Semantic Mapping of Clinical Variables in Heterogeneous Datasets using Word Embeddings

Ornella Mansilla, Roxana Martínez, Agustín Luján, Nicolás Censi

Centro de Altos Estudios en Tecnología Informática (CAETI)
Universidad Abierta Interamericana (UAI)
Buenos Aires, Argentina
{roxana.martinez}@uai.edu.ar;
{ OrnellaLujan.Mansilla, AgustinEzequiel.LujanBidondo,
NicolasMartin.Censi}@alumnos.uai.edu.ar

Abstract. The integration of heterogeneous clinical datasets remains a challenging task due to the variability in the naming of clinical variables across data sources. Traditional approaches based on textual similarity rely on predefined dictionaries and lexical matching, which are often rigid and difficult to maintain. This work proposes a prototype for semantic mapping of clinical variables using vector-based representations derived from Word Embeddings. The system employs a pretrained Word2Vec model to represent column labels as vectors and uses cosine similarity to identify semantic relationships between dataset attributes and target clinical variables. Additionally, a bidirectional prefix expansion strategy is incorporated to address out-of-vocabulary terms and improve matching robustness. The prototype was implemented in C++ and evaluated using a publicly available clinical dataset, demonstrating that semantic similarity metrics provide a flexible and scientifically grounded alternative to dictionary-based methods. The results support the feasibility of semantic mapping as an effective mechanism for improving early-stage data integration processes in clinical data workflows.

Keywords: Word Embeddings, Semantic Mapping, Natural Language Processing, Data Integration, Cosine Similarity.

Mapeo Semántico de Variables Clínicas en Datasets Heterogéneos utilizando Word Embeddings

Resumen. La integración de datasets clínicos heterogéneos continúa siendo un desafío debido a la variabilidad en la denominación de las variables clínicas entre distintas fuentes de datos. Los enfoques tradicionales basados en similitud textual dependen de diccionarios predefinidos y mecanismos de coincidencia

léxica que resultan rígidos y difíciles de mantener. En este trabajo se presenta un prototipo para el mapeo semántico de variables clínicas utilizando representaciones vectoriales derivadas de Word Embeddings. El sistema emplea un modelo preentrenado Word2Vec para representar etiquetas de columnas como vectores y utiliza similitud del coseno para identificar relaciones semánticas entre atributos del dataset y variables clínicas objetivo. Adicionalmente, se incorpora una estrategia de expansión bidireccional de prefijos para el tratamiento de términos fuera de vocabulario, mejorando la robustez del proceso de identificación. El prototipo fue implementado en C++ y evaluado utilizando un dataset clínico de acceso público, demostrando que las métricas de similitud semántica constituyen una alternativa flexible y científicamente fundamentada frente a métodos basados en diccionarios. Los resultados evidencian la viabilidad del mapeo semántico como mecanismo para mejorar procesos iniciales de integración de datos clínicos.

Palabras clave: Word Embeddings, Mapeo Semántico, Procesamiento del Lenguaje Natural, Integración de Datos, Similitud del Coseno.

1 Introducción

En datasets heterogéneos, una misma variable clínica (ej. glucosa) puede aparecer como “glu”, “glucose” o “glucosa”, lo que impide un análisis automático a escala. Los métodos tradicionales para identificar correspondencias entre variables dependen frecuentemente de diccionarios hardcodeados (por ejemplo, en formato JSON), los cuales resultan rígidos, difíciles de escalar y costosos de mantener a medida que se incorporan nuevas fuentes de datos.

Este trabajo propone un sistema basado en identificación semántica inspirado en la literatura reciente sobre ontologías biomédicas y representaciones vectoriales en procesamiento del lenguaje natural. Estas técnicas han demostrado ser efectivas en tareas de procesamiento del lenguaje natural. El objetivo es demostrar que la similitud vectorial permite realizar un mapeo dinámico entre variables clínicas, sustentado en fundamentos matemáticos y justificable científicamente. En los últimos años, los avances en técnicas de representación semántica han permitido modelar relaciones entre términos mediante espacios vectoriales multidimensionales [4].

En particular, los modelos de Word Embeddings permiten representar palabras como vectores numéricos, donde la distancia entre vectores refleja relaciones semánticas entre conceptos. Este enfoque ofrece ventajas frente a métodos basados exclusivamente en coincidencias textuales, al permitir capturar similitudes implícitas entre términos relacionados conceptualmente.

