Propuesta de proyecto en procesamiento de lenguaje natural

Título	**Análisis de documentos científicos en español**
Organización/Grupo de investigación	Grupo de investigación FLAG[1] – TICsW. Departamento de Ingeniería de Sistemas y Computación. Universidad de los Andes.
Experto	Rubén Manrique. Profesor del Departamento de Ingeniería de Sistemas y Computación Uniandes. Doctor en Ingeniería Informática Universidad de los Andes.

1. Descripción.

El crecimiento sostenido de la producción científica ha incrementado la necesidad de métodos automáticos para el análisis y la organización de textos académicos. Las técnicas de PLN han permitido avances relevantes en tareas como la clasificación de documentos, la extracción de información y la generación de resúmenes científicos, principalmente para textos en inglés. Estos desarrollos apoyan procesos como la revisión de literatura, la evaluación de la producción académica y la gestión del conocimiento. Sin embargo, la mayoría de los modelos, recursos y conjuntos de datos existentes han sido diseñados para el inglés, lo que limita su aplicación en otros idiomas.

El español es una de las lenguas más utilizadas en la comunicación científica, en particular en tesis, informes técnicos y revistas regionales. A pesar de ello, los textos científicos en español han recibido una atención limitada en la investigación en PLN. La escritura académica en español presenta desafíos específicos, como una mayor variabilidad estructural entre tipos de documentos, convenciones retóricas dependientes del dominio y una escasa disponibilidad de corpus anotados. Estas características dificultan la transferencia directa de modelos entrenados en inglés y justifican el desarrollo de enfoques específicos para este idioma.

Este trabajo aborda el análisis automático de documentos científicos en español a través de dos tareas complementarias. La primera es la segmentación y clasificación retórica, cuyo objetivo es identificar la función discursiva de secciones o fragmentos de texto dentro de un documento,

[1] https://flaglab.github.io/index.html

tales como introducción, metodología, resultados o discusión. La identificación precisa de la estructura retórica constituye una base necesaria para la comprensión del contenido científico y para el desarrollo de tareas de análisis a nivel de documento.

La segunda tarea es la extracción automática de contribuciones científicas, entendidas como unidades textuales que describen los aportes originales de un trabajo, incluyendo métodos propuestos, resultados relevantes o avances conceptuales. La identificación de contribuciones es fundamental para la revisión de literatura, la evaluación académica y la organización del conocimiento científico. En los textos académicos en español, las contribuciones suelen expresarse de forma indirecta y distribuida a lo largo del documento, lo que incrementa la complejidad de esta tarea.

Al combinar la segmentación retórica con la extracción de contribuciones, este estudio analiza el impacto de la información discursiva en la identificación de los aportes científicos en documentos académicos escritos en español. El objetivo es aportar modelos, datos y evidencia empírica que contribuyan al estudio del PLN científico más allá del inglés, con un enfoque específico en el español como lengua con recursos limitados en este dominio.

Este proyecto tiene como objetivos principales:

- Caracterizar un corpus a gran escala de documentos científicos en español, extraído de CORE.
- Definir un esquema de anotación retórica y de contribuciones científicas adecuado para textos académicos en español, junto con una guía de anotación que permita evaluar la consistencia y el acuerdo entre anotadores.
- Desarrollar modelos de segmentación y clasificación retórica entrenados específicamente para textos científicos en español, utilizando arquitecturas de tamaño reducido o medio, y evaluar cuantitativamente su desempeño.
- Desarrollar modelos automáticos para la extracción de contribuciones científicas, considerando enfoques de clasificación de oraciones y detección de fragmentos textuales, y analizar su dependencia de la estructura retórica del documento.
- Evaluar el uso de modelos de lenguaje de gran escala consumidos vía API como aproximación alternativa para ambas tareas, analizando su desempeño bajo diferentes estrategias de prompting y su capacidad de generalización.
- Comparar el desempeño, costo computacional y estabilidad de resultados entre los modelos de lenguaje de gran escala y los clasificadores entrenados, considerando métricas cuantitativas.

Tarea 1: Segmentación y clasificación retórica de documentos científicos

En esta tarea, el objetivo es identificar y clasificar las unidades textuales de un documento científico en español de acuerdo con su función retórica dentro de la estructura típica de un artículo académico. A diferencia de una segmentación basada únicamente en encabezados, la tarea busca asignar etiquetas funcionales que reflejen el propósito real del contenido.

