UNIVERSIDAD NACIONAL DE COLOMBIA

MÉTODO PARA LA EVALUACIÓN AUTOMÁTICA DE LA ORGANIZACIÓN DE TEXTOS ARGUMENTATIVOS

FABIÁN TRINIDAD ROPERO MONTEJO

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial

Bogotá, Colombia

2014

MÉTODO PARA LA EVALUACIÓN AUTOMÁTICA DE LA ORGANIZACIÓN DE TEXTOS ARGUMENTATIVOS

FABIÁN TRINIDAD ROPERO MONTEJO

Tesis presentada como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas y Computación

Director:
Luis Fernando Niño Vásquez, Ph.D.

Grupos de Investigación:

Laboratorio de Investigación en Sistemas Inteligentes (LISI)

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial

Bogotá, Colombia

2014

Dedicada a mí esposa, Leidy,
y a mi hija, Laura Sofía.

Agradecimientos

Al profesor Luis Fernando Niño Vásquez, Director del Grupo de Investigación LISI, quien durante el desarrollo de este proyecto me brindó el mejor soporte técnico y humano.

A la profesora Julia Marlén Baquero Velásquez y a Sergio Jiménez Vargas quienes me brindaron su importante asesoría en el desarrollo de este trabajo.

A los compañeros de mi grupo de investigación, quienes me aportaron ideas de mucha ayuda.

Resumen

El uso de preguntas abiertas como herramientas para la evaluación de la educación y de competencias en general está ganando gran importancia en el contexto nacional e internacional. Sin embargo, este tipo de preguntas tiene mayores costos para su calificación que su contraparte las preguntas cerradas. Por esta razón el apoyo de herramientas computacionales a la calificación de las respuestas a las preguntas abiertas es un reto y una demanda de primera importancia. Entre los distintos tipos de pregunta abierta se encuentran los ensayos, que son tareas que solicitan a los estudiantes escribir textos de mediana longitud y cuyo fin es, principalmente, evaluar la calidad de la escritura más que los conocimientos conceptuales de los estudiantes. En esta investigación se explora mediante un caso de estudio la aplicación de un método inspirado en los *modelos de contenido* para el reconocimiento y evaluación de la organización de los textos, siendo esta última uno de los atributos del texto de mayor relación con la calidad de la escritura. Específicamente, se desarrolló un método que relaciona e integra algunas técnicas de procesamiento de lenguaje natural, de agrupamiento (*clustering*) y modelos de Markov como una solución de bajo costo, dependiente del dominio, para la evaluación automática de la organización en textos argumentativos. Los resultados obtenidos mostraron que existe una relación directa entre la calificación automática asignada por la metodología propuesta y la calificación asignada por humanos.

Palabras clave: calificación automática de ensayos, CAE, calificación de la organización de ensayos, identificación de temas no supervisada

Abstract

Using open ended questions as an education and skills evaluation tool is getting preponderance in the colombian and international scope. However these kind of questions are more costly than their counterpart, the close questions. That is why the support of computational tools for scoring responses to open ended questions is a challenge and need of great importance. One of the types of open ended questions are the essays; this is a task asking the students to write a medium size text whose purpose is to evaluate the writing skills instead of the conceptual knowledge of students. In this work, the application of a method inspired by *content models* for the recognition and evaluation of text organization is developed as a case study; since organization is one of attributes of text more highly correlated with the quality of writing. As the result of this research, a method that combines natural language processing, clustering and Markov models as a domain dependent low cost solution for automatic evaluation of argumentative text organization is articulated. The obtained results showed that there exists a relation between the automatic score given by this methodology and the score given by humans.

Keywords: automated essay scoring, AES, essay organization scoring, unsupervised topic detection

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Lista de abreviaturas

Abreviaturas

Abreviatura	Término
AES	Automatic Essay Scoring (Calificación Automática de Ensayos)
HMM	Hidden Markov Models (Modelos Ocultos de Markov)
MC	Markov Chain (Cadenzas de Markov)
NLP	Natural Languaje Processing
PLN	Procesamiento de Lenguaje Natural
GMAT	Graduate Management Admission Test
POS	Part of Speach

Introducción

Tradicionalmente las pruebas en los programas de evaluación, como los de ingreso a la universidad o los de la calidad de la educación, se han realizado mediante preguntas de selección múltiple. Esto se debe a que con ellas la calificación de las respuestas es mucho más rápida y barata, en comparación con su contraparte las preguntas abiertas. Adicionalmente, los puntajes asignados por computador este caso se consideran más confiables. Sin embargo, existen habilidades que solo pueden ser medidas adecuadamente usando respuestas libres. Para medir estas habilidades, se prefiere el uso de preguntas abiertas, las cuales piden al estudiante declarar, explicar o argumentar [1]. Algunos ejemplos de estas habilidades son las llamadas *habilidades del siglo XXI (21st-century skills)*: pensamiento crítico y solución de problemas, comunicación, colaboración, creatividad, e innovación [2]. Típicamente las preguntas de respuesta larga solicitan al estudiante escribir un texto argumentativo, por lo cual se les llama ensayos.

Por las razones mencionadas, en la actualidad se presenta un importante movimiento internacional hacia la evaluación del aprendizaje mediante preguntas abiertas. Entre las pruebas internacionales que promueven la evaluación de competencias por medio de preguntas abiertas se encuentra el *Programme for International Student Assessment (PISA)*, de la *Organisation for Economic Co-operation and Development (OECD)* [3]. Colombia no es ajeno a esta tendencia internacional y en la prueba de evaluación para los egresados de la educación superior, saber pro, se incluyen algunos ítems para medir la calidad de la escritura.

La calificación de preguntas abiertas ha sido una labor tradicionalmente humana. Los humanos realizan la evaluación con base en su entendimiento del lenguaje natural y su conocimiento previo del dominio. Sin embargo, la calificación humana presenta algunos obstáculos, especialmente cuando se aplican pruebas a gran escala, como son: costos, demanda de recursos y tiempo. Para abordar dichos obstáculos se ha recurrido a la

calificación por computador, dando origen al área: *Calificación Automatizada por Computador* (*Computer-Automated Scoring*, en inglés). Esta es considerada un área relativamente nueva de investigación, cuyos primeros desarrollos se publicaron en 1966 [10]. Los sistemas modernos de CAS, además de resolver los obstáculos mencionados de la calificación humana, tienen otras ventajas tales como: reproducibilidad, consistencia, granularidad, objetividad, fiabilidad, y eficiencia [4].

Esta investigación se enfocó en analizar la organización de los textos, por ser esta la característica que en la investigación realizada en [11] encontraron que tenía más peso en la calificación de la calidad de la escritura. Se diseñó e implementó una metodología capaz de reconocer y de evaluar la organización de textos argumentativos basándose en un corpus de ensayos bien escritos. La metodología propuesta articula técnicas de procesamiento de lenguaje natural, el algoritmo de agrupamiento no jerárquico *k*-medoids y modelos de Markov para el reconocimiento, modelamiento y evaluación de la organización de textos. El método propuesto es dependiente del contenido, por lo tanto es también dependiente del dominio; es decir, no se busca construir un modelo de aplicación general sino que se construye un modelo para cada dominio de aplicación específico. Sin embargo, debido a que su baja dependencia en recursos costos de adquirir como las anotaciones humanas o reglas complejas, el método propuesto es de menor costo económico que los que hacen uso intensivo anotaciones humanas y de mucha menor complejidad que los que hacen uso de representaciones de mayor generalidad.

Este documento se encuentra organizado de la siguiente manera: en el primer capítulo se hace el planteamiento del problema y de los objetivos de la investigación; en el segundo capítulo se describen los principales sistemas para la calificación automática de ensayos en la actualidad, en el tercer capítulo se presentan trabajos previos relacionados con el reconocimiento y análisis de la organización en textos. Estos trabajos son de especial importancia para el desarrollo de esta investigación ya que fueron fuente de inspiración y de algunos recursos utilizados. En el cuarto capítulo se describe detalladamente la metodología para la evaluación de la organización en textos argumentativos, la cual es uno de los principales resultados del presente trabajo. En el

capítulo cinco se presentan y se analizan los resultados obtenidos. Finalmente se presentan las conclusiones y recomendaciones.

1. Planteamiento del problema

Tanto a nivel de la educación superior como a nivel de la educación básica y media se presenta la necesidad de evaluar la calidad de la escritura, con el fin de cualificar las competencias de los estudiantes. Las pruebas nacionales e internacionales en que participa el país indican que existen algunas deficiencias en las competencias de lecto-escritura entre los estudiantes colombianos.

Existe por tanto la necesidad de reforzar el desarrollo de las competencias de escritura en el proceso educativo. Sin embargo, los docentes se enfrentan al problema de la alta demanda de tiempo que implica la calificación de los ensayos. Adicional a esto, la calificación de los ensayos en pruebas a gran escala presenta problemas como reproducibilidad, consistencia, granularidad, objetividad, fiabilidad, y eficiencia.

Para resolver estos inconvenientes, en países como Estados Unidos, se han desarrollado métodos y herramientas para la calificación de ensayos por medio del computador. No obstante, la calificación de los ensayos no es un problema resuelto ya que los métodos actuales no son infalibles, son de alto costo, como se explica en detalle más adelante, y no existen herramientas construidas para apoyar el proceso de calificación de ensayos específicamente para las realidades de nuestro país y de nuestras aulas de clase.

Como se describió detalladamente en los antecedentes, los expertos en evaluación de las competencias de escritura en general consideran que la organización de los ensayos es una de las características textuales conceptualmente más importantes, si no la más importante, para evaluar la calidad de escritura de un ensayo. Adicionalmente los modelos construidos automáticamente como en E-rater v2® asignan pesos altos a esta característica.

Pese a su importancia, esta es una de las características que se ha desarrollado menos en la mayoría de los sistemas de calificación automática de ensayos. Esto se debe a que actualmente no hay conjuntos de datos públicamente disponibles, tales como ensayos calificados por humanos preferiblemente con base en la organización o al menos holísticamente. Obtener dichos conjuntos de datos es costoso en tiempo y en dinero ya que implica recolectar un número suficiente de ensayos, escritos sobre un mismo conjunto de preguntas abiertas, e implica los costos de uno o varios calificadores humanos.

Enfoques actuales para AES, como E-rater v2®, en general y para calificar la organización en particular han utilizado métodos supervisados. Esto es, métodos que resuelven el problema de manera indirecta intentando aprender el criterio de evaluación seguido por los calificadores humanos, el cual es llamado *rejilla* (*rubric,* en inglés) seguido por los calificadores humanos. Usualmente la calidad de los resultados usando dicho enfoque depende de la calidad y cantidad de los datos anotados cuya consecución es un asunto problemático. Cabe resaltar que en este punto nos referimos a dos tipos de anotaciones, la primera consiste en el puntaje o calificación, la segunda se refiere a marcas dentro de los textos para delimitar cada una de las estructuras; es decir, los tipos de oraciones y párrafos que se hayan presentes. Estas últimas son aún más difíciles y costosas.

Otra de las deficiencias de los métodos implementados en los sistemas actuales es que en general se construye un modelo por cada pregunta o grupo de preguntas muy similares. Esto aplica para las diferentes características textuales como el contenido (qué tan relacionado está el ensayo con el dominio) y la organización (qué tan bien estructurado está el texto).

Por estas razones, consideramos que es importante investigar la calificación de la organización siguiendo un enfoque semi-supervisado o no supervisado. Estos métodos han sido utilizados con éxito en otras aplicaciones de procesamiento automático de texto tales como traducción automática y la clasificación de textos. Estos métodos se basan en técnicas de procesamiento estadístico como *Statistical Machine Translation*, los cuales

son muy estudiados en la actualidad ya que logran resultados iguales o mejores que las técnicas basadas en reglas, a la vez que no incurren en los costos del desarrollo manual de reglas lingüísticas.

Dado, por una parte, el alto costo para obtener ensayos anotados para construir modelos supervisados para la calificación automática de la organización de ensayos, y por otra, la gran disponibilidad de ensayos publicados sin anotaciones de calificación ni de estructura pero que han sido cuidadosamente revisados o escritos por escritores profesionales, la pregunta de investigación es: ¿un modelo de calificación que explote la información de ensayos bien escritos de manera no supervisada o semi-supervisada podrá obtener resultados comparables a los obtenidos con los métodos supervisados a un costo considerablemente menor?.

La principal contribución a realizar en el presente proyecto de investigación consiste en desarrollar un método de calificación basado en un conjunto de textos de un dominio en particular llamado *corpus* no anotado (sin requerir calificaciones ni marcas de estructura realizadas por humanos). Este sería un aporte significativo al área ya que permitiría la construcción de sistemas de calificación de bajo costo. Esto permitiría reducciones en tiempo y costo para la calificación automática de ensayos. Cabe aclarar que el alcance del presente proyecto no abarca la calificación de todas las características de los ensayos sino solo la organización.

1.1 Objetivo general y objetivos específicos

1.1.1 Objetivo general

Proponer un método para la evaluación de la organización de textos argumentativos basado en corpus.

1.1.2 Objetivos específicos

Diseñar una nueva forma de representación de la organización de textos, o implementar una representación encontrada en el estado del arte y que se ajuste a los requerimientos del prototipo.

Construir un modelo de la organización de textos argumentativos mediante la aplicación de un método de aprendizaje de máquina y el procesamiento de corpus.

