Ensamble de Modelos Sintácticos y Semánticos para la Evaluación Automática de Textos en forma de Ensayos

Alumno: Diego Andrés Palma Sánchez
Profesor: John Atkinson

Enero 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 que define las conexiones semánticas entre las partes de un texto (por ejemplo: oraciones, párrafos) y está asociada con las representaciones mentales que tiene una persona al leer un determinado texto. 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 propender al aumento de 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 disminuir 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*. Los primeros métodos propuestos para esta tarea intentan evaluar características relacionadas a la calidad del texto tales como la dicción, uso de vocabulario, coherencia, entre otras. Para ello, utilizan la ocurrencia de términos (conteo de signos de puntuación, largo de las palabras, entre otros).

Sin embargo, este método de evaluación automática de textos tiene algunas debilidades [10][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 diferentes las oraciones.

- No se considera la cohesión de un texto, lo que trae como consecuencia que no se evalúa la calidad del 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 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"*.

- 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.

Debido a los problemas mencionados, nace la necesidad de poder evaluar automáticamente qué tan bien están expresadas las ideas en un texto y cuán coherentes son. La *evaluación automática de coherencia textual* [34] es un problema de investigación que aún se encuentra abierto y tiene múltiples aplicaciones, como por ejemplo: generación automática de resúmenes [35][36], traducción automática, generación automática de texto, entre otros.

Existen métodos para evaluar coherencia textual basados en la teoría de centrado [37], la cual 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.

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, también se ha concluido que los distintos métodos evalúan propiedades complementarias de coherencia [34].

1.1. Hipótesis

Un método 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 de estas características (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
  - Establecer una representación de textos con la que se pueda modelar la sintática y semántica del contenido textual.
  - 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. Estado del Arte

La primeras técnicas de evaluación automática de textos (como PEG [1]) modelan los textos como una combinación lineal de sus características intrínsicas (dicción, contenido, fluidez, etc.). Estas características intrínsicas se estiman mediante características superficiales (*proxes*) tales como: la cantidad de palabras, largo del ensayo, cantidad de signos de puntuación utilizados, largo de las oraciones, etc. El método consta de dos etapas, una de entrenamiento y una de evaluación. En la etapa de entrenamiento se utilizan textos que ya tienen un puntaje asignado. Luego, existe una fase de extracción de características para posteriormente aplicar una regresión lineal (multi-variable), de manera de ajustar las ponderaciones (pesos) de cada característica. 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. 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 cual no deja al método exento de críticas [10]. El uso medidas indirectas deja al método vulnerable a engaños, ya que los estudiantes podrían mejorar sus puntajes obtenidos mediante trucos (por ejemplo escribir un texto más largo). Por otro lado, se ha argumentado que el uso de estas medidas indirectas 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 esctrito ni en las relaciones que hay entre las ideas expresadas en el texto, lo cual es un problema pues no se está asignando un puntaje acorde al contenido del texto.

Para poder evaluar el contenido de un texto, posteriores investigaciones se centraron en utilizar 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. Este modelo considera un corpus (conjunto de documentos), el cual tiene asociado un vocabulario, que son las palabras que aparecen en los documentos del corpus (dimensiones del espacio). Teniendo este espacio vectorial, se puede representar un documento a partir de la frecuencia de aparición de los términos que lo componen (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. Los términos dependerán de la aplicación y pueden ser palabras, oraciones, párrafos, etc. Con esta representación vectorial se puede establecer una medida de similitud entre dos textos, como por ejemplo similitud coseno, que se muestra en 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. Utilizando esta medida de similitud y, teniendo a disposición un conjunto de ensayos pre-evaluados, se pueden evaluar automáticamente ensayos nuevos, buscando para ello el ensayo pre-evaluado más similar.