En este contexto, el presente trabajo propone un prototipo para el mapeo semántico de variables clínicas utilizando representaciones vectoriales derivadas de modelos Word2Vec [1]. El sistema desarrollado permite comparar etiquetas de columnas con un conjunto de variables clínicas objetivo mediante métricas de similitud del coseno, incorporando además estrategias para el tratamiento de términos fuera de vocabulario. La propuesta busca aportar evidencia sobre la viabilidad de enfoques semánticos co-

mo mecanismo para mejorar procesos tempranos de integración de datos en entornos clínicos heterogéneos.

En este contexto, las principales contribuciones de este trabajo son: (i) el diseño de un prototipo para el mapeo semántico de variables clínicas basado en representaciones vectoriales, (ii) la incorporación de una estrategia de tratamiento de términos fuera de vocabulario mediante expansión de prefijos, y (iii) la evaluación preliminar del enfoque sobre un dataset clínico público, evidenciando su aplicabilidad en escenarios de integración de datos heterogéneos.

Este tipo de soluciones resulta particularmente relevante en contextos de interoperabilidad en salud, donde la integración eficiente de datos heterogéneos constituye un desafío central.

2 Trabajo Relacionado

La integración semántica de datos clínicos ha sido ampliamente utilizada mediante enfoques basados en ontologías y diccionarios controlados, como UMLS y SNOMED CT, que permiten establecer correspondencias entre conceptos clínicos mediante estructuras jerárquicas y relaciones semánticas explícitas. Sin embargo, estos enfoques requieren mantenimiento constante y presentan limitaciones frente a variaciones léxicas no contempladas.

Por otro lado, los métodos basados en similitud textual utilizan técnicas de comparación léxica, tales como distancia de edición o coincidencias parciales, que, si bien resultan eficientes en términos computacionales, no logran capturar relaciones semánticas implícitas entre términos.

En los últimos años, el avance de técnicas de procesamiento del lenguaje natural ha permitido el uso de representaciones distribuidas, como Word Embeddings, para modelar relaciones semánticas en espacios vectoriales. En el dominio biomédico, modelos como BioWordVec y BioSentVec han demostrado mejoras en la representación de términos clínicos, permitiendo capturar similitudes contextuales más complejas [1] [2] [3].

En este trabajo se adopta este enfoque vectorial como alternativa a los métodos tradicionales, con el objetivo de evaluar su aplicabilidad en el mapeo automático de variables clínicas en datasets heterogéneos.

3 Metodología y Proceso del Sistema

El prototipo fue desarrollado en lenguaje C++, utilizando una arquitectura local sin dependencia de APIs externas. El sistema se diseñó con el objetivo de permitir la identificación semántica automática de variables clínicas en datasets heterogéneos, priorizando eficiencia computacional y control sobre los componentes utilizados.

La elección de la similitud del coseno se fundamenta en su capacidad para comparar vectores independientemente de su magnitud, lo cual resulta especialmente adecuado en espacios semánticos donde la orientación representa la relación conceptual entre términos.

3.1. Espacio Semántico y Embeddings

Se empleó un modelo preentrenado basado en *Word2Vec* [1], en el cual cada palabra es representada como un vector numérico dentro de un espacio multidimensional [2] [4]. Esta representación permite que términos con afinidad semántica se ubiquen en regiones próximas del espacio vectorial, facilitando su comparación matemática.

Para los fines experimentales de este prototipo, se utilizó una base de embeddings curada manualmente a partir de recursos preentrenados disponibles públicamente [6], en la cual se incorporaron deliberadamente vectores correspondientes a términos clínicos clave como *polyphagia* y *age*.

En el dominio biomédico, modelos especializados como BioWordVec y BioSentVec han demostrado mejoras significativas en la representación semántica de términos clínicos y frases médicas, permitiendo capturar relaciones contextuales complejas [2], [3].

3.2. Métrica de Similitud del Coseno

Para cuantificar el grado de relación semántica entre las variables clínicas objetivo y las etiquetas de columnas presentes en los datasets, se utilizó la métrica de similitud del coseno. Esta métrica es ampliamente utilizada en tareas de procesamiento del lenguaje natural y recuperación de información, debido a su capacidad para medir la proximidad angular entre representaciones vectoriales en espacios multidimensionales.