Proponer un método para evaluar los ensayos del conjunto de pruebas, con base en el modelo generado.

2. Calificación automática de ensayos

Los ensayos, es decir, las preguntas de respuesta larga, por lo general no solicitan a las personas evaluadas demostrar conocimientos concretos sino narrar, describir o argumentar una tesis. Se diferencia de las preguntas de respuesta corta en que su objetivo es evaluar la calidad de la escritura mientras que el objetivo de las últimas es medir conocimientos. Por lo anterior, las preguntas abiertas pueden ser divididas en dos grandes tipos, los ensayos y las preguntas de respuesta corta. Los métodos utilizados para la calificación automática de estos dos tipos de pregunta abierta se diferencian significativamente entre sí. El presente trabajo se ubica en el área del desarrollo de métodos computacionales para la evaluación de textos largos, conocida como *Calificación Automática de Ensayos* (*Automated Essay Scoring* - AES, en inglés).

La *Calificación Automática de Ensayos* es definida como la tecnología informática que evalúa y asigna puntajes a la prosa escrita. Estos sistemas se utilizan principalmente para resolver los asuntos relacionados con la calificación de textos escritos, como son: tiempo, costo, fiabilidad y generalización [10].

Actualmente los sistemas de AES no leen y entienden los ensayos como lo hacen los humanos. Mientras que los humanos, para producir el puntaje de los ensayos, evalúan directamente variables intrínsecas de interés (*trins*), tales como, dicción, fluidez, gramática y orden, los sistemas AES utilizan medidas que se aproximan o correlacionan con las variables intrínsecas (*proxes*) [11].

En la actualidad algunos de estos sistemas son utilizados como complemento a los calificadores humanos, en lugar de sustituirlo, buscando incrementar la fiabilidad y consistencia de los puntajes; por ejemplo, e-rater es utilizado como un segundo calificador, combinado con calificadores humanos, para la evaluación de la prueba *Graduate Managment Admision Test* - GMAT [11]. Además de la calificación, un aporte significativo de algunas herramientas es una retroalimentación rápida, incluso en línea,

que apoye a los estudiantes y profesores en el aprendizaje y enseñanza de las competencias de escritura. Algunos de los sistemas descritos acá cuentan con versiones Web para la retroalimentación, como por ejemplo *Criterion* (del ETS) [11].

El escepticismo y la crítica han acompañado por años el desarrollo de los sistemas AES, relacionados con el hecho de que las máquinas no pueden entender el texto escrito. Page y Petersen identificaron tres objeciones generales a los AES: humanística, defensiva y constructiva [11].

- "Objeción humanística (*humanistic objection*): El juicio de los computadores debe ser rechazado de plano ya que nunca apreciarán o entenderán los ensayos del mismo modo que los humanos. Esta objeción es difícil de reconciliar ya que discute las bases mismas de las capacidades y limitaciones de la inteligencia artificial. Las mejoras en los sistemas, la investigación empírica, y mejores evaluaciones de los sistemas pueden ayudar a incrementar su uso. Mientras tanto algunos sistemas se utilizan en combinación con humanos para la calificación de los ensayos.
- Objeción defensiva (*defensive objection*): Los sistemas AES pueden utilizarse exitosamente solo para evaluar ensayos de buena fe. Estudiantes que escriban ensayos de mala fe pueden engañar el sistema. Se requieren más estudios para delimitar las capacidades de los AES en este sentido.
- Objeción constructiva (*construct objection*): Los *proxes* medidos por el computador no son lo que realmente importa en un ensayo. En respuesta a esto, la capacidad de algunos sistemas (como Criterion) de proveer retroalimentación específica se presenta como evidencia de la disminución de la brecha entre los *trins* y los *proxes*" [11].

Debido a estas objeciones el desarrollo de los sistemas AES ha estado acompañado de múltiple esfuerzos por demostrar su validez. En los últimos años se ha buscado que tanto los sistemas AES como los métodos de validación sean conceptualmente más significativos, es decir, más cercanos a los fundamentos aplicados por los humanos para la calificación de la calidad de escritura, con el fin de aumentar su aceptación [11]. Los enfoques actuales para la validación de los sistemas de AES se presentan a continuación.

2.1 Sistemas y enfoques representativos de la calificación automática de ensayos

Project Essay Grader™ (PEG)

Fue desarrollado en 1966 por Ellis Page, a quien se le reconoce como el iniciador de los sistemas de AES. Se basó en la correlación de medidas lingüísticas para predecir la calidad intrínseca de los ensayos [10]. Como profesor de inglés, el creía que los computadores podrían proveer rápidamente retroalimentación a los estudiantes para mejorar su escritura [13].

Page aprovechó las capacidades computacionales de su tiempo para hacer un análisis estadístico de las características superficiales de la escritura. Aplicando regresión lineal múltiple sobre las características textuales (lingüísticas), descubrió que algunas de ellas tenían la capacidad de predecir el puntaje asignado por humanos tales como la longitud de las palabras, el número de comas, el número de preposiciones y el número de palabras no comunes. Page creyó que las características del texto extraídas por el computador aproximaban las características intrínsecas valoradas por los humanos, de modo que nombró a las primeras *proxes* y a las últimas *trins* [12]. Un aspecto muy importante de este enfoque es que no se requiere y no pretende entender el significado del contenido del ensayo para realizar la calificación [4].

El enfoque de Page ha sido criticado por ignorar los aspectos semánticos de los ensayos y enfocarse en las estructuras superficiales. Ya que no detecta las características relacionadas con el contenido, no provee retroalimentación para instruir a los estudiantes. La primera versión era débil para predecir el puntaje y podía ser engañada con solo escribir ensayos más largos. Sin embargo, en la década de los 90 se modificó el sistema para agregarle nuevas esquemas de clasificación [10].

Intelligent Essay Assessor™ (IEA)

Fue desarrollado en 2000 por Thomas K. Landauer, Darrell Laham, Peter W. Foltz, y otros. Para el análisis y puntuación de los textos utiliza la técnica llamada *Análisis de*

Semántica Latente (*Latent Semantic Analisys – LSA*, en inglés), desarrollada por los mismos autores. IEA™ fue adquirido por *Pearson Knowledge Analysis Technologies* y actualmente se ofrece como un servicio comercial para la calificación de pruebas [10].

En LSA el texto es representado como una matriz. Cada fila representa a un término y cada columna representa el contexto (o documento). Cada celda contiene el valor de frecuencia del término u otro valor basado en la frecuencia. Cada celda denota el grado en que el término transmite información en el dominio del discurso [10].

Para construir el modelo, el sistema procesa textos sobre el dominio de la pregunta, con lo cual establece el espacio semántico. Luego aprende mediante el procesamiento de un conjunto de ensayos calificados por humanos (datos de entrenamiento). Finalmente se aplica el modelo para calificar nuevos ensayos (datos de prueba). IEA compara la similitud de contenido entre los ensayos, por ello se dice que su énfasis principal es el contenido.

IEA no calcula el puntaje basándose solo en el contenido sino que también califica la gramática, el estilo y la mecánica. Además de asignar un puntaje a los ensayos provee retroalimentación a los estudiantes sobre estas características [12].

E-rater V2 ®

Desarrollado por la empresa *Educational Testing Service (ETS)*. Es utilizado en combinación con calificadores humanos para la puntuación del GMAT AWA [10].

En su primera versión, el enfoque del e-rater era la medición de una base de características lingüísticas para luego determinar por medio de regresión paso a paso los pesos asignados a dichas características y seleccionar las que correlacionan con la calificación asignada por humanos. Este es llamado por Bent-Simon y Elliot como enfoque empírico por fuerza bruta (brute-empirical, en inglés) [14]. Sin embargo, en la versión 2 se cambia el enfoque significativamente hacia uno híbrido que se basa en el control crítico del modelo (*judgmental control*, en inglés) [11]. Es decir, que el modelo ya

no es generado automáticamente por el computador sino que interviene el humano. La combinación humano-computador se puede realizar de dos maneras: una, el humano define el conjunto de características a evaluar y el computador asigna automáticamente los pesos; o dos, el humano, además de definir las características, ajusta los pesos asignados por el computador de acuerdo a su criterio de cuales son la variables que conceptualmente deberían tener mayor incidencia en el puntaje [14]. En [14] se presenta un análisis detallado de los experimentos realizados con estos tres enfoques.

Desde sus inicios, los sistemas AES se basaron en un gran número de características que correlacionaban con la calificación humana pero que no tenían una relación intuitiva con las dimensiones de la calidad de escritura. En e-rater v2 se hace un cambio importante de este enfoque hacia uno más significativo conceptualmente, lo que se espera que contribuya a aumentar la validez del sistema. El nuevo enfoque en e-rater v2 es utilizar un conjunto pequeño de características directamente relacionadas con la calidad de la escritura y, por tanto, conceptualmente significativas [11].

El conjunto de características en e-rater v2 se basa en las características sobre las cuales se hace la retroalimentación a los estudiantes en Criterion. Estas son medidas sobre la gramática, uso, mecánica, estilo, organización, desarrollo, complejidad léxica, y el uso del vocabulario específico del dominio de la pregunta [11].

El puntaje en e-rater se calcula como el promedio ponderado de los valores estandarizados de las características, luego de aplicar una transformación lineal para alcanzar una escala deseada [11]. Parte de la definición del modelo es establecer los pesos para cada una de las características.

IntelliMetric™

Este sistema fue desarrollado por la empresa *Vantage Learning*. Se dice que fue el primero que se basó en inteligencia artificial. Adicionalmente utiliza procesamiento de lenguaje natural al igual que E-rater [10]. Es utilizado para calificar la sección *Analytic Writing Assessment*, del examen GMAT [4].

IntelliMetric™ utiliza el motor de aprendizaje de máquina *Quantum Reasoning*™ para inferir la rejilla que aplicaron los humanos para la asignación de los puntajes. En otras palabras, busca reconocer las características que valoraron los humanos para asignar la calificación. Para el entrenamiento del sistema se utiliza un conjunto de ensayos calificados (anotados) por humanos. Por medio de una herramienta de PLN *CogniSearch*™ el sistema procesa los ensayos de entrenamiento para extraer un conjunto de más de 300 características lingüísticas, que constituyen los datos de entrada para el módulo de aprendizaje de máquina. Dichas características lingüísticas se clasifican en cinco categorías: 1) foco y coherencia, 2) organización, 3) desarrollo y elaboración, 4) estructura de las oraciones, 5) mecánica y convenciones [10].

Este sistema asigna una calificación holística pero también brinda retroalimentación a los estudiantes sobre varias dimensiones del ensayo tales como gramática, uso, ortografía y convenciones [10].

Bayesian Essay Test Scoring sYstem (BETSY™)

Este sistema fue desarrollado por Lawrence M. Rudner y TahungLiang en la Universidad de Maryland [15]. No se trata de un sistema comercial como los anteriores, sino una herramienta de investigación [10]. Se basa en la suposición de la independencia condicional de los términos (por ejemplo, las palabras) para encontrar la clase que es más probable para los ensayos; es decir, asume la calificación de los ensayos como un problema de clasificación de textos y para ello utiliza el clasificador bayesiano ingenuo.

Este sistema permite aplicar dos modelos de clasificador bayesiano: el modelo multivariado de Bernoulli y el modelo multinomial. La diferencia entre los modelos radica en que el modelo de Bernoulli representa los documentos mediante un vector de variables booleanas que indican si los términos está presente o no en el documento, mientras que el modelo multinomial no tiene en cuenta la ausencia de los términos sino que representa cada documento como un vector de las frecuencias de sus términos [15].

Para el entrenamiento del sistema se utiliza un conjunto de ensayos, calificados por humanos. El sistema extrae las características y, dependiendo del modelo aplicado, computa las probabilidades a posteriori que cada término aporta para que cada una de las clases sea la clase del ensayo [15].

Si se realizara el cómputo de las probabilidades sobre todos los términos del vocabulario de los ensayos de entrenamiento, la dimensionalidad (tamaño de los vectores) del modelo bayesiano resultante sería igual al tamaño del vocabulario. Por esta razón, el sistema aplica un algoritmo de selección de características, como *entropía*, para reducir la dimensionalidad a una cantidad determinada de los términos que hacen el mayor aporte a la calificación [15]. La reducción de la dimensionalidad mejora el desempeño, especialmente del modelo Bernoulli.

Para la calificación de un ensayo, el sistema extrae los términos del ensayo que también están presentes en el vocabulario extraído en la fase de entrenamiento. Luego se computa por cada clase la probabilidad que aportan los términos extraídos para que dicha clase sea la correcta. Al final se selecciona la clase que obtenga el mayor valor de probabilidad. En los experimentos realizados por los autores del sistema reportaron que alcanzaron una precisión del 80% [15].