Se ha logrado una correlación entre los puntajes asignados por la técnica y los asignados 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 "palabras claves" las cuales podrían no aparecer explícitamente en los ensayos. Esto es un problema, pues existen palabras diferentes que tienen un mismo significado (sinónimos) y que podrían no aparecer lo suficiente en un documento, lo cual traería como consecuencia una mala evaluación, según el método. También hay problemas con polisemia, pues un caso posible es tener un documento que utilice una palabra que por ocurrencia es importante, pero que esté siendo utilizada en un contexto diferente al que se quiere evaluar.

Para abordar problemas como la aparición de sinónimos, y la alta dimensionalidad del modelo de espacio vectorial, se ha propuesto el uso de técnicas de reducción dimensional como el *Análisis Semántico Latente* (LSA) [29]. El Análisis Semántico Latente es una técnica que intenta extraer y representar los significados de las palabras. La técnica toma como supuesto que existe algo que subyace latente en la estructura semántica de los datos, y que está parcialmente oculto a causa de la elección aleatoria de palabras. LSA puede utilizarse para medir coherencia textual [9] [11] [21] [30], y evaluar ensayos[10]. Para evaluar ensayos, se entrena LSA con un corpus de documentos, entre ellos ensayos pre-evaluados, y se obtiene un espacio semántico. Los ensayos nuevos son llevados a este espacio y se evalúan de acuerdo a algún criterio de similaridad como los descritos previamente[11].

LSA tiene varias debilidades, como por ejemplo:

- No aborda el problema de la polisemia (palabra con diferentes significados). Podrían considerarse similares documentos que hablan de temas diferentes.
- No se considera el contexto de las palabras (esto es una consecuencia de utilizar un modelo de bolsa de palabras).
- Es difícil interpretar los significados en el espacio semántico, y esto es un problema pues se quiere asegurar una evaluación similar a la que daría un humano.
- No se define un modelo probabilístico en la ocurrencia de términos. Esto es un problema, pues no se asocia una palabra a un determinado tópico, y por ende, la efectividad del método está fuertemente ligada a los documentos que se utilicen al entrenar.

Para solventar parcialmente el problema de utilizar un modelo de bolsa de palabras se ha propuesto un método denominado [14]: Generalized Latent Semantic Analysis (GLSA). Esta método representa los textos utilizando ffrecuencia n-gramas en vez de palabras. La ventaja que tiene esto, es que se hace distinción en segmentos de texto como por ejemplo "dióxido de carbono" con "Carbono de dióxido", lo que LSA convencional consideraría como equivalente. Para aplicar esta GLSA en la evaluación automática de ensayos, se realiza el mismo procedimiento que con LSA. Experimentalmente se ha obtenido una correlación 0.88 entre el puntaje asignado por humanos y el asignado por el método [14]. Algunos problemas de de este método:

- Alta dimensionalidad.
- La descomposición de valores singulares es costosa computacionalmente.
- LSA convencional ha obtenido correlaciones con humanos de hasta 0.86, lo que indica que GLSA no presenta una mejora significativa para la evaluación automática de textos.

Tanto LSA como GLSA se fundamentan en una técnica matemática denominada descomposición de valores singulares, por lo que sufren del problema de no definir un modelo probabilístico en la courrencia de términos. Para solventar este problemas, se han propuesto métodos como PLSA (*Probabilistic Latent Semantic Analysis*) y LDA (*Latent Dirichlet Allocation*). Estos métodos representan los documentos como mezclas de tópicos, los cuales están representados por una distribución de probabilidad de palabras. Por ejemplo el texto "*El hamster está comiendo brocoli*" podría ser representado por dos tópicos (comida y animales).

Para aplicar PLSA y LDA en la evaluación automática de ensayos se requiere encontrar los tópicos que representan a los documentos. Teniendo los tópicos y la distribución de palabras que los representan, se pueden realizar calculos de similaridad entre documentos, y aplicar métodos como kNN, para dar nota a nuevos ensayos de acuerdo a la similaridad que tengan con ensayos previamente evaluados. Se ha encontrado que PLSA y LDA no obtienen mejores resultados que LSA para la tarea de evaluación automática de ensayos [43].

