo.
- Diseño de un método automático para evaluar ensayos: 7 Marzo - 7 Mayo.
- Ensamble de modelos que consideren sintaxis y semántica: 15 Mayo - 5 Junio
- Crear prototipo para realizar pruebas 10 Mayo - 10 Junio.
- Evaluar rendimiento del modelo y comparar con modelos del estado del arte: 15 Junio - 1 Julio.

[1] http://www.kaggle.com/c/asap-aes/data

Referencias

[1] S. Valenti, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading," *Journal of Information Technology Education*, vol. 2, pp. 319-330, 2003.

[2] T. Miller, "Essay assessment with latent semantic analysis," *Department of Computer Science, University of Toronto*, Toronto, ON M55 3G4, Canada, 2002.

[3] L. M. Rudner and T. Liang, "Automated essay scoring using Bayes' Theorem," *The Journal of Technology, Learning, and Assessment*, vol. 1, no. 2, 2002.

[4] K.M Nahar and L.M. Alsmadi, "The automatic grading for online exams in Arabic with essay questions using statistical and computational linguistics techniques," *MASAUM Journal of Computing*, vol. 1, no. 2, 2009.

[5] S Ghosh and S. S. Fatima, "Design of an Automated Essay Grading (AEG) system in Indian context," in *Proceedings of TENCON 2008-2008 IEEE Region 10 Conference*, pp. 1-6.

[6] Y. Attali, J. Burstein, "Automated essay scoring with e-rater," *The Journal of Technology, Learning and Assessment*, vol. 4, no.3, 2006.

[7] L.M. Rudner, V. Garcia, C.Welch, "An evaluation of the IntelliMetric essay scoring system," *The Journal of of Technology Learning, and Assessment*, vol. 4, no. 4, pp. 1-22, 2006.

[8] P. W. Foltz, D. Laham, T.K. Landauer, "Automated essay scoring: applications to educational technology," in *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 1999, pp.939-944.

[9] B. Lemaire, P. Dessus, "A system to assess the semantic content of student essay," *The Journal of Educational Computing Research*," vol. 24, no. 3, pp. 305-320, 2001.

[10] M.A. Hearst, "The debate on automated essay grading," *Intelligent Systems and their Applications*, IEEE , vol.15, no.5, pp.22-37, Sept.-Oct. 2000.

[11] T. Kakkonen, N. Myller, E. Sutinen, J. Timonen, "Comparison of Dimension Reduction Methods for Automated Essay Grading," *Educational Technology and Society*, 2006, pp. 275-288.

[12] P. Selvi, N.P. Gopalan, "Automated writing Assessment of Student's Open-ended Answers Using the Combination of Novel Approach and Latent Semantic Analysis," *Advanced Computing and Communications* ADCOM 2006. International Conference on , vol., no., pp.370-375, 20-23 Dec. 2006.

[13] G. Russo-Lassner, J. Lin, P. Resnik, "A Paraphrase-Based Approach to Machine Translation Evaluation," *Technical Report* LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland, College Park, August 2005.

[14] M.M. Islam, A.S.M.L. Hoque, "Automated essay scoring using Generalized Latent Semantic Analysis," *Computer and Information Technology (ICCIT)*, 13th International Conference on, pp.358-363, 2010.

[15] H. Chen, B. He, T. Luo, B. Li, "A Ranked-Based Learning Approach to Automated Essay Scoring," *Cloud and Green Computing (CGC)*, Second International Conference on, pp.448-455, 1-3 Nov. 2012.

[16] C.D. Manning, H, Schutze, "Foundations of Statistical Natural Language Processing," Cambridge, MA: MIT Press, 1990.

[17] L. Rudner, L. Tahung, "Automated Essay Scoring using Bayes' Theorem," *The Journal of Technology, Learning, and Assessments*, 3-21, 2002.

[18] H. Breland, R. Jones, Laura J., "The College Board Vocabulary Study," *College Entrance Examination Board*, New York, 1994.

[19] C. Moncayo, F. Julio. "La terminología como elemento de cohesión en los textos de especialidad del discurso económico-financiero", Tesis Doctoral, Facultad de Filosofía y Letras, Universidad de Valladolid, pp. 1-50, 2002.

[20] McCarthy, Philip M.; Briner, Stephen W.; Rus, Vasile y McNamara, Danielle S. "Textual Signatures: Identifying Text-Types Using Latent Semantic Analysis to Measure the Cohesion of Text Structures", Institute for Intelligent Systems, University of Memphis, USA., pp. 1-15, 2005.

[21] E. Kintsch, D. Steinhart, G. Stahl, and the LSA Research Group, Developing Summarization Skills through the Use of LSA-Based Feedback, Interactive Learning Environments, 8:2, pp. 87-109, 2000.

[22] Organización para el Desarrollo Económico (OECD). "Informe sobre los resultados de los estudiantes chilenos en el estudio PISA 2012", pp. 60-78, 2012. Unidad de Curriculum y Evaluación, Ministerio de Educación, Chile.