LightSIDE

Es una herramienta de código abierto de aprendizaje de máquina para texto. Ofrece las capacidades básicas de procesamiento de lenguaje natural como lematización, remover stop words, extraer unigramas y bigramas y asignar etiquetas POS. Por medio de plugins el usuario puede ampliar estas características incorporadas en la herramienta. Ofrece métodos supervisados de aprendizaje de máquina, más específicamente los clasificadores que hacen parte de Weka, como naive Bayes y máquinas de soporte vectorial lineales. Está diseñada como una herramienta de propósito general para el procesamiento de lenguaje natural. Se propone ofrecer a usuarios no expertos una interfaz que les permita utilizar algoritmos de aprendizaje de máquina para distintas tareas de PNL, En el caso específico de aplicación en calificación automática de

ensayos, requiere de un conjunto de datos etiquetados, es decir con una calificación asignada por humanos. Los clasificadores tratan de reconocer los patrones que relacionan las características con las etiquetas asignadas. Se puede concluir que esta herramienta permite construir clasificadores basados en el contenido y dependientes del dominio pero que no tiene en cuenta otras características lingüísticas como gramática, organización, coherencia. Ha sido utilizada en estudios en que se explora una evaluación holística y en otros donde se evalúa la presencia de ciertos conceptos clave en las respuestas de los estudiantes. Como medida de desempeño en estos estudios utilizan los coeficientes de correlación de Pearson o el coeficiente de concordancia Kappa, y afirman que el desempeño del sistema es comparable con el de los humanos [15].

A continuación se presenta un resumen de los sistemas descritos en este capítulo. Las filas uno a la cinco fueron tomadas de [10].

Tabla 2-1: Comparación de sistemas de AES.

Sistema de AES	Desarrollador	Técnica	Enfoque Principal	Aplicación para Instrucción	Número de Ensayos para Entrenamiento
PEG™	Page	Statistical	Style	N/A	100-400
IEA™	Pearson KnowledgeAnalysis Technologies	LSA	Content	N/A	100-300
E-rater®	ETS	NLP	Style and content	CriterionSM	465
IntelliMetric™	Vantage Learning	NLP	Style and content	MY Access!®	300
BETSY™	Rudner	Bayesian text classification	Style and content	N/A	1000
LightSIDE	Mayfield and Rosé	Machine Learning Classifiers	Content	N/A	500

2.2 Métodos Actuales de Validación de Sistemas AES

Como es natural, la confiabilidad de los puntajes asignados por los sistemas AES se debe demostrar a través de métodos de validación estándar. Antes de revisar las prácticas actuales de validación es necesario enunciar el significado de los puntajes

asignados. El puntaje de los ensayos generalmente corresponde a un valor en una escala discreta. Una de las escalas más comúnmente utilizada por los evaluadores es la escala 1-6 donde 1 es el nivel más bajo y 6 el nivel más alto. La cantidad de niveles de la escala es definida por los evaluadores bajo el criterio de buscar la mejor discriminación de la población. De esta forma, los diferentes enfoques de AES y de validación tienen en común el mismo escenario de asignación de puntajes.

Yongwei Yang et al [4] agruparon las prácticas actuales de validación de sistemas de *CAS* en tres tipos o enfoques: el primer enfoque se basa en la relación entre puntajes asignados al mismo instrumento por diferentes calificadores; el segundo enfoque se basa en la relación entre el puntaje y medidas externas; y el tercer enfoque se basa en el proceso de calificación. A continuación se presenta una descripción de dichos enfoques:

Relación entre puntajes asignados al mismo instrumento por diferentes calificadores.

En este enfoque la calificación dada por humanos se considera como el estándar de oro (gold estándar, en inglés) para el sistema. Una manera de medir la relación entre el puntaje asignado por el sistema y el puntaje asignado por un calificador humano es el porcentaje de acuerdo exacto (*percent of exact agreement*, en inglés); el cual consiste en dividir la cantidad de casos en que el puntaje es el mismo por el total de ensayos calificados. Como el acuerdo exacto es difícil de alcanzar, otra medida utilizada es el porcentaje de acuerdo adyacente, en el cual no solo tiene en cuenta los casos en que se asigna el mismo puntaje sino también los casos en que la diferencia entre los puntajes es de 1 punto (suponiendo una escala discreta). Se considera que el puntaje del sistema es válido cuando se obtiene una medida de acuerdo alta. Otra validación realizada es demostrar que el acuerdo entre el puntaje del sistema y el puntaje de un calificador humano es comparable con la medida de acuerdo entre dos calificadores humanos. [4]
Otra medida de acuerdo utilizada es el coeficiente de acuerdo Kappa de Cohen. Esta se considera una medida más robusta ya que realiza ajustes para tener en cuenta el acuerdo por azar entre los calificadores [4].

Relación entre el puntaje y medidas externas.

Los enfoques basados en la relación entre los puntajes asignados por diferentes calificadores al mismo instrumento abordan la consistencia y precisión del puntaje asignado por el sistema. Para abordar otros aspectos de la validez de dicho puntaje se ha utilizado la comparación con una medida externa [4]. Algunas de las medidas externas utilizadas son: los resultados obtenidos en pruebas de escritura con preguntas de selección múltiple y el puntaje obtenido por los mismos estudiantes en respuestas (ensayos) a otras preguntas [11].

Enfoques basados en el proceso de calificación.

Como se explicó anteriormente, hasta ahora la mayoría de los AES utilizan modelos matemáticos para aproximarse a las características intrínsecas de la calidad de escritura y predecir la puntuación del calificador humano. Algunos utilizan cientos de características lingüísticas y combinaciones de ellas. Por tal motivo, algunos enfoques de validación se basan en comprender el significado del modelo utilizado para la calificación. Es decir, no basta con que el modelo prediga con algún nivel de exactitud la calificación holística dada por el humano, sino que es necesario que el modelo utilizado sea conceptualmente significativo, para asegurar la validez del mismo. [4]

3. Métodos computacionales para la detección y análisis de la organización en textos

En este capítulo se describen trabajos previos que están estrechamente relacionados con métodos computacionales para la detección y análisis de patrones de organización de los textos. Estos trabajos fueron los principales inspiradores de la metodología desarrollada en la presente investigación para la evaluación automática de la organización en textos argumentativos.

3.1 Calificación de la organización

La organización se refiere a la estructura de los ensayos. Se considera que un ensayo está bien organizado si introduce un asunto, argumenta una posición y concluye [16]. La organización se puede analizar a nivel del tipo y orden de párrafos presentes en el texto, por ejemplo: introducción, cuerpo, conclusión o refutación; y a nivel del tipo y orden de las oraciones presentes en cada párrafo, por ejemplo: tesis, idea principal, transición, soporte, conclusión, refutación y sugerencia, entre otros [16].

A diferencia de otras dimensiones que forman parte de un sistema de AES, para la organización hasta la fecha se han realizado pocos modelos, debido en parte a la no disponibilidad pública de ensayos calificados con base en esta dimensión [16]. No obstante, en la bibliografía revisada hay consenso en que la organización es una de las dimensiones más importantes para evaluar la calidad de la escritura; es decir, si un ensayo está bien organizado lo más probable es que obtenga un puntaje alto. Una de las razones principales para que un ensayo bien organizado no obtenga un puntaje alto sería que su contenido no estuviese relacionado con el tema (dominio) de la pregunta.

La importancia de la organización puede verse en experimentos realizados con e-rater v2 ®, para comparar diferentes modelos de calificación sobre 10 exámenes diferentes. Por cada examen se construyó un modelo que consistía en los pesos, como porcentajes del puntaje final, asignados a un conjunto fijo de características. En algunos modelos los pesos fueron asignados automáticamente y en otros fueron ajustados por humanos. Una de las características a la que se le dio mayor peso fue la organización, de manera consistente en todos los modelos, tanto en los generados automáticamente, como en los ajustados por humanos [11].

El primer referente e inspiración sobre el que se basa el presente trabajo es la publicación de Persing et al [16]. En esta investigación se utilizaron etiquetas para designar los tipos de párrafo y los tipos de oraciones dentro de los párrafos. Se utilizaron reglas con bases lingüísticas para etiquetar tanto los párrafos como las oraciones. Cada ensayo fue representado como dos secuencias; una secuencia con las etiquetas de los párrafos y otra secuencia con las etiquetas de las oraciones. Para calificar la organización; es decir, para predecir la calificación dada por el humano, se aplicaron métodos de bioinformática y aprendizaje de máquina como alineación de secuencias, kernel *string*, y el kernel de alineación (*alignment kernel*, en inglés).

3.2 Modelos del Contenido

El segundo referente es el enfoque llamado modelos de contenido (*content models*, inglés), descrito en [18]. A diferencia de trabajos previos que presentan métodos independientes del dominio como ítems de esquema (*schema ítems*, en inglés) [24] o relaciones retóricas (*rhetorical relations*, en inglés) [25], este es un enfoque dependiente del dominio ya que se basa directamente en el contenido del texto y no en características más generales, pero a diferencia de los primeros, no requiere de costosas anotaciones humanas para la generación de los modelos del contenido sino que estos se aprenden automáticamente de manera no supervisada, siendo más baratos.

Las autoras plantean la hipótesis de que los modelos de contenido no están limitados a las aplicaciones específicas de generación automática de resúmenes y ordenamiento de información con las cuales presentadas su trabajo; sino que se pueden utilizar de manera general para la representación de la estructura de los textos.

La estructura de los textos se plantea en términos de los temas que se abordan y el orden en que aparecen. Los temas se detectan automáticamente analizando patrones de distribución de las palabras, tomando como principio de base que, para un dominio dado, los distintos tipos de texto se caracterizan por distintos tipos de patrones de ocurrencia de palabras [19]. Se basan también en la suposición de que todos los textos del mismo dominio son generados por un mismo modelo de contenido.

Como herramienta computacional para la representación del modelo del contenido de un conjunto de textos utilizan modelos ocultos de Markov, donde los estados corresponden a los temas y las transiciones de estado corresponden a los cambios de tema que se dan dentro de los documentos. Para la detección automática de los temas, hacen una agrupación inicial de los fragmentos de textos mediante un algoritmo de agrupamiento jerárquico de enlace completo (*complete-link*, en inglés), utilizando el coseno como con medida de distancia. Luego aplican de manera iterativa el algoritmo de Virterbi para reasignar los fragmentos de texto al grupo más probable al que deberían pertenecer. Las iteraciones terminan cuando se estabiliza el modelo, es decir, cuando no se presentan nuevas reasignaciones de los fragmentos de texto.

Para validar el modelo realizan dos estudios mediante la aplicación de este modelo en dos tareas: ordenamiento de información (*information ordering*, en inglés) y generación de resúmenes extractiva (*extractive summarization*, en inglés). Concluyen que los resultados obtenidos por este modelo sobrepasan los enfoques del estado del arte en ese momento.

Los trabajos previos que se describieron en este capítulo fueron los de mayor influencia en el desarrollo de esta investigación. Como se explicará en el capítulo 3, del trabajo en

[16] se tomó la idea de adquirir el corpus ICLE para obtener el conjunto de datos sobre el cual se desarrolla el trabajo, y se hizo uso de las calificaciones asignadas por humanos a un subconjunto de los ensayos de este corpus. Por otro lado, de [18] se tomaron las ideas fundamentales para este trabajo como son: i) representar la organización de los textos mediante un modelo de Markov, si bien en el trabajo original utilizaron modelos ocultos de Markov (HMM, por su sigla en inglés), en el presente trabajo se utilizaron cadenas de Markov para el mismo fin; y ii) implementar un método no supervisado para el reconocimiento de los temas abordados en los textos; si bien en el trabajo original utilizaron un algoritmo de agrupamiento jerárquico, en el presente trabajo se utilizó un algoritmo de agrupamiento no jerárquico.

4. Metodología propuesta para la evaluación automática de textos argumentativos

Proponer una metodología para la evaluación automática de la organización de textos argumentativos es uno de los objetivos fundamentales del presente trabajo. El desarrollo realizado permitió estructurar un método basado en enfoques del procesamiento del lenguaje natural y del aprendizaje de máquina que tienen una sólida relación intuitiva con el objetivo a lograr; esto es, la representación de un modelo de la organización y el uso de este modelo para asignar a ensayos nuevos una etiqueta que representa un nivel dentro de una escala de calidad de la organización.

La metodología que se propone es dependiente del dominio, debido a que se fundamenta en un enfoque basado en el contenido; esto es, se requiere construir un modelo para cada dominio que en que se desee aplicar. Para el caso de la calificación automática de ensayos, este dominio consiste en cada una de las tareas sobre las cuales se solicita a las personas (por ejemplo, estudiantes) escribir sus ensayos. La falta de generalidad de la que adolecen estos métodos es compensada por el hecho de que son mucho más baratos de construir que los enfoques con mayor generalidad conocidos a la fecha; igualmente son mucho más baratos computacionalmente que los métodos que impliquen el procesamiento de complejas representaciones simbólicas de la organización, como los árboles de relaciones.

A continuación se hace una breve explicación de las bases matemáticas de algunas de las técnicas utilizadas. Posteriormente, se describirá como se usan estas técnicas dentro de la metodología propuesta en esta investigación.

4.1 Descripción de algunas de las técnicas utilizadas

4.1.1 Peso tf-idf

En general, para que sea posible aplicar las técnicas de aprendizaje de máquina al procesamiento de lenguaje natural (PNL), los textos deben ser convertidos a una representación vectorial. En este trabajo se implementa una alternativa sencilla de representación vectorial, utilizando los unigramas enriquecidos con etiquetas parte del discurso (POS, por su sigla en inglés, *Part of Speeech*) como las dimensiones de los vectores y calculando las coordenadas de los mismos mediante el peso tf-idf. Esta es una técnica que proviene del campo llamado recuperación de información (*information retrieval*, en inglés). Una de sus fortalezas es que logra un balance entre el número de ocurrencias de un término por documento con el número de documentos en que dicho término aparece, favoreciendo los términos que mejor discriminan entre los tipos de documentos.

