# Predicción de Conocimiento Faltante en Neurocirugía mediante Modelos de Grafos y Procesamiento de Lenguaje Natural

Gabriel Alonso Coro C-412
Mauro Eduardo Campver Barrios C-411
Salma Fonseca Curbelo C-412
Gabriel Herrera Carrazana C-412
José Ernesto Morales Lazo C-412
Adrián Alejandro Souto Morales C-412

**Resumen**

Este trabajo presenta un sistema automatizado diseñado para construir un Grafo de Conocimiento a partir de textos médicos en español y descubrir relaciones entre conceptos que no están explícitas en los documentos. Se diseñó una metodología híbrida que combina un modelo Transformer de Reconocimiento de Entidades Nombradas para la identificación de conceptos médicos, con un conjunto de reglas sintáctico-gramaticales y de proximidad textual para la extracción inicial de relaciones. Posteriormente, se aplicaron algoritmos de Aprendizaje de Máquinas, específicamente modelos de representación vectorial, para predecir y sugerir nuevas conexiones. Los resultados fueron verificados utilizando ontologías médicas, demostrando que el sistema es capaz de encontrar asociaciones relevantes para el apoyo a la investigación clínica.

**Palabras clave:** Grafos de Conocimiento, Neurocirugía, Predicción de Enlaces, Procesamiento de Lenguaje Natural, Aprendizaje de Máquinas.

# Índice

# 1. Introducción

La neurocirugía es una especialidad médica de alta complejidad donde el conocimiento se actualiza constantemente. Sin embargo, la enorme cantidad de artículos, reportes de casos y ensayos clínicos que se publican diariamente hace prácticamente imposible que un especialista pueda leer y relacionar todos los nuevos hallazgos. Esta sobrecarga de información provoca que existan conexiones importantes entre síntomas, tratamientos y enfermedades que, aunque están presentes en los datos, pasan desapercibidas.

El Aprendizaje de Máquinas (Machine Learning) y el Procesamiento de Lenguaje Natural (PLN) ofrecen herramientas capaces de leer y procesar estos textos de manera automática. En este contexto, destacan los Grafos de Conocimiento, los cuales organizan la información como una red de conceptos interconectados. No obstante, un problema común es que los grafos generados automáticamente suelen estar incompletos o tener huecos de información.

El objetivo principal de este proyecto es construir un grafo a partir de literatura de neurocirugía y utilizar modelos matemáticos para identificar las conexiones faltantes. Para lograrlo, se utilizan técnicas de análisis de texto y modelos de predicción que permiten inferir y sugerir nuevas relaciones médicas, ayudando así a integrar mejor el conocimiento disponible.

# 2. Estado del Arte

Para el desarrollo de este trabajo, se investigaron las técnicas actuales de extracción de información médica y completamiento de grafos. A continuación, se describen los fundamentos que respaldan la metodología elegida.

## 2.1. Grafos de Conocimiento en Medicina

Un Grafo de Conocimiento es una estructura que representa conceptos (nodos) y las relaciones que los unen (aristas). En el campo de la salud, esta estructura es clave para integrar datos que suelen estar separados. Según estudios recientes [4], estos grafos permiten mejorar los sistemas de búsqueda y recuperación de información. Sin embargo, el procesamiento de textos médicos en español presenta una dificultad mayor que en inglés debido a la menor disponibilidad de herramientas especializadas.

## 2.2. Reconocimiento de Entidades en el Dominio Biomédico

El primer obstáculo al analizar textos médicos es que el lenguaje clínico es muy distinto al lenguaje cotidiano. Los modelos generales (entrenados con textos como Wikipedia) suelen cometer errores al intentar identificar términos técnicos.

La literatura científica indica que es fundamental utilizar modelos adaptados al dominio específico. Según Miranda Escalada et al. [5], usar modelos entrenados con historias clínicas y artículos médicos mejora notablemente la precisión.

Por este motivo, en este trabajo se utiliza el modelo `spanish_medical_ner`. Este modelo está diseñado específicamente para detectar entidades como Enfermedades, Procedimientos y Sustancias.