La similitud del coseno permite comparar dos vectores independientemente de su magnitud, considerando únicamente la orientación relativa entre ellos. En este contexto, cada etiqueta textual es representada como un vector dentro del espacio semántico generado por el modelo de embeddings, y la similitud entre dos términos se calcula mediante la siguiente expresión: la métrica se define como se muestra en la Figura 1.

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

Fig. 1. Representación matemática de la similitud del coseno.

3.3. Pipeline de Procesamiento

El flujo general del sistema se ilustra en la Figura 2. El proceso inicia con una etapa de normalización léxica, en la cual las etiquetas de las columnas son convertidas a minúsculas para asegurar compatibilidad con el vocabulario del embedding.

El sistema actual se encuentra optimizado para cadenas alfanuméricas, identificándose el tratamiento avanzado de caracteres especiales como una posible línea de trabajo futuro.

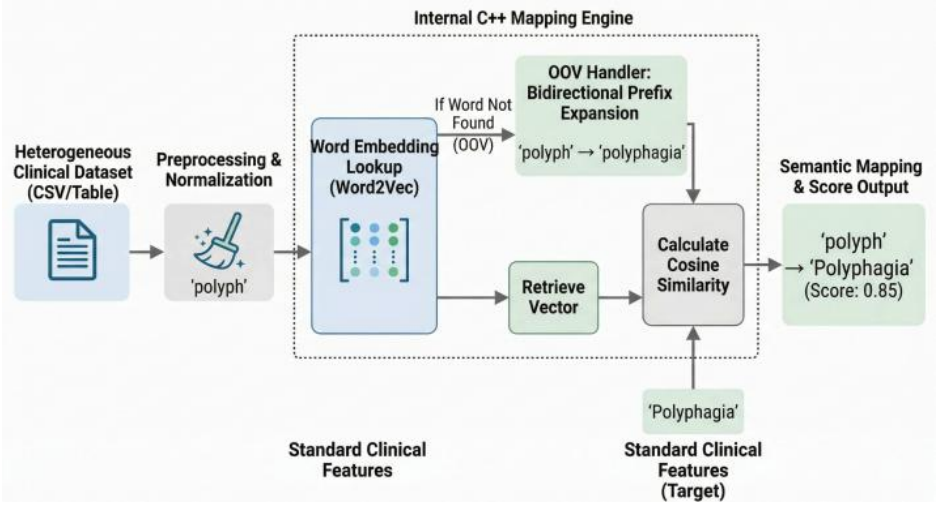

Fig. 2. Pipeline del sistema de identificación semántica de variables clínicas.

3.4. Tratamiento de Términos Fuera de Vocabulario (OOV)

Ante la detección de términos no presentes en el embedding (por ejemplo: *polyph*), se implementó una heurística de expansión de prefijo bidireccional. Esta técnica permite identificar términos completos potenciales (por ejemplo: *polyphagia*) con el fin de aproximar la intención semántica del dato original y permitir su procesamiento posterior.

3.5. Parámetros de Decisión y Penalización

Para controlar el comportamiento del sistema, se definieron parámetros específicos de decisión:

Umbral de similitud (Threshold = 0.6): Este valor fue definido con el objetivo de priorizar la precisión sobre la cobertura, evitando aceptar correspondencias con baja confianza semántica.

Factor de penalización por expansión (0.85): Se aplica este factor al valor de similitud cuando se utiliza el mecanismo OOV. Esta decisión permite evitar que coincidencias indirectas obtengan valores máximos de similitud de manera artificial.

4 Implementación del Prototipo

El sistema fue implementado en lenguaje C++, bajo una arquitectura local sin dependencias externas, con el objetivo de garantizar eficiencia y control sobre el procesamiento de datos.

El prototipo se estructura en módulos que permiten la carga del modelo de embeddings, el preprocesamiento de etiquetas, el cálculo de similitud y la generación de resultados. Esta organización modular facilita la extensibilidad del sistema y su adaptación a distintos escenarios de uso. Además, el diseño adoptado permite incorporar nuevas estrategias de procesamiento semántico sin afectar el núcleo del sistema, lo cual resulta relevante para futuras mejoras y validaciones experimentales.