Uno de los factores del peso *tf-idf* es la llamada frecuencia inversa de documento (*inverse document frequecy,* en inglés), la cual se define como:

$$idf_t = log\ \frac{N}{df_t}$$

donde N es el número de documentos y df_t es el número de documentos en la colección que contienen el término t. Asigna un valor alto a los términos raros y un valor bajo a los términos muy comunes.

El peso *tf-idf* se define como:

$$tf - idf_{t,d} = tf_{td} * idf_t$$

donde tf_{td} es el número de ocurrencias del término t en el documento d.

El peso asignado por *tf-idf* es:

- muy alto cuando el término t, ocurre muchas veces en un pequeño número de documentos;
- bajo cuando t ocurre pocas veces en un documento u ocurre en muchos documentos; y
- muy bajo cuando t ocurre prácticamente en todos los documentos" [20].

4.1.2 Estimar una cantidad adecuada de grupos para k-medoids

Una parte de la metodología propuesta, como se describirá más adelante, es la agrupación de textos. Para esto en este trabajo se utiliza el algoritmo de agrupamiento no jerárquico *k-medoids*; es decir, la versión modificada de *k-means* que selecciona a uno de los elementos del grupo como su representante o medoid. Este algoritmo es una de las técnicas básicas y ampliamente conocidas del aprendizaje de máquina; por lo cual, no se considera necesario describirlo en este documento. Una tarea se suma importancia para el éxito de este algoritmo es la determinación de un valor adecuado para el parámetro k; es decir, la determinación de la cantidad de grupos que se considera apropiada. Con el fin de terminar dicho valor, en el presente trabajo se realizó una implementación de la regla de Hartigan, la cual se define como:

$$hk = \left(\frac{w_k}{w_{k+1}} - 1\right) * (n - k - 1)$$

donde n es el número total de elementos a agrupar (en este caso, el número de párrafos). Iterativamente se incrementa el valor de k hasta la primera vez que se obtenga un valor para hk menor que 10, y se toma el valor de dicha iteración como el valor adecuado. Esta regla hace parte de los enfoques basados en varianza con los cuales se ha abordado el problema de encontrar el valor adecuado para k. La intuición sobre la que se base esta fórmula se puede encontrar [22]. En este mismo trabajo se compararon distintos métodos para obtener el valor de k, con diferentes configuraciones de experimentos, encontrando que el método que obtuvo los mejores resultados para estimar un valor óptimo fue la regla de Hartigan. Con base en los resultados presentados en dicho trabajo se seleccionó este método para determinar el valor de k.

4.2 Metodología propuesta para la evaluación de la organización en textos argumentativos.

En esta propuesta se aborda el problema de asignar una calificación a la organización de los textos como un problema de clasificación. A través de la combinación de técnicas de PLN con técnicas de aprendizaje de máquina se busca la identificación de patrones presentes en el texto que permitan el entrenamiento de un modelo de la organización de los mismos y su posterior aplicación para la calificación de nuevos ensayos. Si bien el objetivo del presente trabajo es la evaluación de la organización de textos argumentativos, este enfoque podría ser generalizado a otros tipos de texto en los cuales sea posible la identificación de temas a partir de los patrones de distribución de palabras.

A continuación se describirá la metodología propuesta. En primer lugar se presenta una visión general de la misma, luego se presentan sus fases de manera detallada.

4.2.1 Descripción general de la metodología

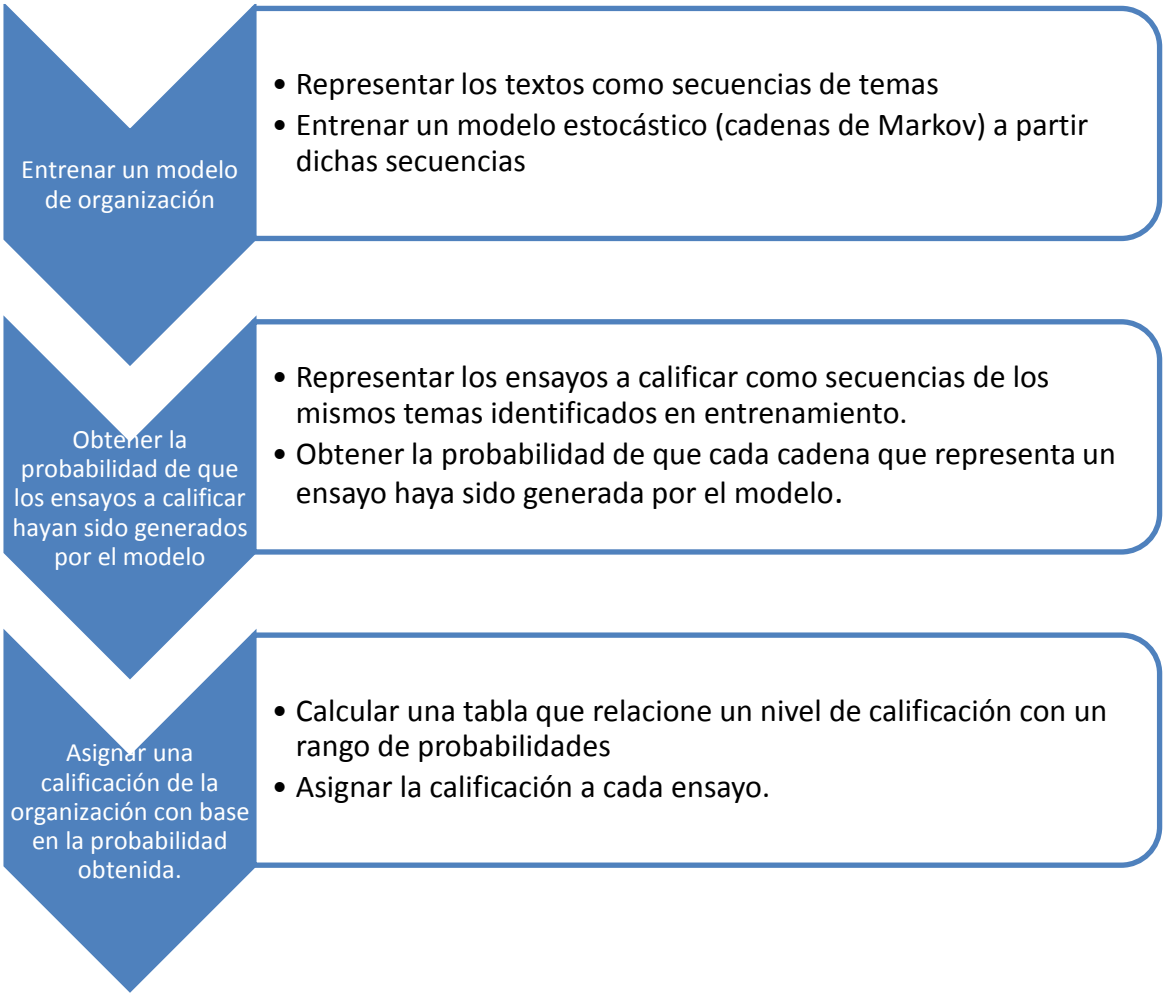
La metodología propuesta se divide en tres grandes fases: entrenar un modelo de organización, aplicar el modelo a los ensayos nuevos para obtener una medida que compare su organización con los patrones aprendidos por el modelo, y finalmente utilizar dicha medida para asignar una calificación. El modelo de la organización de los textos debe ser aprendido sobre un conjunto de ensayos que se consideren bien escritos.

Selección del conjunto de datos para el entrenamiento del modelo de organización.

A fin de probar la validez del modelo propuesto en este trabajo se utilizó una parte de los ensayos que habían recibido la máxima nota (4) por parte del calificador humano. Sin embargo, en caso en que se contara con otra forma de obtener un conjunto de ensayos bien escritos para el dominio en estudio, este método podría ser no supervisado. Debe notarse que solo para la conformación de este conjunto de datos se haría necesaria la intervención humana para asignar etiquetas. No ser requiere de calificación humana en los demás niveles de la escala. Tampoco se requiere de anotaciones para el

reconocimiento de distintos tipos de texto. Esto es realizado por la metodología de manera no supervisada.

Figura 4-1: Metodología general para la evaluación de la organización.



A continuación se describe en detalle la metodología presentada en la gráfica anterior:

Entrenar un modelo de organización.

En esta propuesta la organización de los textos es entendida como el conjunto de los temas que son tratados en los textos y el orden en que estos temas aparecen [18]. Para el entrenamiento de un modelo de la organización se hace necesario, en primer lugar, identificar los temas presentes en los textos y, en segundo lugar, el reconocimiento de

patrones en el orden en que estos temas aparecen. Para esto se requiere de un conjunto de textos que se consideren bien escritos, ya que los temas que se detecten en dichos textos y sus patrones del orden van a ser considerados como el estándar de oro para los textos a calificar.

Para la identificación de los temas se toman los párrafos de los textos como la unidad de trabajo, se construye una representación vectorial de los mismos y se aplica un algoritmo de agrupamiento no jerárquico (*k-medoids*) sobre ellos. Los grupos identificados corresponden a los temas que se tratan en los textos. Esto es posible ya que los textos del mismo tipo tienden a tener patrones de distribución de palabras similares [19].

En esta propuesta se considera a la secuencia en que se presentan los temas en los textos como un proceso estocástico. Para el reconocimiento de los patrones de orden y su representación computacional se utiliza un modelo de Markov, esto es, una cadena de Markov de orden 1.

Obtener la probabilidad de que los ensayos a calificar hayan sido generados por el modelo.

Para asignar la calificación a los ensayos nuevos, en primer lugar se obtiene la probabilidad de que su secuencia de temas haya sido generada por el modelo aprendido. Para hacer esto es necesario llevar los textos a la misma representación que se utilizó para el entrenamiento del modelo. Luego se calcula la probabilidad de las transiciones de estado que representan al ensayo, mediante el producto de las probabilidades de dichas transiciones en la cadena de Markov.

Asignar una calificación de la organización con base en la probabilidad obtenida.

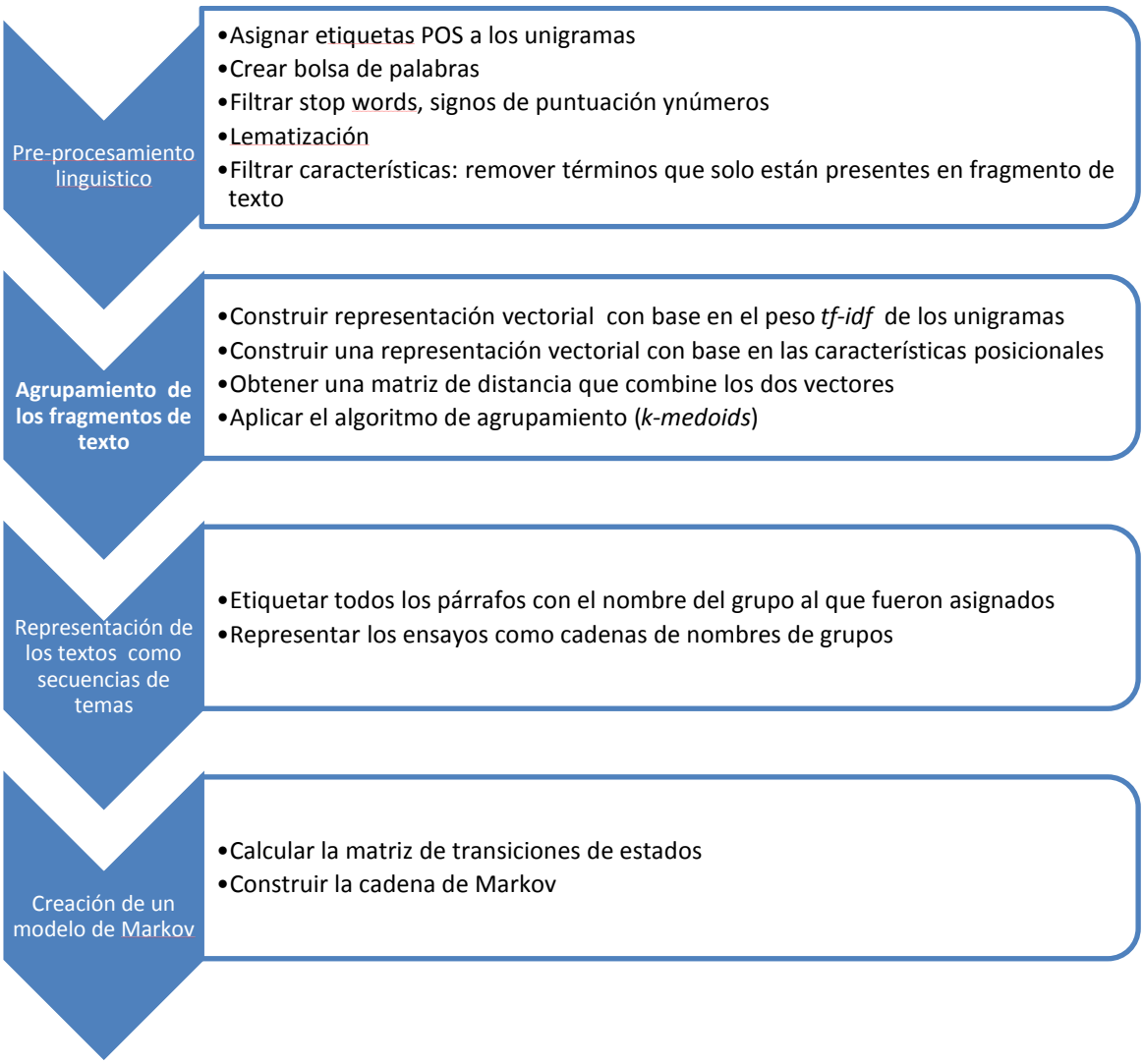
Para la asignación de la calificación a un ensayo nuevo, se le asigna al etiqueta que corresponde a un nivel en de una escala de calificación. Esta etiqueta se determina a partir del valor de la sumatoria de los logaritmos de probabilidades que se haya obtenido para el ensayo. Para esto se construye una tabla de equivalencias entre los niveles de la escala de calificación y los rangos de logaritmos de probabilidades que correspondería

para cada nivel; mediante un análisis estadístico de la distribución de los logaritmos de las probabilidades sobre los distintos niveles de la escala.