## 2.3. Estrategias para la Construcción de Grafos

Una vez identificados los conceptos, el siguiente paso es determinar cómo se relacionan entre sí. Lo ideal sería entrenar un modelo con miles de ejemplos corregidos por médicos, pero al no disponer de ese recurso, se optó por estrategias alternativas:

- **Análisis de la Estructura Gramatical:** Fundel et al. [3] demostraron que las relaciones biomédicas suelen seguir patrones gramaticales fijos. Por ejemplo, analizando quién es el sujeto y quién el objeto de un verbo en una oración, se pueden extraer relaciones con alta fiabilidad sin necesidad de entrenamiento manual.
- **Co-ocurrencia Controlada:** Percha y Altman [6] validaron que, en grandes volúmenes de texto, si dos términos aparecen juntos con mucha frecuencia, es muy probable que exista una relación entre ellos.

Finalmente, para limpiar los posibles errores generados por la automatización, se aplica un proceso de validación semántica. Como sugiere Paulheim [**?**], es necesario contrastar los datos extraídos con bases de conocimiento oficiales (Ontologías) para filtrar aquellas predicciones que no tienen sentido médico.

## 2.4. Predicción de Enlaces (Link Prediction)

Los grafos de conocimiento rara vez están completos. La tarea de Predicción de Enlaces consiste en adivinar qué relaciones faltan basándose en las que ya existen. Una de las técnicas más utilizadas es Modelos de Embedding de Grafos(KGE). Estos modelos convierten los conceptos y relaciones en embeddings para poder operarlos matemáticamente.

- **Modelos Traslacionales:** Modelos como **TransE** [7] interpretan la relación como una suma vectorial. Sin embargo, tienen problemas con relaciones complejas. Variantes más modernas como **RotatE** resuelven esto modelando las relaciones como rotaciones en un espacio complejo, lo que permite entender patrones como la simetría [4].
- **Modelos de Factorización:** Modelos como **DistMult** o **ComplEx** [2] utilizan operaciones de multiplicación para calcular qué tan probable es que una relación sea verdadera.

### 2.5. Evaluación de Modelos de Predicción

La evaluación de los modelos de *Link Prediction* presenta desafíos únicos, especialmente el desequilibrio de clases (existen muchas más relaciones falsas/inexistentes que verdaderas).

Tradicionalmente, se utilizan métricas basadas en el rango (Ranking-based metrics) como **Mean Reciprocal Rank (MRR)** y **Hits@k**, que evalúan la posición de la tripleta correcta frente a un conjunto de tripletas corruptas [1]. Sin embargo, en escenarios de clasificación binaria (¿existe o no esta arista?), métricas como el Área Bajo la Curva ROC (**AUC-ROC**) y la Curva de Precisión-Recall (**AUC-PR**) son preferibles. Akrami et al. [1] señalan que el AUC-PR es más informativo que el AUC-ROC en grafos altamente desbalanceados, ya que penaliza más fuertemente los falsos positivos, proporcionando una visión más realista de la utilidad clínica del modelo.

## 3. Procesamiento de Texto

El paso inicial para crear el grafo consistió en la obtención y limpieza de la información contenida en los documentos PDF de neurocirugía.

### 3.1. Obtención y Limpieza de Datos

Se creó un flujo de trabajo utilizando herramientas de programación (`PyMuPDF` y `pytesseract`) para leer los documentos. Dado que el texto extraído suele contener ruido", se aplicaron las siguientes técnicas de limpieza:

- **Eliminación de elementos irrelevantes:** Se borraron encabezados, números de página y bibliografías que no aportan información clínica.
- **Normalización:** Uso de `unidecode` para normalización de caracteres y eliminación de palabras vacías (*stopwords*) irrelevantes para el dominio médico, manteniendo términos clínicos esenciales.

- **Segmentación:** Se utilizó `NLTK` para la tokenización de oraciones en idioma español.