5 Resultados y Evaluación

El prototipo desarrollado fue evaluado utilizando el dataset público Early Stage Risk Prediction Dataset, disponible en la plataforma Kaggle [5]. Este dataset fue seleccionado debido a su diversidad en la denominación de variables clínicas y su representatividad en escenarios reales de análisis de riesgo en salud. El objetivo principal de la evaluación fue analizar la capacidad del sistema para identificar correspondencias semánticas entre variables clínicas objetivo y atributos presentes en el dataset analizado.

Durante la ejecución del sistema, cada etiqueta de columna fue transformada en una representación vectorial mediante el modelo Word2Vec previamente cargado [1]. Posteriormente, se calcularon valores de similitud del coseno entre los vectores correspondientes a variables clínicas objetivo y atributos del dataset. Este proceso permitió identificar relaciones semánticas incluso en casos donde no existían coincidencias textuales exactas entre los términos analizados.

La Figura 3 presenta un ejemplo de la salida generada por el sistema, donde se muestran valores de similitud semántica asociados a diferentes variables clínicas. En este ejemplo, se observa cómo el sistema logra identificar relaciones semánticas relevantes entre términos abreviados y sus correspondientes formas completas, evidenciando la capacidad del enfoque propuesto para capturar similitudes basadas en significado y no únicamente en coincidencias textuales.

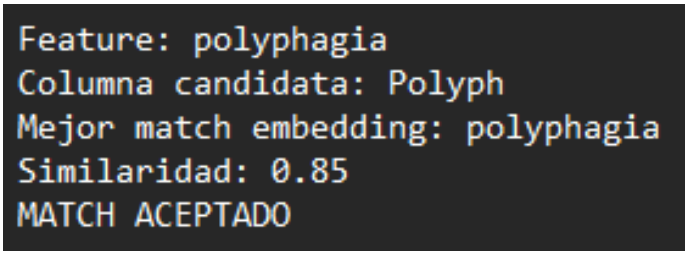

Fig. 3. Ejemplo de salida del sistema mostrando valores de similitud semántica entre variables clínicas y atributos del dataset.

En particular, se observó que el sistema fue capaz de detectar correspondencias semánticas en casos donde las etiquetas presentaban variaciones léxicas o abreviaciones frecuentes en entornos clínicos. Además, la incorporación del mecanismo de expansión de prefijos permitió procesar términos inicialmente no presentes en el vocabulario del modelo, mejorando la cobertura general del sistema.

Los valores obtenidos durante las pruebas experimentales se resumen en la Tabla 1, donde se presentan ejemplos representativos de similitud semántica entre variables clínicas objetivo y atributos del dataset analizado.

El análisis de los resultados permitió observar que valores de similitud superiores al umbral definido indicaron correspondencias semánticas confiables entre variables, mientras que valores inferiores fueron descartados como coincidencias no significativas. Este comportamiento evidencia la efectividad del uso de métricas vectoriales como mecanismo de identificación semántica en entornos con alta variabilidad terminológica.

Table 1. Resultados de la identificación semántica.

Variable Objetivo (Feature)	Columna Dataset	Score Resultante	Estado
age	Age	1.00	Match Aceptado
polyphagia	Polyph	0.85 (penalizado)	Match Aceptado
glucose	-	-1.00	Sin match confiable

Desde un análisis cualitativo, se observa que el sistema presenta un buen desempeño en la identificación de correspondencias semánticas en presencia de variaciones léxicas y abreviaciones, lo cual constituye un escenario frecuente en datasets clínicos reales.

Si bien la evaluación se realizó sobre un conjunto acotado de variables, los casos analizados permiten observar patrones consistentes en el comportamiento del sistema, especialmente en escenarios de abreviación y variabilidad léxica.

Los resultados obtenidos muestran distintos comportamientos del sistema frente a diferentes tipos de coincidencias semánticas. En el caso de la variable age, se obtuvo un valor de similitud igual a 1.00, indicando una coincidencia semántica directa y completa entre la variable objetivo y la columna del dataset.