Las partes retóricas consideradas en este trabajo corresponden a las secciones más comunes en la literatura científica y se definen de la siguiente manera:

- Introducción (INTRO): Presenta el problema de investigación, su motivación, los objetivos del trabajo y, en algunos casos, una descripción general del enfoque propuesto.
- Trabajo relacionado o antecedentes (BACK): Describe el estado del arte, trabajos previos relevantes y el contexto teórico en el que se enmarca la investigación.
- Metodología (METH): Explica el diseño experimental, los métodos, modelos, datos, materiales y procedimientos utilizados para desarrollar el estudio.
- Resultados (RES): Presenta los resultados obtenidos a partir de los experimentos, análisis empíricos o evaluaciones realizadas, generalmente sin interpretación extensiva.
- Discusión (DISC): Interpreta los resultados, analiza sus implicaciones, los compara con trabajos previos y discute su relevancia.
- Contribuciones (CONTR): Identifica explícitamente los aportes del trabajo, tales como métodos propuestos, hallazgos principales o avances conceptuales. Esta etiqueta puede coexistir con otras, especialmente en introducciones y conclusiones.
- Limitaciones (LIM): Describe restricciones del enfoque, supuestos adoptados, posibles fuentes de error o aspectos que limitan la generalización de los resultados.
- Conclusiones (CONC): Resume los principales hallazgos del trabajo y presenta líneas de trabajo futuro.

Tarea 2: Extracción automática de contribuciones científicas

La segunda tarea consiste en identificar oraciones o fragmentos textuales que expresan las contribuciones científicas de un documento. Estas contribuciones suelen introducirse mediante patrones lingüísticos recurrentes, especialmente al inicio de las oraciones, aunque no se limitan a estas formas. A continuación, se presentan ejemplos representativos de inicios de oraciones que comúnmente indican la presencia de una contribución en textos académicos en español:

- Este trabajo propone…
- En este artículo se presenta…
- Se propone un nuevo enfoque para…

- Este estudio introduce…
- La principal contribución de este trabajo es…
- Se desarrolla un método que…
- Los resultados obtenidos demuestran que…
- Este trabajo aporta evidencia empírica sobre…
- Se presenta una metodología novedosa para…
- A diferencia de trabajos previos, este enfoque…

Además de estas formulaciones explícitas, las contribuciones pueden aparecer de forma implícita, integradas en la descripción de resultados o en la discusión, sin marcadores léxicos directos. Por esta razón, la tarea requiere modelos capaces de utilizar el contexto discursivo y la información retórica del documento para distinguir entre contenido descriptivo y aportes originales.

2. Conjunto de datos (*dataset*).

El presente trabajo parte de un conjunto inicial de 1.812.557 documentos científicos en español, extraídos mediante la API de CORE. Este corpus incluye artículos académicos, tesis y reportes técnicos provenientes de múltiples dominios y fuentes institucionales. A partir de este conjunto masivo, se deben construir dos datasets específicos, uno para cada tarea definida en este estudio: segmentación y clasificación retórica (Tarea 1) y extracción de contribuciones científicas (Tarea 2).

Dataset para la Tarea 1: Segmentación y clasificación retórica.

Para la Tarea 1, el grupo construirá un dataset balanceado que contenga al menos 2000 ejemplos por cada etiqueta retórica considerada (INTRO, BACK, METH, RES, DISC, CONTR, LIM, CONC). Cada ejemplo corresponderá a un fragmento textual continúo extraído de un documento científico, con una longitud máxima de 1000 palabras y mínima de 250, con el fin de mantener consistencia entre muestras y facilitar el entrenamiento de los modelos.

La selección inicial de candidatos se realizará mediante una estrategia automática basada en:

- encabezados de sección y variantes léxicas comunes en español académico,
- patrones estructurales del documento (posición relativa del fragmento),
- reglas heurísticas específicas por tipo de sección.

Del total de ejemplos seleccionados para cada etiqueta, al menos el 10% será validado manualmente por miembros del grupo de investigación. La validación consistirá en confirmar o corregir la etiqueta asignada automáticamente, siguiendo una guía de anotación previamente definida por el grupo.

Las muestras validadas por humanos se utilizarán exclusivamente como conjunto de *evaluación*, garantizando que la evaluación final se realice sobre datos con alta confiabilidad. El resto de los ejemplos se empleará para entrenamiento y validación interna.

Para medir la consistencia del proceso de anotación, una fracción de las muestras validadas será anotada de forma independiente por al menos dos anotadores. El acuerdo entre anotadores deberá ser evaluada utilizando métricas estándar, tales como:

- Cohen's Kappa para anotación binaria o de etiqueta única,
- Fleiss' Kappa en los casos en que participen más de dos anotadores,
- Krippendorff's Alpha para analizar robustez frente a clases desbalanceadas.

