**Sistema de identificación y vinculación gen–función a partir de literatura científica** Octubre de 2025

Marian Bastidas Cruz, Catalina Criollo Castelblanco, Jonathan Andres Jimenez Trujillo,

[mbastidasc@unal.edu.co](mailto:mbastidasc@unal.edu.co) , [ccriollo@unal.edu.co](mailto:ccriollo@unal.edu.co) , [jojimenezt@unal.edu.co](mailto:jojimenezt@unal.edu.co)

Universidad Nacional de Colombia

**Resumen** - Este trabajo presenta un avance del desarrollo de un sistema basado en Procesamiento de Lenguaje Natural (PLN) orientado a la identificación y vinculación de genes con sus funciones biológicas a partir de artículos científicos. El proyecto integra un proceso que abarca la recopilación y exploración del corpus, el preprocesamiento lingüístico y la implementación de modelos de aprendizaje supervisado enfocados en el reconocimiento de entidades y la extracción de relaciones. Los resultados parciales evidencian el potencial del enfoque propuesto para estructurar información biomédica dispersa y facilitar la creación de bases de datos que consoliden el conocimiento científico disponible.

*Palabras claves:* Biomedicina, Reconocimiento de Entidades, Extracción de Relaciones, Genes

1. INTRODUCCIÓN

En la actualidad, el crecimiento exponencial de la literatura biomédica representa un desafío significativo para investigadores y profesionales del área de la salud. La cantidad de artículos publicados diariamente en bases de datos como PubMed o PMC dificulta la lectura manual y la extracción eficiente del conocimiento científico disponible. En este contexto, surge la necesidad de desarrollar herramientas automáticas capaces de procesar grandes volúmenes de texto y transformar la información no estructurada en datos organizados y consultables.

El presente proyecto propone el diseño e implementación de un sistema basado en Procesamiento de Lenguaje Natural (PLN) orientado a la identificación automática de genes y sus funciones biológicas en artículos científicos. Mediante el uso de modelos preentrenados especializados en biomedicina, el sistema busca reconocer entidades relevantes (genes y funciones) y extraer las relaciones entre ellas, con el fin de generar una base de datos estructurada que consolide la información dispersa en la literatura científica.

1. ESTADO DEL ARTE

El PLN se ha convertido en una tecnología crucial para manejar la creciente cantidad de literatura biomédica, facilitando la extracción automática y estructuración del conocimiento que se encuentra disperso en textos científicos. Entre las aplicaciones principales están la identificación de entidades biomédicas (genes, proteínas, enfermedades) y la extracción de relaciones funcionales entre estas entidades.

El reconocimiento automático de entidades nombradas (Named Entity Recognition, NER) es la base para detectar genes en los textos, complementado por algoritmos de extracción de relaciones para identificar funciones biológicas y asociaciones entre genes y demás entidades. Se utilizan enfoques tradicionales de PLN junto con modelos de lenguaje avanzados, como los modelos de transformadores basados en Deep Learning, que ahora dominan este campo por su capacidad para comprender contexto y relaciones complejas en el texto biomédico.

Diversos modelos y herramientas han sido desarrollados para estas tareas. Entre ellos destacan SciSpacy y BioBERT, que representan hitos en la aplicación de modelos de lenguaje profundo en dominios biomédicos. SciSpacy es una extensión de SpaCy entrenada con textos científicos, diseñada para reconocer entidades y relaciones con gran precisión . Por su parte, BioBERT, una versión de BERT preentrenada con millones de artículos de PubMed y PMC, ha mostrado resultados sobresalientes en tareas de clasificación semántica y extracción de relaciones genéticas.. Estas arquitecturas superan ampliamente los enfoques tradicionales basados en reglas o diccionarios, al capturar mejor el contexto lingüístico especializado.

A pesar de los avances alcanzados, persisten desafíos significativos en la aplicación del PLN a la biomedicina. La heterogeneidad terminológica, la ambigüedad semántica de los nombres de genes y la variabilidad en los formatos de publicación continúan dificultando la extracción automática de información con alta precisión (Zhang et al., 2019). Asimismo, la integración de información dispersa en bases estructuradas plantea retos técnicos y conceptuales, especialmente en lo relativo a la trazabilidad y la evidencia textual.