Para la variable polyphagia, el sistema logró identificar una correspondencia válida a partir del término abreviado polyph, aplicando el mecanismo de expansión de prefijos y el correspondiente factor de penalización, lo que resultó en un valor final de similitud de 0.85. En contraste, la variable glucose no presentó coincidencias confiables dentro del umbral definido, obteniendo un valor negativo, lo que evidencia el correcto funcionamiento del sistema al evitar asociaciones incorrectas cuando no existe una relación semántica significativa.

No obstante, se identificaron casos en los que el sistema no logra establecer correspondencias, particularmente cuando los términos no se encuentran representados en el

espacio de embeddings, lo que evidencia una limitación asociada a la cobertura del modelo utilizado.

6 Discusión

Los resultados obtenidos permiten analizar las ventajas del enfoque propuesto frente a métodos tradicionales basados en diccionarios. En particular, la utilización de representaciones vectoriales permite capturar relaciones semánticas implícitas, reduciendo la dependencia de estructuras rígidas predefinidas.

Sin embargo, el enfoque presenta ciertas limitaciones. En primer lugar, la calidad de los resultados depende directamente del modelo de embeddings utilizado, lo que puede afectar la cobertura de términos específicos del dominio clínico. En segundo lugar, el sistema no incorpora actualmente información contextual más allá de las etiquetas de las variables, lo cual podría limitar su capacidad de desambiguación.

A pesar de estas limitaciones, el enfoque propuesto constituye una alternativa flexible y escalable para el mapeo semántico en entornos heterogéneos, especialmente en etapas tempranas de integración de datos.

Finalmente, a diferencia de enfoques basados en diccionarios o reglas léxicas, el sistema propuesto no requiere la definición manual de correspondencias, lo que reduce significativamente el esfuerzo de mantenimiento. Sin embargo, no se realizó en esta etapa una comparación cuantitativa directa con dichos enfoques, lo cual constituye una línea de trabajo futuro.

7 Conclusiones

En este trabajo se presentó un prototipo orientado a la identificación semántica automática de variables clínicas en datasets heterogéneos mediante el uso de representaciones vectoriales basadas en Word Embeddings. El sistema desarrollado permite comparar etiquetas textuales utilizando métricas matemáticas de similitud, reduciendo la dependencia de diccionarios rígidos y mejorando la flexibilidad en el proceso de integración de datos clínicos.

Los resultados experimentales obtenidos demostraron la capacidad del sistema para identificar correspondencias semánticas válidas tanto en coincidencias directas como en casos de abreviación parcial. En particular, se observó que el mecanismo de expansión de prefijos permitió reconocer correctamente términos incompletos, como en el caso de polyph, generando un valor de similitud penalizado que se mantuvo por encima del umbral definido.

Desde el punto de vista metodológico, el uso de similitud del coseno sobre representaciones vectoriales demostró ser una estrategia efectiva para abordar problemas de heterogeneidad terminológica en variables clínicas. Este enfoque permite detectar relaciones semánticas más allá de coincidencias textuales exactas, lo que resulta especialmente relevante en entornos clínicos caracterizados por el uso frecuente de abreviaturas y variantes lingüísticas.

Como líneas de trabajo futuro, se propone ampliar la base de embeddings utilizada, incorporar datasets clínicos adicionales y evaluar el comportamiento del sistema bajo distintos valores de umbral de similitud. Finalmente, se considera relevante explorar la integración del prototipo en entornos reales de análisis clínico, con el objetivo de validar su aplicabilidad en escenarios operativos y mejorar su capacidad de generalización.

Referencias

[1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. Proceedings of the International Conference on Learning Representations (ICLR).

[2] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "BioWordVec: Improving biomedical word embeddings with subword information and MeSH," *Scientific Data*, vol. 6, 2019.

[3] Q. Chen, Y. Peng, and Z. Lu, "BioSentVec: Creating sentence embeddings for biomedical texts," in *Proceedings of the 7th IEEE International Conference on Healthcare Informatics*, 2019.

[4] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., draft version, 2026.

[5] I. Dutta, "Early Stage Diabetes Risk Prediction Dataset," Kaggle, 2020. Available: https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset/data

[6] pkugoodspeed, "NLP Word2Vec Embeddings (pretrained) [Dataset]," Kaggle. Available: https://www.kaggle.com/datasets/pkugoodspeed/nlpword2vecembeddingspretrained/data