Por otro lado, se han propuesto nuevas medidas de coherencia semántica para resolver el problema de AEG mediante clasificación [41]. Estas medidas están basadas en un modelo de espacio vectorial como el descrito previamente. Lo que se hace es dividir cada ensayo en partes fundamentales (Introducción, desarrollo, y conclusión), y se extraen características tales como: oraciones más distantes semánticamente, distancia de las partes a un centroide, distancia promedio entre oraciones más cercanas, entre otros. Las pruebas demostraron que agregando estas medidas de coherencia se obtienen resultados mejores que sin considerarlas, y que superan a modelos en la literatura en términos de correlación entre notas asignadas por el sistema y por evaluador humano.

Otras técnicas de evaluación consideran métodos de aprendizaje no supervisado es decir, "no hay datos previamente etiquetados por humanos". Este tipo de técnicas intentan resolver el problema de requerir demasiados datos pre-evaluados. En [38], se utiliza un método de clustering basado *Z-score*. La técnica ha logrado buenos resultados en cuanto a *accuracy* de la evaluación. Los problemas con este método son los siguientes:

- Utiliza modelo de bolsa de palabras, por lo que tiene los mismos problemas que los métodos anteriores.
- El cálculo de los Z-score iniciales del método depende de una medida indirecta (cantidad de palabras "únicas" del ensayo).
- Para interpretar los clusters obtenidos se requiere un historial de notas de ensayos, o suponer una distribución normal.

Existen otros métodos para medir coherencia textual que consideran la sintaxis de un texto [34][35][36] y su fundamentan en teoría de centrado [37]. La teoría de centrado estudia cómo se introducen y discuten entidades dentro de un discurso. En particular, se establece que los segmentos de un discurso que se centran en ciertas entidades, son más coherentes que los que discuten múltiples entidades entre oraciones.

Para la componente sintáctica se considera que el análisis de coherencia gira entorno a patrones de transiciones de entidades locales que especifican cómo el foco del discurso cambia entre las oraciones. El supuesto clave es que ciertos tipos de transiciones son probables que aparezcan discursos coherentes. Para exponer estos patrones de transición de entidades, se representará un texto mediante una matriz de entidades. Las columnas de esta matriz corresponden a las entidades del discurso, mientras que las celdas corresponden a oraciones del mismo. 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 1: 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, se observa que en la oración 3 es un sujeto, en la oración 6 es un objeto, pero se encuentra ausente en el resto de las oraciones. El cálculo de estas matrices es simple si es que se tiene un parser preciso a disposición. En este caso se utiliza el parser de Collins' [39], que es un parser estadístico para identificar las entidades de discurso y sus roles gramaticales.

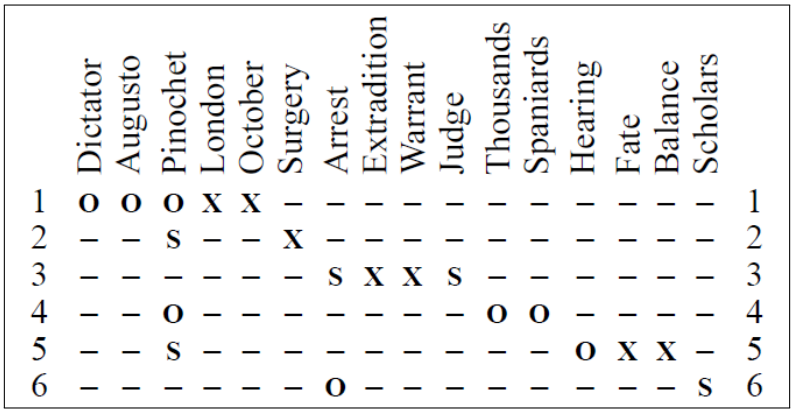

Figura 2: Una matriz de entidades

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

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

El modelo para ser entrenado requiere textos coherentes [37]. Luego, puede utilizarse para predecir P en textos nuevos. Se tendrá que $P_{coherence}(T)$ será mayor para textos que se consideren más coherentes que los que tengan un $P_{coherence}(T)$ menor.