El dataset se dividirá en conjuntos de entrenamiento, validación y evaluación, respetando estrictamente el criterio de no solapamiento a nivel de documento. En particular, no se permitirá que fragmentos provenientes del mismo documento aparezcan en diferentes particiones. Por ejemplo, si la sección metodológica de un artículo se asigna al conjunto de entrenamiento, la introducción, los resultados y cualquier otro fragmento del mismo artículo no podrán formar parte de los conjuntos de validación o prueba.

Dataset para la Tarea 2: Extracción de contribuciones científicas.

Para la Tarea 2, deberán construir un dataset que contenga al menos 1000 fragmentos que expresen explícitamente contribuciones científicas, y donde cada fragmento debe estar atado a su artículo y la ubicación dentro del mismo. No hay una longitud máxima del fragmento positivo para esta tarea. También deben buscar al menos 1000 fragmentos de entre 250 y 500 palabras en los que no se identifique ninguna formulación explícita de contribución científica.

Los fragmentos candidatos se seleccionarán automáticamente utilizando patrones lingüísticos frecuentes en el español académico, tales como expresiones introductorias de aportes ("este trabajo contribuye", "se presenta un nuevo método", "se introduce un marco comparativo" entre otras), así como información contextual derivada de la estructura retórica identificada en la Tarea 1. Debe plantear alguna estrategia inteligente para este proceso, así como de corte del fragmento para no perder información del aporte del artículo. Puede usar modelos de lenguaje si lo requiere.

De manera análoga a la Tarea 1, al menos el 10% de los fragmentos identificados automáticamente será validado manualmente por el grupo de investigación. La validación consistirá en confirmar que el fragmento seleccionado representa efectivamente una contribución científica y no una reformulación de antecedentes, metodología estándar o resultados sin novedad.

Asimismo, los fragmentos identificados como carentes de contribuciones explícitas serán revisados manualmente en una muestra representativa. Las muestras validadas por humanos se reservarán para *evaluación*, mientras que las restantes se emplearán para entrenamiento y

validación de los modelos. El acuerdo interanotador se evaluará utilizando las mismas métricas definidas para la Tarea 1, adaptadas a un esquema binario (contribución / no contribución).

3. Actividades para la Tarea 1: Segmentación y clasificación retórica

A1. Preparación y curaduría del dataset.

La primera actividad que deberán realizar consiste en la preparación y curaduría del dataset a partir del conjunto inicial de 1.812.557 documentos en español extraídos de la API de CORE. En esta etapa se filtran los documentos para retener únicamente aquellos que presentan una estructura académica reconocible, tales como artículos científicos, tesis y reportes técnicos. Posteriormente, se realiza un proceso de normalización del texto que incluye la eliminación de referencias bibliográficas, figuras, tablas y otros elementos no textuales. Los documentos deben segmentarse en unidades textuales continuas, correspondientes a secciones o bloques coherentes, y se deben aplicar heurísticas basadas en encabezados, posición relativa y patrones léxicos para asignar etiquetas retóricas preliminares. A partir de estos candidatos se construye un dataset balanceado con al menos 2000 ejemplos por etiqueta, cada uno con una longitud máxima de 1000 palabras, garantizando que las particiones de entrenamiento, validación y evaluación no compartan fragmentos provenientes del mismo documento.

A2. Validación humana y análisis de acuerdo.

Una vez construido el dataset preliminar, deberán seleccionar al menos el 10% de los ejemplos por cada etiqueta retórica para validación manual por parte del grupo de investigación. Esta validación se realiza siguiendo una metodología de anotación previamente definida por el grupo, que establece criterios claros para cada categoría retórica y proporciona ejemplos representativos. Cada ejemplo debe ser revisado por más de un anotador con el fin de medir el grado de consistencia del proceso. El acuerdo interanotador se cuantifica mediante métricas estándar como Cohen’s Kappa, Fleiss’ Kappa o Krippendorff’s Alpha, según corresponda. Las muestras validadas por humanos se reservan exclusivamente para la evaluación final de los modelos, NO para entrenamiento.

A3. Desacoplamiento entre patrones de selección y etiquetas finales.

Para mitigar el riesgo de que los modelos aprendan directamente los patrones utilizados en la construcción del dataset, deberán eliminar o neutralizan explícitamente dichos patrones. Esto incluye la supresión de expresiones que uso para encontrar las secciones (“Metodología”, “Introducción”, “Estado del Arte”, “Trabajos relacionados”, etc) o su sustitución por marcadores neutrales durante el entrenamiento, la idea es obligar al modelo a basar sus decisiones en el contenido semántico y no en las señales usadas para la construcción del dataset. *Debe entregar el dataset antes y después de este paso y ya particionado.*

A4. Desarrollar clasificadores basados en encoders entrenados para español científico.