4.2.2 Entrenamiento del modelo de la organización.

A continuación se describe con mayor detalle la primera fase de la metodología propuesta para la evaluación de la organización en textos argumentativos.

Figura 4-2: Metodología para el entrenamiento del modelo de la organización.



Pre-procesamiento lingüístico.

Un paso estándar en cualquier aplicación de procesamiento de lenguaje natural es la aplicación de técnicas de lingüística computacional, a fin de depurar los datos y de conformar un conjunto de características enriquecidas que constituyan una representación adecuada de los textos de acuerdo con el problema en cuestión. En el presente trabajo se aplicaron las siguientes técnicas de lingüística computacional:

- Asignar etiquetas POS. Para esto se utilizó el modelo etiquetador opennlp framework version 1.5.2 que hace parte de KNIME [33]. Este etiquetador hace uso del conjunto de etiquetas para el idioma inglés llamado Penn Treebank [34].
- Crear bolsa de palabras.
- Remover las palabras vacías (*stop words*, en inglés), signos de puntuación, números: a fin de aplicar una primera reducción de las características menos relevantes.
- Lematización: reducir las diferentes formas de una palabra a su raíz. Para esto se utilizó el algoritmo de lematización para el idioma inglés llamado *porter* [35], el cual es provisto por la librería *Snowball Stemmer* y hace parte de *KNIME*.
- Filtrar características: remover términos que solo están presentes en fragmento de texto. Dado que estos términos no serían significativos para discriminar un tipo de texto de otro.

Como resultado de este paso se obtiene un conjunto de características consistente en una bolsa depurada de unigramas enriquecidos con etiquetas POS.

Representación vectorial de las oraciones o párrafos.

Para identificar los temas presentes en los textos lo primero que se requiere es subdividir los textos en unidades más pequeñas o fragmentos, los cuales se supone que trasmiten la información de algún tema en particular. Estos fragmentos podrían ser las oraciones, o fragmentos aún mayores como los párrafos, o menores como las clausulas. Para el presente trabajo se escogió trabajar con párrafos ya que a menor número de fragmentos de texto menor el costo computacional (tiempo de procesamiento) que toma la

agrupación de los mismos. Encontramos que los párrafos ofrecían un balance adecuado entre contener una cantidad de información asimilable a un tema y el costo de procesamiento.

Acorde con la técnica computacional seleccionada para la detección de los temas de los textos (agrupamiento no jerárquico), estos deben ser convertidos a una representación vectorial. En el presente trabajo los textos son representados mediante dos vectores, uno con base en las características léxicas (unigramas enriquecidos con etiquetas POS) y otro con base en características posicionales, es decir el lugar que ocupan los párrafos dentro del texto.

Vector de características textuales (unigramas).

Consiste en un vector cuyas dimensiones son cada uno de los unigramas enriquecidos con etiquetas POS, luego de aplicar las técnicas de filtrado y selección de características. Los valores de dichas dimensiones, es decir, las coordenadas son calculadas mediante el peso *tf-idf*.

Vector de características posicionales.

Para los seres humanos es intuitivo que el orden en que se presentan los temas en algún documento está relacionado con la posición dentro del mismo. Es decir, para un dominio en particular ciertos temas tienden a presentarse al inicio, otros hacia el final, o en el cuerpo o a lo largo de todo el texto. En [29] se realizó un análisis de la distribución de categorías o clases (similares a los temas) dentro de un corpus conformado por informes clínicos de estados de pacientes, encontrando que algunas ocurrían frecuentemente hacia el inicio, otras hacia el final y otras a lo largo de los textos. Tomando como inspiración los resultados de dicho análisis, en el presente trabajo se combinan las características léxicas de los documentos con un vector de características posicionales que enriquezcan el criterio para el reconocimiento de los temas. Las características posicionales utilizadas en este trabajo son tres indicadores: inicio, mitad y final, las cuales tendrán un valor de 1 si el párrafo se encuentra en la sección

correspondiente o 0 si lo contrario. Para determinar los límites de dichas secciones se utilizó la siguiente proporción:

- Inicio: el 10% inicial de los párrafos de un texto
- Final: el 10% final de los párrafos de un texto
- Mitad: el resto de los párrafos del texto.

Estas proporciones fueron establecidas de manera intuitiva ya que no se encontró en la literatura una regla o fórmula para definirlas.

Agrupamiento (clustering) de los fragmentos de texto.

Para la detección automática de los temas presentes en los textos se utilizó el algoritmo de agrupamiento no jerárquico *k-medoids*. La entrada para este algoritmo es una matriz de distancias de todos los vectores entre sí. Para su estimación se utilizó la distancia de coseno, la cual es comúnmente usada en los problemas de procesamiento de lenguaje natural. Como se describió anteriormente, para la representación de los párrafos se utilizaron dos tipos de vectores, uno de características léxicas y otro de características posicionales, de tal manera que la estimación de las distancias entre los párrafos debe contemplar los dos tipos de vectores. Para logar esto, se realizó un paso intermedio que consistió en calcular dos matrices de distancia: la matriz de distancias de los vectores de unigramas y la matriz de distancias de los vectores de características posicionales. Finalmente, se calculó la matriz de distancias definitiva combinando las dos anteriores, mediante la siguiente formula:

$$DISTANCIACOMBINADA = 0.8 * DISTANCIATEXTUAL + 0.2 * DISTANCIAPOSICIONAL$$

donde DISTANCIATEXTUAL es la matriz de distancias de los vectores de características léxicas (o unigramas enriquecidos con etiquetas POS) y DISTANCIAPOSICIONAL es la matriz de distancias de los vectores de características posicionales. En ambos casos se utiliza la función coseno para medir la distancia entre pares de vectores. DISTANCIACOMBINADA es la matriz que finalmente representa las distancias de los párrafos entre si y que constituye los datos de entrada para el algoritmo de agrupamiento. Con esta fórmula se busca que la distancia entre los párrafos (definida

como 1-similitud) tenga en cuenta que tan similares son en el contenido, lo cual es dado por las características léxicas que tengan en común, y que tenga en cuenta si se encuentran o no la misma posición dentro del texto. Se utiliza para ello una combinación lineal que da un peso a la distancia basada en el contenido, que para el presente trabajo fue del 80%, y un peso a la distancia basada en la posición, que para el presente trabajo fue del 20%.

Para determinar el número adecuado de grupos (valor de k), en el presente trabajo se implementó la regla de Hartigan, como se explicó en el punto 4.1.2.

Representación de los textos como secuencias de temas

Una vez se tienen todos los párrafos clasificados en grupos se procede a generar una representación de los ensayos requerida para el entrenamiento del modelo de Markov. Específicamente, un texto se puede considerar como una secuencia de párrafos. Para generar su representación, los párrafos son reemplazados por los grupos a los que pertenecen. De este modo, cada ensayo es convertido en una secuencia de símbolos, donde cada símbolo corresponde a un grupo y cada grupo a su vez representa uno de los temas que se abordan en los textos.

Creación de un modelo de Markov

A partir de las cadenas que representan a los ensayos se construye un modelo de Markov; en el cual los estados corresponden a los temas identificados y las probabilidades de las transiciones de estado corresponden a las probabilidades de pasar de un tema a otro.

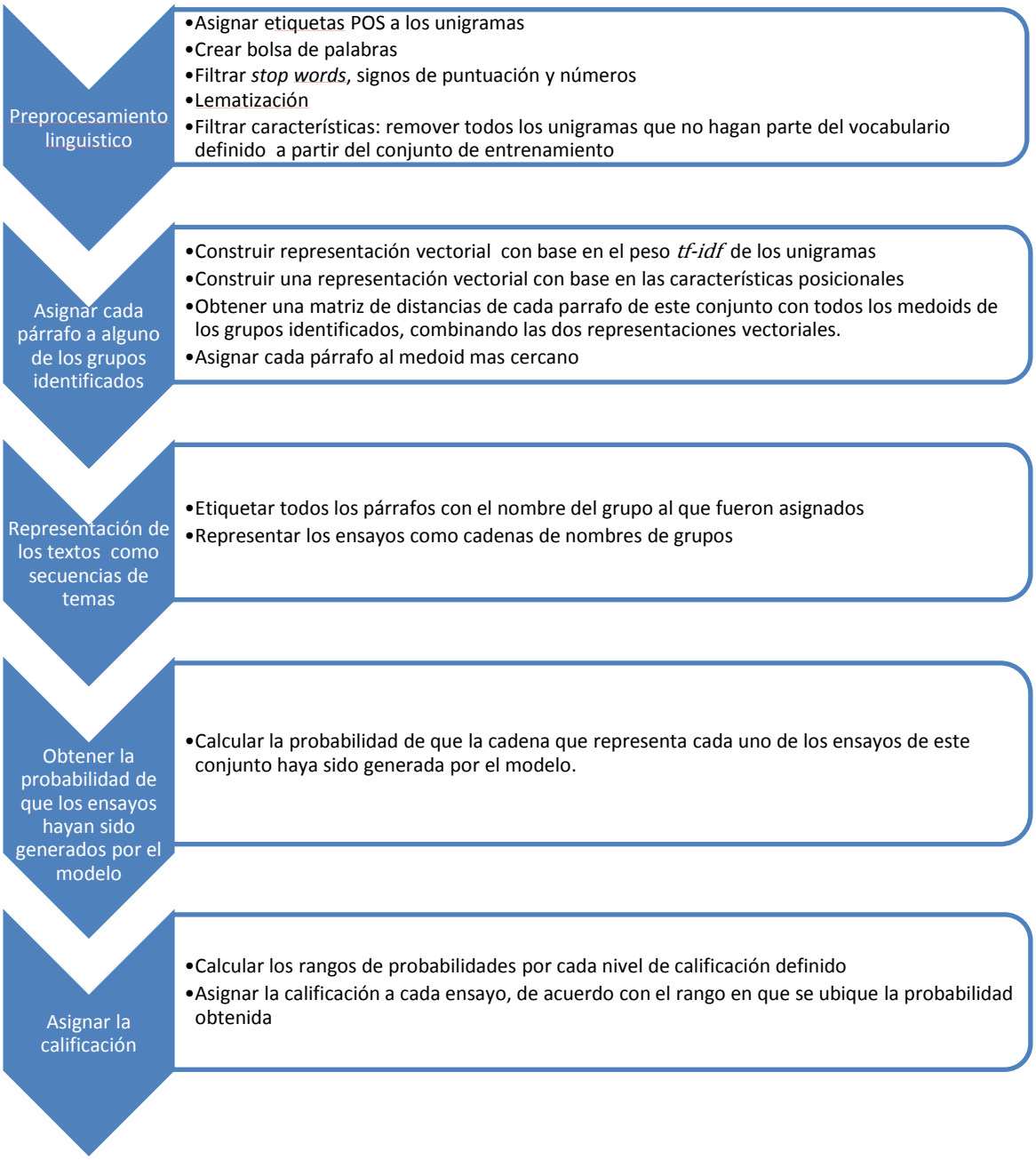
Debe notarse que si una transición de estados no hace parte de las transiciones posibles según el conjunto de datos de entrenamiento, la probabilidad asignada sería de 0. Si esta transición ocurre dentro de uno de los ensayos a evaluar, la probabilidad que asignaría el modelo a toda la cadena sería de 0. Para evitar que esto ocurra se utiliza una técnica llamada suavizado (*smoothing*, en inglés), la cual consiste en disminuir una proporción de probabilidad a las transiciones válidas para repartirla de manera uniforme entre las

transiciones no identificadas en el entrenamiento. De esta manera ninguna transición de estados en el modelo aprendido tiene probabilidad 0 de ocurrir. Existen diferentes técnicas de smoothing que se aplican para el procesamiento de lenguaje natural. En este trabajo se aplicó una de las técnicas más sencillas llamada suavizado de Laplace (*Laplacian Smoothing*, en inglés). Los detalles de esta técnica así como de los distintos métodos de suavizado pueden encontrarse en [31] y [32].

4.2.3 Evaluación de la organización

La evaluación de la organización de los textos es realizada en dos pasos. En primer lugar se aplica el modelo aprendido en la etapa anterior a los ensayos nuevos con el fin de determinar la probabilidad de generación de las secuencias que los representan. Luego se establece una tabla que relaciona cada nivel de calificación a un rango de probabilidades y se asigna una calificación a cada ensayo con base en dicha tabla.

Figura 4-3: Metodología para evaluación de la organización



Pre-procesamiento lingüístico.

Se aplican las mismas técnicas de lingüística computacional que se aplicaron para el entrenamiento del modelo. Siendo la única diferencia que el vocabulario para la representación de los ensayos a calificar se restringe al vocabulario identificado en la

fase anterior esto es, solo se tienen en cuenta los unigramas etiquetados que hagan parte de las características seleccionadas en el entrenamiento.