Para la componente semántica existen se puede modelar la forma en que las oraciones y frases se enlazan. Se representa la cohesión léxica a través de un modelo *cadenas léxicas* [40], es decir, secuencias de palabras relacionadas que abarcan una unidad textual. Unidades textuales coherentes tendrán una alta concentración de estas cadenas. La premisa fundamental detrás de las cadenas léxicas es que textos coherentes contendrán una gran cantidad de palabras relacionadas semánticamente. Este supuesto permite realizar una representación que no considere la sintaxis del texto, y que no considere el orden de las palabras. Por tanto, se puede representar cada texto como una bolsa de palabras. Se puede representar cada oración como un conjunto de palabras. Luego, para medir la coherencia local de un texto se debe cuantificar la relación semántica entre oraciones adyacentes. Por lo tanto, para medira la coherencia de un texto T se tomará el promedio de las similitudes entre oraciones:

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

Se tiene que $sim(S_i, S_{i+1})$ es una medida de similaridad entre las oraciones S_i y S_{i+1}. Se utilizarán distintas medidas de similitud.

Este método fue utilizado para evaluar coherencia de resúmenes generados automáticamente por un computador, tomando como referencia resúmenes escritos por humanos[34]. Se compararon distintos métodos para evaluar coherencia textual como LSA, o métodos basados en Thesauros. Se detectó que no hay correlación entre los métodos, por lo que evalúan distintas componentes de coherencia. Luego, se propuso un ensamble de modelos para evaluar la coherencia textual, obteniendo una correlación con humanos más alta que cualquier método por sí solo.

Tomando lo anterior, un modelo que considere la sintaxis y semántica de un ensayo puede superar a cualquier modelo que utilice medidas indirectas.

3. Metodología Experimental para Validar Hipótesis

1. Se realizará una revisión bibliográfica de métodos que consideren evaluar coherencia a nivel de discurso.

2. Se recopilarán datos de ensayos evaluados por humanos, los cuales se limpiarán y prepararán para utilizarlos en el modelo propuesto. Para este procesamiento y limpieza se utilizarán herramientas existentes como por ejemplo nltk, pyCharm, entre otros. Los datos a utilizar serán los proporcionados por Kaggle[1] en la competencia *Automatic Essay Scoring*. Este conjunto de datos contiene 8 categorías diferentes de ensayo. 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).

3. Se desarrollará un método de evaluación que considere lo propuesto en la hipótesis, es decir, sintaxis y semántica del ensayo a evaluar. Para ello, se estudiarán modelos de evaluación de coherencia textual que se fundamenten en teoría de discursos. Se tomará como base el modelo propuesto en [34], el cual deberá a lo menos ser adaptado al contexto de evaluación automática de ensayos. Para ello se requerirá realizar un ensamble de modelos de evaluación automática de ensayos.

4. Se implementará un prototipo computacional para realizar las pruebas y experimentos que validen la hipótesis, y esto se realizará con herramientas existentes.

5. El modelo se evaluará y luego se comparará con otros modelos en la literatura, se utilizarán las siguientes métricas:

    - *Exact Agreement*: Se define como la proporción de ensayos que fueron calificados igualmente por el evaluador humano y la técnica computacional.
    - *Adjacent Agreement*: Es una medida que se define como el la proporción ensayos que fueron evaluados igual por el evaluador humano y la técnica computacional o que difiere en a lo más 1 punto (de calificación).
    - *Quadratic Weighted Kappa*: es una métrica de error, que mide el grado de acuerdo entre dos evaluadores.

4. 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.