[23] Fernández, Aguistín. "Aprender a Leer: una tarea de todos y de siempre", en Revista Digital Umbral 2000, tomo 13, pp. 1-10, 2003.

[24] Rivera Lam, Mailing. "Estrategias de lecturas para la comprensión de textos escritos: El pensamiento reflexivo y no lineal en alumnos de educación superior", en Revista Digital Umbral 2000, tomo 12, pp. 1-14, 2003.

[25] Chung, Gregory K.W.K., and Eva L. Baker (2003). "Issues in the Reliability and Validity of Automated Scoring of Constructed Responses", p. 23. In: Automated Essay Scoring: A Cross-Disciplinary Perspective. Shermis, Mark D., and Jill Burstein, eds. Lawrence Erlbaum Associates, Mahwah, New Jersey, ISBN 0805839739.

[26] Dikli, Semire (2006). "An Overview of Automated Scoring of Essays". Journal of Technology, Learning, and Assessment, 5.

[27] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "Introduction to Information Retrieval", Cambridge University Press. 2008.

[28] Daniel Jurafsky and James H. Martin, "An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", Second Edition. 2009.

[29] T. Landauer, S. Dumais, "Introduction to Latent Semantic Analysis", in: Discourse Processes 25, pp. 259-284, 1997.

[30] S. Hernández, A. Ferreira, "Evaluación Automática de Coherencia Textual en Noticias Policiales Utilizando Análisis Semántico Latente", Revista de Lingüística Teórica y Aplicada. Concepción (Chile), 48 (2), II Sem. 2010, pp. 115-139.

[31] F. Wild, “Parameters Driving Effectiveness of Automated Essay Scoring with LSA”, IN: Proceedings of the 9th CAA Conference, Loughborough: Loughborough University.

[32] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012). “Foundations of Machine Learning,” The MIT Press. ISBN 978-0-262-01825-8.

[33] Grosz, Barbara J, Pollack, Martha E. y Sidner, Candace L., “The Discourse”, The MIT Press, pp. 437-467, 1989.

[34] M. Lapata and R. Barzilay, “Automatic evaluation of text coherence: models and representations,” In Proceedings of the 19th international joint conference on Artificial intelligence (IJCAI 05). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1085-1090.

[35] 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 (EACL 03), Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 1-8.

[36] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown, “The Pyramid Method: Incorporating human content selection variation in summarization evaluation,” ACM Trans. Speech Lang. Process. 4, 2, Article 4 (May 2007)

[37] B. Grosz, A. Joshi, S. Weinstein, “Centering: A framework for modeling the local coherence of a discourse,” Computational Linguistics, pp. 203-225.

[38] Yen-Yu Chen, Chien-Liang Liu, Tao-Hsing Chang, Chia-Hoang Lee, “An Unsupervised Automated Essay Scoring System,” in Intelligent Systems, IEEE , vol.25, no.5, pp.61-67, Sept.-Oct. 2010.

[39] M. Collins, “Head-driven Statistical Models for Natural Language Parsing,” PhD thesis, University of Pennsylvania, 1998.

[40] J. Morris and G. Hirst, “Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 1(17):21-43, 1991.

[41] K. Zupanc, Z. Bosnic, “Automated Essay Evaluation Augmented with Semantic Coherence Measures,” in Data Mining (ICDM), 2014 IEEE International Conference on , vol., no., pp.1133-1138, 14-17 Dec. 2014

[42] T. Hoffman, “Unsupervised learning by probabilistic latent semantic analysis”, Machine Learning, 42 (1-2), pp. 177-196, 2001.

[43] T. Kakkonen, N. Myller, E. Sutinen, J. Timonen, “Comparison of Dimension Reduction Methods for Automated Essay Grading,” Educational Technology and Society, pp. 275-288, 2008.

[44] F. Gutierrez, D. Dou, S. Fickas, G. Griffiths, “Online Reasoning for Ontology-Based Error Detection in Text”, OTM 2014, LNCS 8841, pp. 562-579, 2014.

[45] F. Gutierrez, D. Dou, S. Fickas, G. Griffiths, “Providing Grades and Feedback for Student Summaries by Ontology-based Information Extraction,” CIKḾ12, 2012.

[46] E. Miltasaki, K. Kukich, “Evaluation of text coherence for electronic essay scoring systems,”. Natural Language Engineering, 2004.

[47] W. H. “Dubay, Smart Language: Readers , Readability , and the Grading of Text”. BookSurge Publishing, 2007.

[48] C. Smith and A. Jönsson, “Automatic Summarization As Means Of Simplifying Texts, An Evaluation For Swedish”, in Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011, B. Sandford Pedersen, G. Nespore, and I. Skadina, Eds., 2011, pp. 198?205.

[49] A. Mellor, “Essay Length , Lexical Diversity and Automatic Essay Scoring,” Memoirs of the Osaka Institute of Technology, vol. 55, no. 2, pp. 1?14, 2011.

[50] Y. Attali, “A Differential Word Use Measure for Content Analysis in Automated Essay Scoring,” ETS Research Report Series, vol. 36, 2011.