Asignar cada párrafo a alguno de los grupos identificados.

Los párrafos de los ensayos a calificar deben ser clasificados utilizando los grupos identificados en el entrenamiento del modelo. En primer lugar se debe construir la representación vectorial de los párrafos. Para esto se utiliza la misma representación de vector de características léxicas y vector de características posicionales utilizadas en la fase anterior. Para la construcción de la matriz de distancias, en lugar de medir las distancias de todos los párrafos entre sí, se mide la distancia de cada uno de los párrafos con todos los medoids identificados; aplicando la misma combinación lineal de las dos tipos de distancias (léxica y posicional). Finalmente se asigna cada párrafo al grupo representado por el *medoid* más cercano.

Representación de los textos como secuencias de temas.

Una vez se han agrupado todos los párrafos, cada uno de los ensayos a calificar es representado como una secuencia de los nombres de grupos a los que pertenecen sus párrafos, conservando el orden en que aparecen el texto, es decir, es representado como una secuencia de temas.

Obtener la probabilidad de que los ensayos hayan sido generados por el modelo.

Se obtiene la probabilidad de que las transiciones de temas representadas por cada cadena haya sido generada por el modelo aprendido:
Dada una secuencia de estados $S_1, S_2, \ldots, S_n$, la probabilidad de que esta secuencia de haya sido generada por el modelo se calcula como:

$$P(S_1, S_2, \ldots S_n) = \prod_{k=1}^{n} P(S_k | S_{k-1})$$

donde el S_0 corresponde al estado *inicio*.

Como este valor puede ser muy pequeño o incluso puede sobrepasar la precisión de los computadores durante su cálculo, se utiliza en su lugar la suma de logaritmos de las probabilidades, ya que esta es una función de la probabilidad, directamente correlacionada con ella y tiene la característica de ser monótona. Por lo anterior, resulta conveniente para realizar cálculos basados en la probabilidad:

$$log(P(S_1, S_2, \dots S_n,)) = \sum_{k=1}^{n} log(P(S_k|S_{k-1}))$$

Este paso de la metodología se basa en el supuesto de que a mayor similitud de la organización de los ensayos a calificar con la organización de los ensayos tomados como estándar de oro, mayor va a ser el valor obtenido por esta fórmula y viceversa.

Asignar la calificación

La asignación de las calificaciones, es decir, la asignación del nivel en la escala que le corresponde a cada ensayo, es realizada por una función de los logaritmos de probabilidad. En este trabajo se diseñó una tabla de equivalencias entre cada nivel de la escala y un rango de logaritmos de probabilidad, como implementación de dicha función.

4.2.4 Herramientas computacionales utilizadas

En este trabajo se utilizó la plataforma para el análisis de dato estadístico llamada *KNIME* [33] para la implementación de todos los pasos que hacen parte de esta metodología, excepto para la representación de las cadenas de Markov. Esta herramienta ofrece numerosos recursos para realizar procesamientos estadísticos y minería de datos tales como: implementación de técnicas de procesamiento de lenguaje natural, la representación vectorial de documentos, el cálculo de matrices de distancia, una implementación de *k-medoids*.

Como herramienta computacional para la representación del modelo de la organización se utilizó una librería para el software de procesamiento estadístico *R*, llamada *markovchain package* [30]. Esta librería es marco de trabajo que ofrece las estructuras y

métodos para la representación de cadenas de Markov así como los cálculos básicos requeridos. Se encuentra disponible bajo la licencia GPL-2 (*General Public License*).

5. Resultados y discusión

En este capítulo se presenta un caso de aplicación de la metodología propuesta mediante el cual se pudieron comprobar las hipótesis y los supuestos en que se basa el presente trabajo, estos son: la posibilidad de la identificación no supervisada de los temas de los textos, la construcción de un modelo que represente los patrones de orden en la aparición de los temas, la relación que existe entre la probabilidad de generación de las secuencias de temas de los ensayos por parte del modelo y la calificación asignada por los humanos a la organización de los mismos.

5.1 Descripción de los datos

En [16] además de que exploraron una técnica basada en una heurística basada en reglas para el modelamiento de la organización en ensayos, realizaron una contribución adicional y muy valiosa al publicar un conjunto de calificaciones asignadas por humanos específicamente a la dimensión de la organización, para un subconjunto de los ensayos del *International Corpus of Learner English (Version 2)* - ICLE [17]. Las anotaciones humanas son un recurso difícil de encontrar y costoso de construir. Por tal motivo, en el presente trabajo se utilizan las calificaciones publicadas en [16] con dos propósitos: para la selección del conjunto de ensayos bien escritos a partir del cual se construye el modelo y para la validación de los resultados al comparar la calificación automática con la calificación asignada por humanos. Adicionalmente se adquirió el corpus ICLE [17], lo cual era necesario para tener acceso a los textos de ensayos y de tareas (*prompts*) que asignadas a los estudiantes. Entre las diferentes tareas que hacen parte de este corpus, para el presente estudio se escogió la tarea para la cual fue publicado un mayor número de ensayos calificados. El texto de la tarea es: "*Some people say that in our modern world, dominated by science and technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?*".

La cantidad de ensayos calificados fue de 491, de los cuales a 42 les fue asignada la máxima nota en organización por alguno de los calificadores humanos. Una parte de estos 42 ensayos (el 60%) fue utilizada para el entrenamiento del modelo. El resto de los ensayos calificados por humanos fue utilizado para la validación de la calificación automática asignada.

5.2 Entrenamiento del modelo

El entrenamiento del modelo consta de dos partes principales: en primer lugar, el reconocimiento de los temas y la representación de los ensayos como secuencias de temas. En segundo lugar, la estimación de los parámetros de una cadena de Markov a partir de la representación de los ensayos anteriormente mencionada.

5.2.1 Agrupamiento de los párrafos

Para la construcción de un método de detección automática de los tema se dividieron los ensayos en fragmentos de texto, para este caso en párrafos. El algoritmo utilizado fue k-*medoids*. Para la estimación del número adecuado de grupos se utilizó la regla de Hartigan, la cual es una regla heurística. A continuación se presenta una tabla con los pasos dados por esto algoritmo para hallar el valor de k.

Estimación de un valor adecuado de k

Tabla 5-1: Estimación de un valor adecuado de k.

w_k	hk	k
11.251	16.937	10
10.045	14.138	11
9.123	14.328	12
8.271	11.993	13
7.609	11.233	14
7.033	11.222	15
6.497	**9.972**	**16**
6.050		

donde w_k es la sumatoria de distancias de todos los elementos del grupo a su *medoid*, hk es el valor de la heurística y k el valor del parámetro k en cada iteración. El algoritmo termina cuando hk es menor que 10.

Grupos identificados

Los 16 grupos obtenidos representan 16 temas distintos presentes en el texto a partir de los patrones de distribución de palabras. Para los datos de entrada del algoritmo se construyó la representación vectorial y el cálculo de la matriz de distancia descrita en la metodología. Los grupos obtenidos fueron:

Tabla 5-2: Grupos identificados.

Grupo	Ensayo	Párrafo	Tamaño de la partición
dream[VB(POS)]_dream[VBG(POS)]_time[NNP(POS)]_life[NN(POS)]_dream[NN(POS)]	DBAN2036	3	18
dream[VBZ(POS)]_dream[NNS(POS)]_dream[VBG(POS)]_dream[NN(POS)]_chang[VBN(POS)]	ITVE3001	4	14
dream[VBP(POS)]_dream[VBZ(POS)]_dream[NNS(POS)]_dream[VBD(POS)]_dream[VBN(POS)]	FIAB3004	2	13
modern[JJ(POS)]_domin[VBN(POS)]_world[NN(POS)]_technolog[NN(POS)]_dream[NN(POS)]	NOUO2015	1	13
nt[RB(POS)]_live[NNS(POS)]_peopl[NNS(POS)]_ca[MD(POS)]_televis[NN(POS)]	SPM07017	3	13
develop[NN(POS)]_believ[VBP(POS)]_develop[VBG(POS)]_civilis[NN(POS)]_develop[VBN(POS)]	RUMO4002	5	10
dream[NNS(POS)]_dream[VBZ(POS)]_life[NNP(POS)]_imagin[NN(POS)]_life[NN(POS)]	SPM03002	7	10
dream[NNS(POS)]_dream[NN(POS)]_human[JJ(POS)]_histori[NN(POS)]_life[NNP(POS)]	RUMO5028	1	8
centuri[NNS(POS)]_art[NN(POS)]_centuri[NN(POS)]_individu[NN(POS)]_artist[NNS(POS)]	SPM01017	8	8
imagin[JJ(POS)]_watch[VBG(POS)]_dream[NN(POS)]_dream[NNS(POS)]_watch[NN(POS)]	BGSU1108	3	8

conclus[NN(POS)]_dream[NNP(POS)]_dream[VBG(POS)]_race[NN(POS)]_dream[NN(POS)]	FRUC3009	7	7
space[NN(POS)]_fantasi[NN(POS)]_imagin[NN(POS)]_scienc[NN(POS)]_technolog[NN(POS)]	SWUL1013	5	7
play[NN(POS)]_children[NNS(POS)]_charact[NNS(POS)]_book[NN(POS)]_brain[NNS(POS)]	TSNO1431	4	7
product[NN(POS)]_advanc[NNS(POS)]_invent[NN(POS)]_imagin[NN(POS)]_product[NNS(POS)]	SPM01017	5	6
human[JJ(POS)]_stress[VB(POS)]_chang[VB(POS)]_featur[NN(POS)]_nt[RB(POS)]	RUMO5028	5	5
machin[NN(POS)]_dream[NNP(POS)]_dream[VBG(POS)]_speed[NN(POS)]_live[NNS(POS)]	BGSU1108	1	5

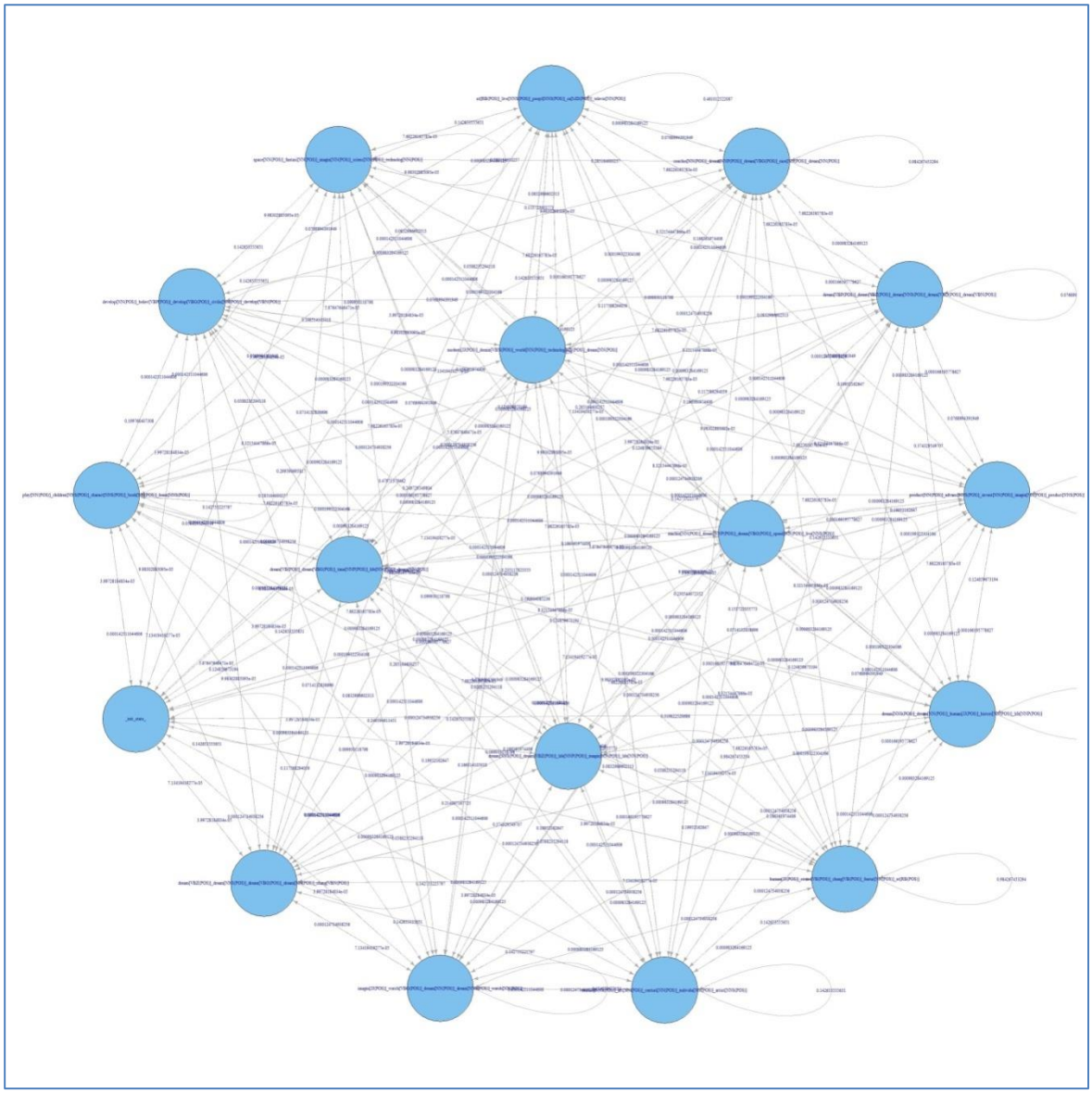
En la Tabla 5-2 se muestran los grupos identificados. Las columnas mostradas son, en su orden: el nombre del grupo, el nombre del ensayo al cual pertenece el *medoid* que representa el grupo, el número del párrafo que corresponde al *medoid* y la cantidad de elementos del grupo.