# 4. Construcción del Grafo de Conocimiento

La transformación del texto plano a una estructura de grafo matemático $G = (V, E)$ se realizó combinando inteligencia artificial y reglas lingüísticas.

## 4.1. Extracción de Entidades (Nodos)

Para identificar los nodos del grafo, se utilizó un modelo de Reconocimiento de Entidades Nombradas (NER) basado en la arquitectura Transformers. Específicamente, se usó `HUMADEX/spanish_medical_ner`, especializado en medicina. Para evitar tener nodos duplicados (como tumor y tumores), se aplicó un proceso de lematización que convierte las palabras a su forma base.

## 4.2. Extracción de Relaciones (Aristas)

Las conexiones iniciales se crearon basándose en la gramática y la proximidad de las palabras. Se establece una conexión si:

1. Existe una relación gramatical directa (ancestro-descendiente) en el análisis sintáctico de la oración.
2. Ambas entidades comparten el mismo verbo principal.
3. Aparecen juntas en una ventana de texto pequeña (menos de 5 palabras de distancia).

## 4.3. Validación y Filtrado

Para asegurar la calidad del grafo, se aplicaron dos filtros:

1. **Análisis de Importancia:** Se calcularon métricas de centralidad para eliminar nodos aislados que no aportaban información útil a la red.
2. **Validación con Ontologías:** Se implementó un validador que consulta bases de datos médicas oficiales (como HPO, DOID y NCIT). Esto permite verificar si los conceptos y relaciones extraídos tienen sentido médico real.

## 4.4. Caracterización Topológica del Grafo Base

Antes de proceder con el entrenamiento de los modelos de predicción, se realizó un análisis cuantitativo de la estructura del grafo resultante. Este análisis es fundamental para entender la conectividad de los conceptos médicos extraídos y seleccionar las estrategias de aprendizaje adecuadas.

### 4.4.1. Métricas Globales

El grafo construido presenta las siguientes características topológicas:

- **Densidad ($0{,}0017$):** El valor extremadamente bajo confirma que se trata de un *grafo disperso*.
- **Coeficiente de Clustering ($0{,}4823$):** A pesar de la baja densidad, el coeficiente de agrupamiento es relativamente alto ($\approx 48\,\%$). Esto indica una fuerte tendencia a formar comunidades locales; es decir, si el concepto A está conectado con B y C, es muy probable que B y C también estén conectados entre sí, formando clústeres temáticos específicos.
- **Grado Promedio ($5{,}89$):** En promedio, cada entidad médica está conectada con otros 6 conceptos, lo que permite una propagación de información eficiente en las capas convolucionales de la GCN.
- **Componentes Conectados ($98$):** El grafo no es totalmente conexo; existen 98 islas de conocimiento. Sin embargo, dado el tamaño del grafo, esto sugiere una componente gigante principal que agrupa la mayoría de los términos relevantes, con pequeñas islas periféricas.

### 4.4.2. Nodos Centrales

El análisis de centralidad de grado permitió identificar los conceptos más influyentes dentro de la red.

Tabla 1: Top 10 Nodos con mayor grado en el Grafo

| Entidad (Nodo) | Grado (Conexiones) |
|---|---|
| lesión | 330 |
| tumor | 206 |
| quirúrgico | 171 |
| resección | 167 |
| tumoral | 152 |
| complicación | 149 |
| cirugía | 141 |
| endoscópico | 137 |
| tratamiento | 124 |
| meningioma | 118 |

# 5. Predicción de Aristas Faltantes

Una vez construido el grafo base, el objetivo central fue predecir el conocimiento faltante, es decir, inferir relaciones probables que no estaban escritas explícitamente pero que son lógicas en el contexto médico.

## 5.1. Modelo mediante embeddings (KGE)

Se utilizó la librería `PyKEEN` para entrenar y comparar múltiples modelos de incrustación, buscando capturar diferentes patrones geométricos:

- **RotatE:** Modela relaciones como rotaciones en un espacio complejo (útil para simetría/antisimetría).
- **DistMult:** Utiliza una función de puntuación bilineal, eficaz para relaciones simétricas.
- **ComplEx:** Extiende DistMult al dominio complejo para modelar relaciones asimétricas.