En esta actividad deben entrenar y evalúan clasificadores supervisados basados en modelos encoder preentrenados en español científico, con especial énfasis en arquitecturas como SciBETO-large. El problema se formula como una tarea de clasificación de fragmentos textuales, considerando esquemas de etiqueta única o multi-etiqueta según lo decida el grupo (ojo esta decisión es importante). Los modelos se ajustan mediante fine-tuning sobre el conjunto de entrenamiento, explorando configuraciones relacionadas con el tamaño del contexto y la representación del fragmento. El desempeño de estos clasificadores se evalúa utilizando el conjunto de evaluación validado manualmente.

A5. Modelos de lenguaje de gran escala open-weight (1–8B).

De forma paralela, debe evaluar el uso de modelos de lenguaje de gran escala de pesos abiertos, con tamaños comprendidos entre 1 y 8 mil millones de parámetros. Estos modelos se utilizan en modo de inferencia, sin entrenamiento adicional, con el fin de reflejar escenarios de uso práctico donde no se dispone de datos anotados a gran escala. Para esta evaluación debe diseñar estrategias de prompting orientadas a la clasificación retórica, incluyendo enfoques zero-shot y few-shot, así como formatos de salida estructurados. Se analiza la sensibilidad del desempeño a factores como la longitud del fragmento, el número de ejemplos incluidos en el prompt y la formulación de las definiciones de las etiquetas.

A6. Modelos comerciales consumidos vía API.

Adicionalmente, se deben evaluar modelos comerciales de lenguaje de gran escala consumidos vía API, tales como GPT-5 o Gemini. Estos modelos se emplean bajo estrategias de prompting equivalentes a las utilizadas con los modelos open-weight, con el fin de asegurar una comparación justa. Se controlan los parámetros de generación para reducir la variabilidad de las respuestas y se evalúa su desempeño exclusivamente sobre el conjunto de evaluación validado por humanos. Esta actividad permite analizar el comportamiento de modelos de propósito general en una tarea especializada y en un idioma distinto del inglés.

A7. Evaluación comparativa de desempeño.

Los distintos enfoques deben comparse utilizando métricas estándar de clasificación, tales como Macro F1, Micro F1 y precisión por etiqueta. Además del análisis cuantitativo global, debe realizar un estudio detallado de las matrices de confusión y de los errores más frecuentes, con especial atención a las secciones retóricas que presentan mayor ambigüedad.

A8. Análisis de costo y eficiencia.

Más allá del desempeño predictivo, deben realizar un análisis del costo y la eficiencia de cada enfoque. En el caso de los clasificadores basados en encoders, se consideran los costos asociados al entrenamiento y a la inferencia. Para los modelos de lenguaje de gran escala, se estima el costo

por documento procesado, la latencia y la estabilidad de las predicciones, especialmente en el caso de los modelos consumidos vía API.

4. Actividades para la Tarea 2: Extracción automática de contribuciones científicas

A1. Selección inicial de documentos y fragmentos candidatos.

Para la Tarea 2 se construirá un dataset compuesto por al menos 1000 fragmentos que expresen explícitamente contribuciones científicas, donde cada fragmento estará asociado de forma explícita a su artículo de origen y a su ubicación dentro del documento. No se impondrá una longitud máxima para los fragmentos positivos, con el fin de preservar el contexto necesario para comprender adecuadamente el aporte científico. De manera complementaria, se identificarán al menos 1000 fragmentos negativos, con longitudes comprendidas entre 250 y 500 palabras, en los que no se identifique ninguna formulación explícita de contribución científica.

A2. Anotación humana y redefinición de etiquetas.

A partir del dataset construido en A1, al menos el 10% de los fragmentos positivos y negativos será validado manualmente por el grupo de investigación. En el caso de los fragmentos positivos, los anotadores verificarán que el texto seleccionado represente efectivamente una contribución científica y no una reformulación de antecedentes, metodología estándar o resultados sin novedad. Para los fragmentos negativos, se confirmará la ausencia de formulaciones explícitas de contribución. Estas muestras serán anotadas de manera independiente por más de un anotador, y el acuerdo interanotador se evaluará utilizando métricas estándar adaptadas a un esquema binario, tales como Cohen’s Kappa y Krippendorff’s Alpha. Las muestras validadas por humanos se reservarán exclusivamente como conjunto de evaluación, mientras que el resto del dataset se utilizará para entrenamiento y validación interna.