El nombre de los grupos está compuesto por los cinco términos de mayor influencia en la formación de cada grupo. La selección de estos cinco términos por grupo se realizó con base en la suma del valor del peso *tf-idf* de cada término en todos los vectores de características léxicas asignados al grupo.

5.2.2 Entrenamiento del modelo de Markov

En el siguiente paso de la metodología se genera una representación de los ensayos que consiste en la secuencia de los nombres de los grupos a que pertenecen sus párrafos, en el mismo orden en que aparecen originalmente. Es decir, en una cadena de los temas que aparecen en los ensayos. Con dicha representación como entrada se construyó una cadena de Markov, como se describió en la sección 4.2.2. El grafo correspondiente a este modelo se muestra en la Figura 5-1.

Figura 5-1: Grafo del modelo de Markov



La Figura 5-1, que muestra la estructura completa del modelo de Markov se muestra únicamente con fines ilustrativos. Los patrones que representan la organización no pueden ser vistos fácilmente en esta gráfica, ya que todos los nodos están conectados entre sí, debido a que se aplicó la técnica de suavizado de Laplace. El modelo de Markov se construyó como una herramienta para obtener los valores de probabilidad a partir de

los cuales se asigna una calificación automática, pero no para explicar por sí mismo la organización de los textos.

5.3 Evaluación de la organización

La segunda fase de la metodología consiste en aplicar el modelo entrenado para comparar la organización de los ensayos a calificar con los patrones de organización aprendidos.

5.3.1 Estimación de la probabilidad de que el ensayo haya sido generado por el modelo.

Como se explicó en la metodología, para el entrenamiento del modelo se utilizó un conjunto de ensayos que se considera bien escritos. Estos son el estándar de oro contra el cual se comparan los ensayos para su calificación en organización. El modelo de Markov entrenado permite comparar la organización de los ensayos nuevos con la organización de los ensayos del conjunto de entrenamiento. Dicha comparación consiste en obtener la probabilidad de que la secuencia de temas de los ensayos nuevos haya sido generada por el modelo entrenado. A mayor probabilidad, mayor similitud de la organización y a menor probabilidad, menor es la similitud de la organización. Por conveniencia se utilizó el logaritmo de la probabilidad, en lugar de utilizar directamente el valor de la misma, como se explicó en la sección 4.2.2.

Cálculo del logaritmo de probabilidad de generación por el modelo.

Tabla 5-3: Estadísticas descriptivas de los logaritmos de la probabilidad por cada clase, en la escala original.

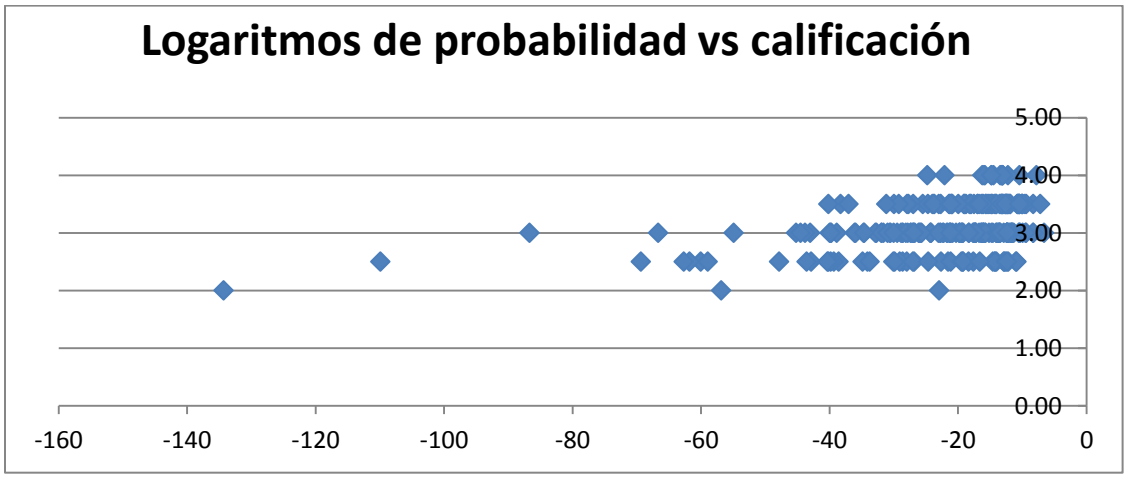
Nivel	N	Mínimo	Máximo	Promedio	Desviación Estándar
2.0	3	-134.307	-22.942	-71.358	57.087
2.5	45	-109.881	-10.863	-31.085	19.771
3.0	118	-86.671	-6.600	-22.464	11.953
3.5	86	-40.142	-7.203	-16.870	6.846
4.0	17	-24.758	-7.764	-14.598	3.930

En la Tabla 5-3 se presentan las estadísticas que describen la distribución de los valores de logaritmo de probabilidad por cada uno de los niveles de la escala de calificación original. Estos niveles se pueden considerar como las clases a las que pertenecen los ensayos. En la tabla se muestra, en su orden, la cantidad ensayos por clase (N), los valores mínimo y máximo de los logaritmos de probabilidad entre todos los ensayos de la clase, así como el promedio y la desviación estándar de dichos valores.

En ella se puede observar una notoria cercanía de los promedios de logaritmos de probabilidad entre las clases 3.5 y 4, así como entre las clases 2.5 y 3. Con base en lo anterior se infiere que cada una de estas parejas de clases se puede fusionar entre si para formar una nueva clase. Adicionalmente, las desviaciones estándar de ambas parejas presentan una relativa cercanía, por lo cual se infiere que las nuevas clases generadas mantendrían una distribución de logaritmos de probabilidad similar a las dos clases fusionadas, y que por lo tanto no se estaría haciendo una alteración de fondo a la escala. Reducir la escala de cinco a tres clases disminuye la dificultad y la complejidad de la asignación de la calificación automática dadas las características que presentan los datos del presente estudio, porque de esta manera se puede obtener una mayor diferenciación en la distribución de los logaritmos de probabilidad entre los niveles de la escala.

La Tabla 5-3 se complementa con la Figura 5-2, en la cual se muestra gráficamente la dispersión de los valores del logaritmo de probabilidad por cada clase.

Figura 5-2: Dispersión del logaritmo de probabilidad por cada clase, en la escala original



En la Figura 5-2 se puede apreciar visualmente la similitud de la distribución de los valores de logaritmo de probabilidad entre las clases 2.5 y 3 y entre las clases 3.5 y 4.

Transformación de la escala de calificación

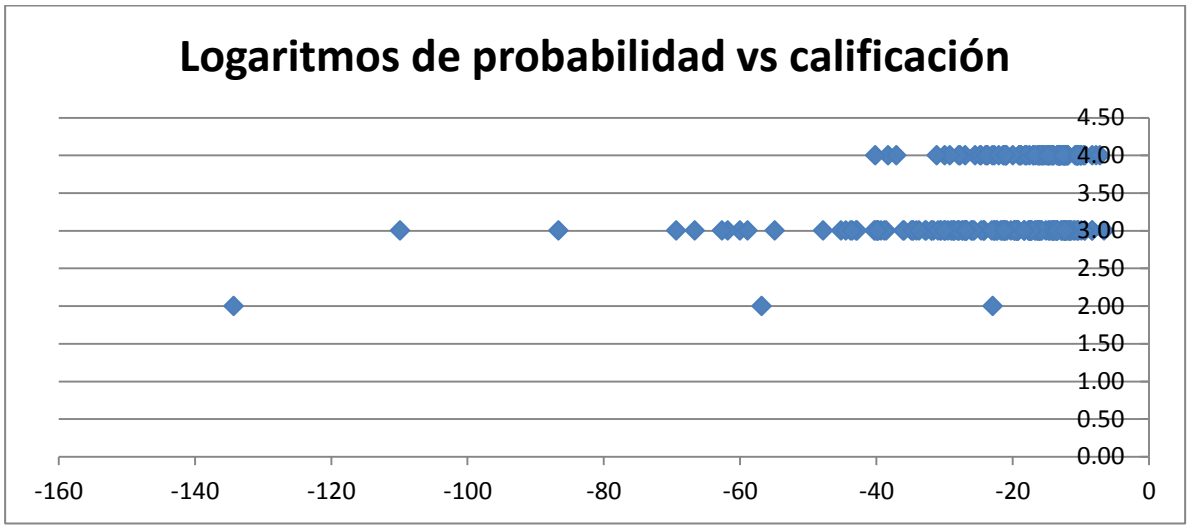
Con el fin de realizar la transformación de la escala de calificación original en una escala reducida para la asignación de la calificación automática, siguiendo el criterio descrito en el punto anterior, se estableció una nueva escala conformada por tres clases en la cual la clase 2 se mantiene igual a la escala original, la clase 3 resulta de fusionar las clases 2 y 2.5 de la escala original y la clases 4 resulta de fusionar las clases 3.5 y 4 de la escala original. Con la expresión fusionar se quiere significar que los ensayos a los que les fue asignada una clase en la escala original (por ejemplo, 2.5) se les asigna una clase en la nueva escala (por ejemplo, 3). Debe tenerse presente que estas calificaciones siguen siendo la calificación asignada por humanos, que se utilizará más adelante para validar los resultados del método de calificación automática.

Tabla 5-4: Estadísticas descriptivas de los logaritmos de probabilidad por clase en la escala transformada.

Nivel	N	Mínimo	Máximo	Promedio	Desviación Estándar
2.0	3	-134.307	-22.942	-71.358	57.087
3.0	163	-109.881	-6.600	-24.844	14.977
4.0	103	-40.142	-7.203	-16.495	6.496

En la Tabla 5-4 se muestra las estadísticas descriptivas resultantes para la escala de calificación transformada. Se puede observar que el promedio y la desviación estándar de la clase 3 es cercano a los valores para la misma clase en la escala original; igualmente que estos valores para la clase 4 se mantienen cercanos a los obtenidos para la escala original. Por lo anterior se puede afirmar que el significado de las clases en la escala transformada es similar su significado en la escala original.

Figura 5-3: Dispersión de los logaritmos de probabilidad por cada clase en la escala transformada



En la Figura 5-3 se puede observar la distribución de logaritmos de probabilidad para la escala transformada.

5.3.2 Asignación de la calificación automática de la organización de los textos

Luego de transformar la escala de las calificaciones original en una escala de tres clases, se procedió a definir una función para la asignación de la calificación automática con base en los logaritmos de probabilidad. En esta metodología no se propone que se establezca una función de aplicación universal sino que esta función deberá definirse para cada caso (tarea o *prompt*) con base en el análisis estadístico de los datos. Para el presente caso de estudio la función de asignación de la calificación automática se definió mediante el algoritmo de la Figura 5-4.

Figura 5-4: Algoritmo para la asignación de la calificación

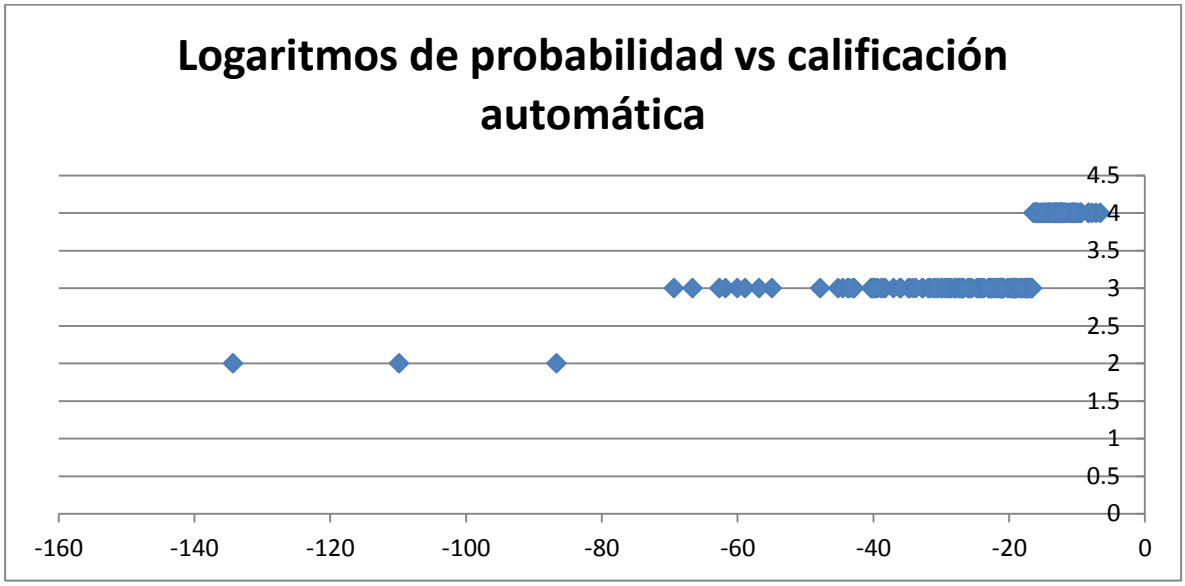
1. sea **T**el conjunto de ensayos a calificar y n el tamaño de este conjunto
2. mean_2 := media de la suma de logaritmos de probabilidad para las ensayos que tienen calificación de 2;
3. mean_4 := media de la suma de logaritmos de probabilidad para las ensayos que tienen calificación de 4;
4. **for**i := 1 **to** n **do begin**
5. **if T**[i].suma_log_probabilidad<mean_2 **then**
6. calificacion_asignada := 2
7. **Else if T**[i].suma_log_probabilidad< mean_4 **then**
8. calificacion_asignada := 3
9. **else**
10. calificacion_asignada := 4
11. **end if**
12. **end for;**

Tabla 5-5: Estadísticas descriptivas de los logaritmos de probabilidad por cada clase de la calificación automática.