1. ENTENDIMIENTO DEL NEGOCIO
2. Objetivos

El presente proyecto tiene como objetivo desarrollar una herramienta de Procesamiento de Lenguaje Natural (PLN) aplicada al análisis de literatura biomédica, con el fin de identificar automáticamente genes y sus funciones biológicas en artículos científicos. Esta herramienta permitirá estructurar el conocimiento disperso en grandes volúmenes de textos y almacenarlo en una base de datos consultable que relacione genes, funciones y evidencia textual.

Desde la perspectiva del negocio, este proyecto representa el punto de partida para una solución tecnológica innovadora orientada a optimizar los procesos de revisión, análisis y extracción de información biomédica. Su implementación puede beneficiar a laboratorios farmacéuticos, grupos de investigación, universidades y empresas del sector biotecnológico, al facilitar la integración y sistematización del conocimiento científico.

1. Descripción del conjunto de textos seleccionado

El conjunto de textos seleccionado está conformado por artículos científicos especializados en biología molecular y genética, centrados en la descripción de genes y sus funciones. Los documentos pertenecen principalmente a estudios relacionados con familias génicas como RAS y PLA2G2A, involucradas en diversos procesos celulares y patologías.

Entre los títulos que componen el corpus se encuentran, por ejemplo:

* *“The Secretory Phospholipase A2 Gene Is a Candidate”*
* *“RAS Is Regulated by the let-7 MicroRNA Family”*
* *“Multiple Forms of Phospholipase A2 in Arthritic Synovial Fluid”*
* *“NRAS Mutation Causes a Human Autoimmunelymphoproliferative Disorder”*

De cada artículo se extrajeron los siguientes elementos:

* Autor(es): responsables de la publicación científica.
* Título: nombre del artículo, indicador principal del tema tratado.
* Abstract: resumen del contenido, empleado para la detección automática de entidades (genes y funciones).
* Referencias: conjunto de fuentes bibliográficas citadas, que permiten establecer relaciones entre trabajos y autores.

Estos textos conforman el corpus de entrenamiento y análisis sobre el cual se aplican las técnicas de PLN para la identificación de entidades y la extracción de relaciones gen-función.

1. Proceso de carga u obtención de textos

El proceso de obtención de los textos se realizó a partir de la plataforma OMIM (Online Mendelian Inheritance in Man) [1], específicamente desde la sección Gene Map, que recopila información genética y publicaciones científicas asociadas a genes humanos y sus mutaciones.

Para cada gen de interés, se consultó su entrada en la Gene Map, desde la cual se accedió a los artículos vinculados. Los documentos fueron descargados en formato PDF, organizados en carpetas correspondientes a cada gen (por ejemplo, *RAS* y *PLA2G2A*), y posteriormente procesados para la extracción automática de metadatos relevantes: autor, título, abstract y referencias.

1. ENTENDIMIENTO Y PREPARACIÓN DEL CONJUNTO DE TEXTO
2. Análisis exploratorio del conjunto de textos.

El análisis exploratorio del corpus tuvo como propósito comprender la estructura, extensión y características lingüísticas de los textos recopilados. Los documentos se componen de resúmenes científicos (abstracts) obtenidos de artículos relacionados con genes y sus funciones biológicas.

A partir del código desarrollado, los textos fueron cargados en un marco de datos (DataFrame) mediante la biblioteca pandas, lo que permitió examinar la cantidad de registros, longitud promedio de los abstracts y presencia de valores nulos o duplicados.

No se detectaron vacíos significativos de información, garantizando la integridad del conjunto de datos. Asimismo, se verificó la existencia de términos técnicos recurrentes en el ámbito biomédico, los cuales resultan de interés para las posteriores etapas de extracción de entidades.

1. Análisis descriptivo

El análisis descriptivo se enfocó en la frecuencia y distribución de las palabras dentro del corpus, con el fin de identificar patrones léxicos característicos del dominio biomédico. Para ello, se aplicaron métricas básicas como el conteo total de tokens, el número de vocablos únicos y la frecuencia media de aparición.