## 5.2. Refinamiento Semántico con LLMs

Adicionalmente, se integró el modelo de lenguaje **Mistral AI**. Mientras que los modelos matemáticos anteriores predicen si existe una conexión, Mistral se utilizó para inferir qué tipo de relación basándose en el contexto.

# 6. Evaluación de la Predicción

Evaluar estos modelos es un reto porque en un grafo existen muchas más conexiones inexistentes que reales (desequilibrio de clases).

## 6.1. Métricas Utilizadas

Para medir el rendimiento, se separó un 15 % de las conexiones para usarlas como prueba:

- **AUC-ROC:** Mide la capacidad general de clasificación.
- **AUC-PRC:** Área bajo la curva Precisión-Recall. Esta métrica se priorizó porque es más honesta cuando hay desequilibrio de datos, ofreciendo una visión más realista de la utilidad clínica.
- **F1-Score, Precisión y Recall:** Métricas estándar para evaluar la exactitud de las predicciones.

# 7. Resultados Obtenidos

En esta sección se presentan los resultados cuantitativos de los modelos de representación de grafos (KGE) bajo diferentes configuraciones de entrenamiento.

## 7.1. Resultados de Modelos de Embedding (KGE)

Se evaluaron tres arquitecturas distintas: **RotatE**, **DistMult** y **ComplEx**. Para analizar la estabilidad y convergencia de los modelos, se realizaron experimentos con 500 y 1000 épocas de entrenamiento.

### 7.1.1. Escenario 1: Entrenamiento a 500 Épocas

En esta configuración inicial, el modelo RotatE mostró una superioridad notable en la capacidad de recuperar enlaces correctos (Hits@k) en comparación con los modelos de factorización.

Tabla 2: Resultados de KGE con 500 Épocas

| Métrica | RotatE | DistMult | ComplEx |
|---|---|---|---|
| **MRR** | **0.0985** | 0.0255 | 0.0030 |
| Hits@1 | 0.0384 | 0.0045 | 0.0005 |
| Hits@3 | 0.1066 | 0.0188 | 0.0013 |
| **Hits@10** | **0.2173** | 0.0627 | 0.0038 |
| ROC AUC | 0.7502 | 0.7571 | 0.5057 |
| PR AUC | 0.0384 | 0.0423 | 0.0101 |

### 7.1.2. Escenario 2: Entrenamiento a 1000 Épocas

Al duplicar el tiempo de entrenamiento, se observó un comportamiento mixto. Mientras que DistMult mejoró marginalmente, el rendimiento de RotatE en métricas de ranking disminuyó, aunque mantuvo su AUC.

Tabla 3: Resultados de KGE con 1000 Épocas

| Métrica | RotatE | DistMult | ComplEx |
|---|---|---|---|
| **MRR** | 0.0785 | 0.0297 | 0.0050 |
| Hits@1 | 0.0289 | 0.0058 | 0.0015 |
| Hits@3 | 0.0810 | 0.0213 | 0.0030 |
| **Hits@10** | 0.1759 | 0.0675 | 0.0073 |
| ROC AUC | 0.7681 | **0.7783** | 0.5331 |
| PR AUC | 0.0432 | 0.0511 | 0.0109 |

## 7.2. Interpretación y Análisis de Resultados

El análisis comparativo de los resultados arroja tres conclusiones fundamentales sobre la naturaleza del grafo médico y los modelos utilizados:

1. **Superioridad de RotatE:** El modelo RotatE superó consistentemente a DistMult y ComplEx en las métricas de ranking (MRR y Hits@k). Con 500 épocas, RotatE logró posicionar la relación correcta dentro del top 10 en el 21.73 % de los casos, frente al 6.27 % de DistMult.