Nivel	N	Mínimo	Máximo	Promedio	Desviación Estándar
2.0	3	-134.3074	-86.6715	-110.2867	23.8205
3.0	144	-69.3596	-16.6138	-28.3037	11.1121
4.0	122	-16.4007	-6.6004	-12.7544	2.1905

En la Tabla 5-5 se presentan las estadísticas descriptivas de los logaritmos de probabilidad por cada una de las calificaciones automáticas asignadas. Se puede observar que los promedios de logaritmos de probabilidad para las clases 3 y 4 de la calificación automática son similares a los valores respectivos para la calificación asignada por humanos. En los rangos de valores mínimo y máximo y en la desviación estándar se puede observar un efecto de la función de asignación de la calificación automática, ya que esta función hace que los logaritmos de probabilidad en que se basa no se puedan traslapar entre las clases.

Figura 5-5: Dispersión de logaritmos de probabilidad por nivel de calificación automática



En la Figura 5-5 se observa la dispersión de los logaritmos de probabilidad por clase que se obtienen para la calificación automática. En ella se puede observar cómo la función de asignación de la calificación hace que este valor no se traslape entre las distintas clases; es decir, mediante la calificación automática se produce una clara separación entre las clases.

5.4 Medidas de desempeño

Para validar el desempeño en este caso de estudio de la metodología propuesta para la asignación de la calificación automática de la organización se utilizaron algunos de los métodos descritos en la sección 2.2. Estos son: i) usar la calificación usada por humanos como el estándar de oro contra el cual se compara la calificación automática; ii) utilizar la correlación de Pearson y el coeficiente de acuerdo Kappa de Cohen para medir el grado de acuerdo entre la calificación automática y el calificador humano.

Matriz de confusión.

En este trabajo se abordó el problema de la calificación automática de ensayos como un problema de clasificación. Para estos casos, la generación de la matriz de confusión es un procedimiento estándar de evaluación de clasificadores. En la Tabla 5-6 se presenta la matriz de confusión obtenida.

Tabla 5-6: Matriz de confusión

		Calificación estimada		
		Nota 2	**Nota 3**	**Nota 4**
Calificación humana	**Nota 2**	1	2	0
	Nota 3	2	106	55
	Nota 4	0	36	67

En la Tabla 5-6 la diagonal representa los casos en que hubo acuerdo entre la calificación automática y el calificador humano. Se puede observar que para las clases 3 y 4, los casos en que hubo acuerdo superan ampliamente los casos en que no lo hubo. La cantidad de ensayos que pertenecen a la clase 2 es muy baja (3 en total) y eso hace difícil que se pueda obtener un porcentaje de acuerdo alto para ella.

Medidas del acuerdo de la calificación automática con la calificación humana.

Como se explicó en la sección 2.2, el porcentaje de acuerdo exacto es uno de los métodos estándar para medir el desempeño de los sistemas de AES, comparándolos con la calificación asignada por humanos. Para el presente caso de estudio este valor fue de 0,65 (174 acuerdos de los 269 ensayos calificados). Si bien el valor obtenido no es tan alto como se pudiera desear para un sistema que pueda ser comparable a los calificadores humanos, es aceptable para comprobar que existe una relación entre la calificación asignada por la metodología propuesta y la calificación asignada por el humano. Haciendo un análisis más a fondo, esto significa que la medida con la cual se compara la organización de los ensayos a calificar con la organización de los ensayos bien escritos, la cual se obtiene mediante el modelo de Markov, está relacionada con la calificación asignada por los humanos. De tal manera que esto comprueba la hipótesis que se planteó en esta investigación de que es posible construir un modelo que reconozca y evalúe la organización de textos argumentativos de manera automática, basado en un conjunto de entrenamiento conformado por ensayos bien escritos, sin hacer uso de anotaciones humanas como son las marcas de distintos tipos de estructura en los textos, o ensayos calificados en todos los niveles de la escala de calificación, o de métodos complejos como los árboles de relaciones retóricas o de reglas heurísticas basadas en conocimiento lingüístico.

Otra medida estándar del desempeño de los sistemas EAS mencionada en la sección 2.2 es el coeficiente de acuerdo Kappa de Cohen. Este coeficiente es importante porque tiene en cuenta qué tanto se debe al azar el acuerdo entre calificadores. A partir de la matriz de confusión de la Tabla 5-6 se obtuvo un coeficiente de acuerdo Kappa de 0.6455**.** Este es un valor cercano al porcentaje de acuerdo exacto obtenido. Por lo tanto se puede concluir que el acuerdo entre la calificación asignada por esta metodología con la calificación asignada por humanos no se debió al zar. El coeficiente de Kappa obtenido refuerza las conclusiones que se describieron arriba para el porcentaje de acuerdo exacto.

6. Conclusiones y recomendaciones

6.1 Conclusiones

Se alcanzó el objetivo general de desarrollar un método para la evaluación automática de textos argumentativos basados en un conjunto de textos en el mismo dominio, considerados como bien escritos. La metodología propuesta permite reconocer automáticamente temas abordados en los textos y patrones de orden de presentación de dichos temas. Es capaz también de comparar la organización de ensayos nuevos con la organización de los ensayos tomados como el estándar de oro.

Los vectores de unigramas enriquecidos con etiquetas POS y el peso *tf-idf*, son un método sencillo pero la representación los textos, que permite obtener buenos resultados en problemas de procesamiento de lenguaje natural mediante la aplicación de técnicas estadísticas, como se comprobó en esta investigación.

En esta investigación fue posible implementar un método capaz de reconocer los temas abordados en los textos mediante en reconocimiento de patrones de la distribución de palabras, sin utilizar para ello costosas anotaciones humanas o de esquemas complejos.

Los modelos estocásticos son una herramienta computacional adecuada para la representación y el análisis de los patrones de organización de los textos.

La calificación automática de ensayos mediante un método como el propuesto en este trabajo, bien sea de manera holística o considerando una característica de los ensayos como la organización, presenta aspectos difíciles de controlar como la escasez de los datos, o la falta de calidad en los ensayos mediante los cuales se evalúa la calidad de escritura de estudiantes.

La metodología propuesta fue desarrollada para el objetivo que se había planteado de evaluar la organización de textos argumentativos. Sin embargo, este enfoque para el reconocimiento, representación y evaluación de la organización podría ser aplicado para otros tipos de texto, ya que los supuestos en que se basa no están limitados solo a textos argumentativos.

6.2 Recomendaciones

Se recomienda realizar otros casos de estudio para aplicar y validar la metodología propuesta, lo cual seguramente permitiría enriquecer y refinar esta metodología. Esto no fue posible de realizar en esta investigación debido a las restricciones de tiempo para el desarrollo del proyecto y a la dificultad para encontrar conjuntos de datos apropiados.

Se recomienda evaluar el desempeño de la metodología probando otras técnicas de base; como por ejemplo, un algoritmo de agrupamiento jerárquico en lugar k-*medoids* u otros métodos gráficos probabilísticos, en lugar de las cadenas de Markov.

Se recomienda desarrollar una aplicación informática e interfaz de usuario final que articule los distintos elementos de esta metodología, como una herramienta de fácil uso para hacer análisis de la organización de los textos.

Bibliografía

[1] Klein, S. P. (2008). Characteristics of hand and machine assigned scores to college students' answers to open-ended tasks. IMS Collections, 2, 76-89.

[2] Partnership for 21st Century Skills. (2009). 21st CENTURY STUDENT OUTCOMES. Retrieved from http://www.p21.org/

[3] ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. (1999). MEASURING STUDENTS KNOWLEDGE AND SKILLS A New Framework for Assestment.

[4] Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., &Bhola, D. S. (2002). A Review of Strategies for Validating Computer-Automated Scoring. Applied Measurement in Education, 15(4), 391-412.

[5] Mohler, M., &Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09, 567-575. Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1609067.1609130

[6] Sukkarieh, J. Z., &Stoyanchev, S. (2009). Automating Model Building in c-rater. Proceedings of the 2009 Workshop on Applied Textual Inference - TextInfer '09, (August), 61. Morristown, NJ, USA: Association for Computational Linguistics.

[7] Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards Robust Computerised Marking of Free- Text Responses . Proceedings of the Sixth International Computer Assisted Assessment Conference. Loughborouh, UK

[8] Rosé, C. P., Roque, A., Bhembe, D., &Vanlehn, K. (2003). A hybrid text classification approach for analysis of student essays. Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing -, 2, 68-75. Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1118894.1118904

[9] Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2003). Auto-marking : using computational linguistics to score short , free text responses. Paper presented at the 9th Annual Conference of the International Association for Educational Assessment (pp. 1-15). Manchester, UK.

[10] Dikli, S. (2006). An Overview of Automated Scoring of Essays. The Journal of Technology, Learning, and Assessment, 5(1).

[11] Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater V.2. The Journal of Technology, Learning, and Assessment, 4(3).

[12] Wang, J., & Brown, M. S. (2007). Automated Essay Scoring Versus Human Scoring: A Comparative Study. The Journal of Technology, Learning, and Assessment, 6(2), 29.

[13] Shermis, M. D., Burstein, J., Higgins, D., &Zechner, K. (2010). Automated Essay Scoring: Writing Assessment and Instruction. International Encyclopedia of Education (3rd edition).

[14] Ben-simon, A., & Bennett, R. E. (2007). Toward More Substantively Meaningful Automated Essay Scoring. The Journal of Technology, Learning, and Assessment, 6(1).

[15] Mayfield, Elijah and Rosé, Carolyn(2013). LightSIDE: Open source machine learning for text (pp. 124-135), Handbook of automated essay evaluation; current applications and new directions.Book; New York: Routledge

[16] Persing, I., Davis, A., & Ng, V. (2010). Modeling Organization in Student Essays. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 229-239). MIT, Massachusetts, USA.

[17] Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. International Corpus of Learner English (Version 2). Presses universitaires de Louvain.

[18] Regina Barzilay and Lillian Lee. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. 2004.

[19] Regina Barzilay and NoemieElhadad. 2003. Sentence alignment for monolingual comparable corpora. In Proceedings of the 2003 conference on Empirical methods in natural language processing (EMNLP '03). Association for Computational Linguistics, Stroudsburg, PA, USA, 25-32. DOI=10.3115/1119355.1119359 http://dx.doi.org/10.3115/1119355.1119359.

[20] Christopher D. Manning, PrabhakarRaghavan, and HinrichSchütze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

[21] HARTIGAN, J. A. (1975), Clustering Algorithms, New York: J. Wiley & Sons.

[22] Mark Ming-Tso Chiang and Boris Mirkin. 2010. Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. J. Classif. 27, 1 (March 2010), 3-40. DOI=10.1007/s00357-010-9049-5 http://dx.doi.org/10.1007/s00357-010-9049-5

[23] Shermis, Mark D. (Ed); Burstein, Jill (Ed). 2013Handbook of automated essay evaluation; current applications and new directions.Book; New York: Routledge

[24] K. R. McKeown. 1985. Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text. Cambridge University Press, Cambridge, UK.

[25] W. C. Mann, S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. TEXT, 8(3):243–281.

[26] D. Marcu. 1997. The rhetorical parsing of natural language texts. In Proceedings of the ACL/EACL, 96–103.

[27] Sauper, Christina and Regina Barzilay. Automatically Generating Wikipedia Articles: A Structure-Aware Approach. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 208–216, Suntec, Singapore, 2-7 August 2009.

[28] Z. Harris. 1982. Discourse and sublanguage. In R. Kittredge, J. Lehrberger, eds., Sublanguage: Studies of Language in Restricted Semantic Domains, 231–236. Walter de Gruyter, Berlin; New York.

[29] Pablo Duboue and Kathleen R. McKeown Empirically Estimating Order Constraints for Content Planning in Generation Proceeding of the ACL/EACL, 2001.

[30] Giorgio Alfredo Spedicato. markovchain: discrete time Markov chains made easy. 2014

[31] Christopher Manning and Hinrich Schuetze. (1999). Foundations of Statistical Natural Language Processing, The MIT Press. 322

[32] Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL '96). Association for Computational Linguistics, Stroudsburg, PA, USA, 310-318. DOI=10.3115/981863.981904 http://dx.doi.org/10.3115/981863.981904

[33] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias K\&\#246;tter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. 2009. KNIME - the Konstanz information miner: version 2.0. DOI=10.1145/1656274.1656280 http://doi.acm.org/10.1145/1656274.1656280

[34] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the penn treebank. Comput. Linguist. 19, 2 (June 1993), 313-330.

[35] M. F. Porter. 1997. An algorithm for suffix stripping. In Readings in information retrieval, Karen Sparck Jones and Peter Willett (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 313-316.