Se utilizó la biblioteca NLTK para realizar el conteo y analizar la frecuencia de los términos más comunes. Los resultados mostraron una alta ocurrencia de palabras científicas relacionadas con procesos genéticos y moleculares.

1. Visualización de los textos

Con el fin de representar visualmente la distribución de los términos más frecuentes, se emplearon herramientas gráficas basadas en matplotlib y wordcloud. Las visualizaciones permitieron observar las palabras más representativas del corpus, evidenciando el predominio de conceptos asociados a la regulación genética y la expresión de proteínas.

La nube de palabras generada reveló una alta densidad en términos como tumor, mutation, chromosome y cell, lo cual valida la homogeneidad temática y proporciona una visión intuitiva de la terminología clave.

Estas representaciones gráficas facilitaron la detección de posibles sesgos léxicos y orientaron las decisiones de limpieza y normalización del texto en la fase de preprocesamiento.

1. Preprocesamiento en NLP

Este proceso incluyó varias operaciones secuenciales:

* Normalización de texto: conversión de todos los caracteres a minúsculas y eliminación de signos de puntuación, números y caracteres especiales, con el fin de reducir la variabilidad léxica.
* Tokenización: división de cada documento en unidades léxicas (tokens) mediante herramientas de NLTK.
* Eliminación de stopwords: supresión de palabras vacías en inglés, como the, is, of y and, que no aportan información semántica relevante.
* Lematización: aplicación del lematizador para reducir las palabras a su forma canónica, permitiendo agrupar variaciones morfológicas de un mismo término.
* Reconstrucción del corpus limpio: una vez aplicadas las transformaciones, los textos procesados fueron almacenados en una nueva columna del DataFrame, denominada clean\_text, lista para su uso en las etapas posteriores de representación vectorial y modelado.

Este flujo de preprocesamiento permitió obtener un corpus homogéneo, libre de ruido y optimizado para la extracción de características.

1. TRANSFORMACIÓN DEL TEXTO PARA EXTRACCIÓN DE CARACTERÍSTICAS
2. Implementación de estrategia de extracción de características de los textos

Una vez finalizado el proceso de preprocesamiento y limpieza del corpus, se procedió a transformar los textos en representaciones numéricas adecuadas para su análisis por parte de los modelos de aprendizaje supervisado. Este proceso permitió convertir los documentos en vectores que reflejan la frecuencia y relevancia de los términos presentes. Cada texto preprocesado fue reconstruido a partir de sus tokens lematizados, generando una columna denominada texto\_final.

1. Selección de características/vectorización

La primera estrategia implementada fue Bag of Words (BoW), mediante la clase CountVectorizer. Este enfoque se centra en contabilizar la frecuencia de aparición de cada palabra dentro del corpus, generando una matriz dispersa en la que las filas representan documentos y las columnas corresponden a los términos más frecuentes. Para limitar la dimensionalidad y evitar el sobreajuste, se estableció un máximo de 1000 características representativas. El análisis de las palabras seleccionadas mostró la predominancia de términos asociados con procesos biológicos y moleculares, como expression, mutation, protein y activity.

La segunda estrategia fue la vectorización mediante TF-IDF, implementada con la clase TfidfVectorizer. Este método pondera cada término considerando su frecuencia en el documento y su rareza en el conjunto total, de manera que las palabras comunes reciben un peso menor y las más informativas adquieren mayor relevancia. Al igual que en BoW, se consideraron 1000 características principales. Los valores TF-IDF resultantes fueron visualizados mediante un gráfico de barras que muestra las veinte palabras con mayor peso promedio, evidenciando los términos con mayor aporte informativo dentro del corpus.

Esta fase de vectorización permitió disponer de una representación cuantitativa del lenguaje científico contenida en los artículos, lista para ser utilizada en las etapas de entrenamiento y validación de los modelos de reconocimiento de entidades y relaciones gen–función.

REFERENCIAS

1. Online Mendelian Inheritance in Man (OMIM). [En línea]. Disponible en: <https://www.omim.org/>