   *Explicación:* Esto se debe a que RotatE modela las relaciones como rotaciones en un espacio complejo ($h \circ r \approx t$), lo que le permite capturar patrones lógicos

como simetría, antisimetría e inversión. En el dominio médico, muchas relaciones son complejas y direccionales (ej. .ᴬ causa B"no implica "B causa A"), algo que DistMult (que asume simetría pura) no puede modelar eficazmente.

2. **Fallo de ComplEx:** El modelo ComplEx obtuvo resultados cercanos al azar (ROC AUC $\approx$ 0.50-0.53). Esto sugiere que la dimensionalidad o la escasez del grafo actual no permite que este modelo explote adecuadamente las interacciones complejas en forma de producto hermitiano.

3. **Sobreajuste en Entrenamiento Prolongado:** Al comparar 500 vs. 1000 épocas, se observa que el rendimiento de RotatE *empeoró* en métricas de ranking (MRR bajó de 0.098 a 0.078) al aumentar las épocas, a pesar de que el ROC AUC subió ligeramente. Esto indica un posible sobreajuste: el modelo se volvió muy bueno clasificando negativos fáciles (subiendo el AUC), pero perdió precisión fina para ordenar la respuesta correcta entre las opciones más probables. Por tanto, **el modelo entrenado a 500 épocas es el más robusto** para tareas de recomendación práctica.

# 8. Conclusiones

El presente trabajo demostró la viabilidad de aplicar técnicas de Aprendizaje de Máquinas para mitigar la fragmentación del conocimiento en el dominio de la neurocirugía. A través de una metodología que integró Procesamiento de Lenguaje Natural y Aprendizaje de Representaciones en Grafos, se lograron las siguientes contribuciones:

- **Generación Automatizada de Grafos:** Se implementó exitosamente un pipeline capaz de extraer entidades médicas complejas y sus relaciones a partir de texto no estructurado en español, superando las barreras lingüísticas habituales en la bioinformática.

- **Identificación de Patrones Semánticos:** El análisis topológico reveló que, aunque el grafo es disperso (densidad 0.0017), posee una estructura de comunidades definida (clustering 0.48), lo que valida que los términos médicos se agrupan lógicamente según patologías y tratamientos.

- **Eficacia de los Modelos Geométricos:** En la tarea de predicción de enlaces, el modelo **RotatE** demostró ser superior a los enfoques de factorización, validando la hipótesis de que las relaciones médicas requieren modelos capaces de entender direccionalidad y antisimetría. Se determinó que un entrenamiento

moderado (500 épocas) ofrece un mejor equilibrio entre generalización y precisión que entrenamientos prolongados.

En conclusión, el sistema desarrollado ofrece una herramienta prometedora para asistir a investigadores y neurocirujanos, sugiriendo conexiones latentes entre conceptos que podrían pasar desapercibidas en una revisión manual, facilitando así el descubrimiento de conocimiento en la literatura médica.

# Referencias


[1] Akrami, F., Saeef, M. S., Zhang, Q., Hu, W., & Li, C. (2020). *Realistic re-evaluation of knowledge graph completion methods: An experimental study*. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data.

[2] Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., & Duan, Z. (2020). *Knowledge graph completion: A review*. IEEE Access, 8, 192435-192456.

[3] Fundel, K., Küffner, R., & Zimmer, R. (2007). *RelEx—Relation extraction using dependency parse trees*. Bioinformatics, 23(3), 365-371.

[4] Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2021). *A survey on knowledge graphs: Representation, acquisition, and applications*. IEEE Transactions on Neural Networks and Learning Systems, 33(2), 494-514.

[5] Miranda-Escalada, A., Farré-Maduell, E., & Krallinger, M. (2020). *Named entity recognition, normalization and coding of clinical mentions in Spanish: The CANTEMIST track*. IberLEF@SEPLN, 2020.

[6] Percha, B., & Altman, R. B. (2018). *A global network of biomedical relationships derived from text*. Bioinformatics, 34(15), 2614-2624.

[7] Wang, M., Qiu, L., & Wang, X. (2021). *A Survey on Knowledge Graph Embeddings for Link Prediction*. Symmetry, 13(3), 485.