A3. Desacoplamiento entre patrones de selección y etiquetas finales.

Dado que los patrones lingüísticos utilizados en la selección automática pueden introducir sesgos no deseados en el aprendizaje del clasificador, se implementarán estrategias explícitas para desacoplar el mecanismo de recuperación de candidatos del proceso de entrenamiento. En particular, una fracción controlada de los fragmentos positivos deberá ser presentada al modelo con las expresiones usadas atenuadas, eliminadas o parafraseadas, nuevamente la idea es obligar al clasificador a basar sus decisiones en el contenido semántico del fragmento y no en la presencia de marcadores superficiales. De manera análoga, los fragmentos negativos en lo posible *deberán* incluirán ejemplos que contengan expresiones similares a las utilizadas en los patrones de selección (es permitido manipularlos de forma inteligente para agregar estas frases), pero que no correspondan a contribuciones científicas reales, funcionando como negativos difíciles.

A5. Entrenamiento de clasificadores basados en encoders especializados en español

Deberán entrenar clasificadores supervisados basados en modelos encoder preentrenados en español, utilizando el conjunto de entrenamiento. Deben explorar configuraciones que incorporen información contextual adicional, como la etiqueta retórica del fragmento obtenida en la Tarea 1 o la posición relativa dentro del documento. Recuerde que los datos validados manualmente solo deberán ser usados para evaluación no para el entrenamiento.

A6. Evaluación con modelos de lenguaje de gran escala open-weight

De manera complementaria, deberán evaluar el uso de modelos de lenguaje de gran escala de pesos abiertos, con tamaños comprendidos entre 1 y 8 mil millones de parámetros. Estos modelos se utilizarán en modo de inferencia, mediante estrategias de prompting diseñadas para la clasificación binaria de fragmentos. Los prompts incluirán definiciones claras de lo que constituye una contribución científica, pero evitarán ejemplos explícitos que reproduzcan los patrones utilizados en la selección automática. Para esta actividad solo utilice el conjunto de evaluación validado por humanos.

A7. Evaluación con modelos comerciales consumidos vía API

Deberá para esta tarea también evaluar modelos comerciales de lenguaje de gran escala, tales como GPT-5 o Gemini, utilizando estrategias de prompting equivalentes a las empleadas con los modelos open-weight.

A10. Análisis cuantitativo y cualitativo de errores

Realice el análisis de los resultados obtenidos utilizando métricas estándar de clasificación binaria como precisión, recall y F1, complementadas con un análisis cualitativo de errores. Este análisis se centrará en identificar patrones recurrentes de falsos positivos y falsos negativos, la influencia de la longitud del fragmento y el impacto de la información retórica. Haga un análisis de los errores cometidos en fragmentos negativos que no contienen contribuciones explícitas. Compare lo enconder tuneados vs los modelos de lenguaje pequeños y versus los comerciales.

5. Despliegue y demostrador interactivo

Como parte del proyecto deberá desarrollar un demostrador interactivo que permitirá explorar de manera integrada los resultados de la Tarea 1 (segmentación retórica) y la Tarea 2 (identificación de contribuciones científicas explícitas). El objetivo del despliegue será no solo validar los modelos en un entorno de uso realista, sino también facilitar el análisis cualitativo de sus decisiones y la comparación sistemática del comportamiento de diferentes arquitecturas de modelos frente a una misma entrada textual.

La interfaz deberá permitir al usuario seleccionar dinámicamente el modelo a utilizar entre tres categorías: (i) clasificadores basados en encoders ajustados para español académico, (ii) modelos

de lenguaje de menor tamaño utilizados como decoders en modo de inferencia, y (iii) modelos comerciales de gran escala consumidos vía API.

El sistema deberá permitir ingresar como entrada un texto completo o parcial de un artículo académico en español. A partir de esta entrada, el backend ejecutará primero el modelo correspondiente a la Tarea 1 para identificar las partes retóricas del documento. Los resultados se presentarán de forma visual, segmentando el texto por secciones y asignando a cada uno su etiqueta retórica (por ejemplo, introducción, antecedentes, metodología, resultados o discusión), utilizando códigos de color u otros recursos gráficos que faciliten la interpretación por parte del usuario.

Sobre esta segmentación, el sistema aplicará el modelo seleccionado para la Tarea 2, evaluando cada párrafo de manera independiente para determinar si contiene o no una contribución científica explícita. El resultado se mostrará junto a cada párrafo, indicando la predicción binaria y, cuando el modelo lo permita, una puntuación de confianza.