Análisis Visual de Problemas de Granularidad y Ambiguedad en el Agrupamiento de Intenciones con LLMs

DATASET CLINC150

Cecilia Vilca Alvites

Contexto

En la actualidad organizar grandes volúmenes de texto no etiquetado representa un desafío clave en el procesamiento de lenguaje natural. El agrupamiento de intenciones de usuario, común en asistentes virtuales, se ve afectado por la alta granularidad y ambigüedad semántica de las consultas. Aunque los LLMs han mejorado la representación contextual del texto, sus procesos de clustering siguen siendo poco interpretables. Esta falta de claridad, sumada a la escasez de herramientas visuales interactivas, dificulta el análisis y diagnóstico de errores. Ante ello, se justifica el desarrollo de enfoques visuales que permitan explorar y comprender mejor los agrupamientos generados.

Problema

El agrupamiento de textos en lenguaje natural, específicamente las intenciones de usuario, representa un desafío significativo en el ámbito del procesamiento de lenguaje no supervisado. Este desafío se fundamenta en dos propiedades de los conjuntos de datos de intenciones:

Alta Granularidad:

- Muchas intenciones son específicas y semánticamente cercanas.
- Representaciones por LLMs no siempre discriminan intenciones similares.
- Resultado: solapamiento de intenciones en el espacio de embeddings.

Ambigüedad Semántica:

- Frases iguales o similares etiquetadas con intenciones distintas.
- El significado depende fuertemente del contexto.

Esta dificultad se acentúa en contextos no supervisados, donde la ausencia de etiquetas limita la detección de errores. La escasez de herramientas visuales e interactivas dificulta el análisis, haciendo fundamental contar con recursos que faciliten la identificación de intenciones mezcladas y frases ambiguas.

Objetivos

- Usar LLMs para encontrar temas importantes en los textos y así mejorar la forma en que se representan antes de agruparlos.
- Identificar textos inusuales dentro de los grupos, usando tanto el análisis de intenciones como la visualización.
- Implementar métricas internas para evaluar la calidad de los clústeres, considerando coherencia semántica y separación.
- Diseñar e implementar una interfaz visual que permita explorar los grupos de textos creados, cambiar cuántos grupos queremos ver y mejorar los resultados de manera interactiva.

Metodología: Datos

TABLE I
ATRIBUTOS PRINCIPALES DEL DATASET CLINC150 Y SU SIGNIFICADO

Atributo	Descripción	Tipo de Dato	Rango / Valores Posibles
`text`	Representa la **consulta original del usuario** en lenguaje natural. Es la entrada textual que el sistema debe procesar para inferir la intención.	String	Cadenas de texto que varían en longitud desde 2 hasta 136 caracteres. Ejemplos incluyen "what is my account balance" o "can you please tell me how much money I have left in my primary checking account after deducting all pending transactions?".
`intent`	Es la **etiqueta de la intención** subyacente asociada a la consulta de texto. Sirve como la verdad fundamental (*ground truth*) para la clasificación.	String	**150 categorías de intención** distintas y finamente granularizadas. Incluye intenciones como `pay_bill`, `transfer`, `balance`, `greeting`, `goodbye`, `translate`, `money_transfer`, y la categoría `out_of_scope` (OOS).
`text_length`	Atributo derivado que representa la **longitud de la consulta de texto** en número de caracteres. Se utiliza para análisis exploratorio.	Entero	Valores entre 2 y 136 caracteres. La mayoría de las consultas se agrupan alrededor de los 39 caracteres, con una mediana de 37.
`split`	Indica a qué subconjunto pertenece la instancia, utilizado para la división estándar del dataset para entrenamiento, validación y prueba de modelos.	String	`train` (entrenamiento), `val` (validación), `test` (prueba).
`embeddings`	Atributo derivado, no original del dataset, pero crucial para este proyecto. Son las **representaciones numéricas densas** de cada consulta de texto, generadas por un Large Language Model (LLM) (específicamente, `all-MiniLM-L6-v2`).	Vector	Vector de 384 dimensiones. Cada valor en el vector es un número flotante.
`tsne_x`, `tsne_y`	Atributos derivados de la reducción de dimensionalidad de los *embeddings*. Representan las **coordenadas 2D** de cada consulta en un espacio visual, obtenidas mediante t-SNE, facilitando su visualización e interpretación.	Flotante	Valores que varían según la distribución en el espacio 2D, generalmente en un rango de números reales.

Metodología: Pipeline

CLINC150

text, intent

Generación de Embeddings

LLM: all-MiniLM-L6-v2

Embeddings Semánticos (384D)

Reducción de Dimensionalidad

t-SNE

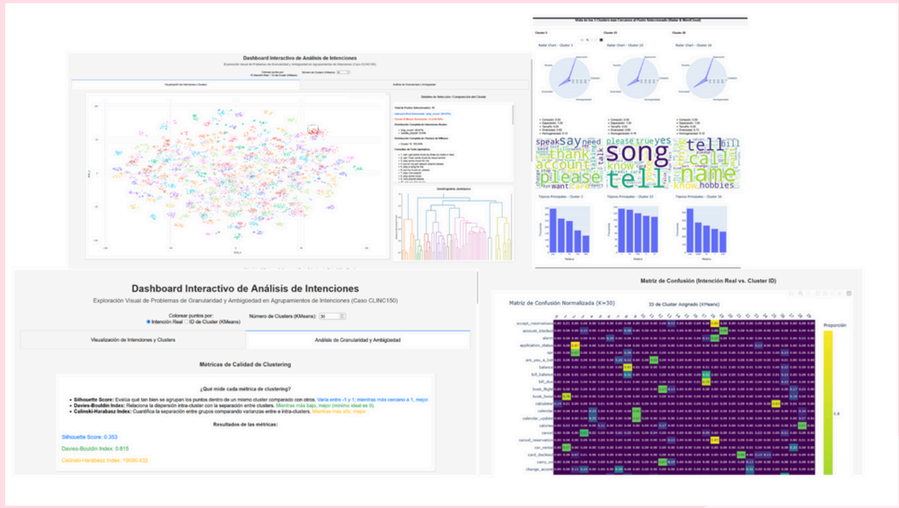
Coordenadas 2D (tsne_x, tsne_y)

Agrupamiento de Intenciones

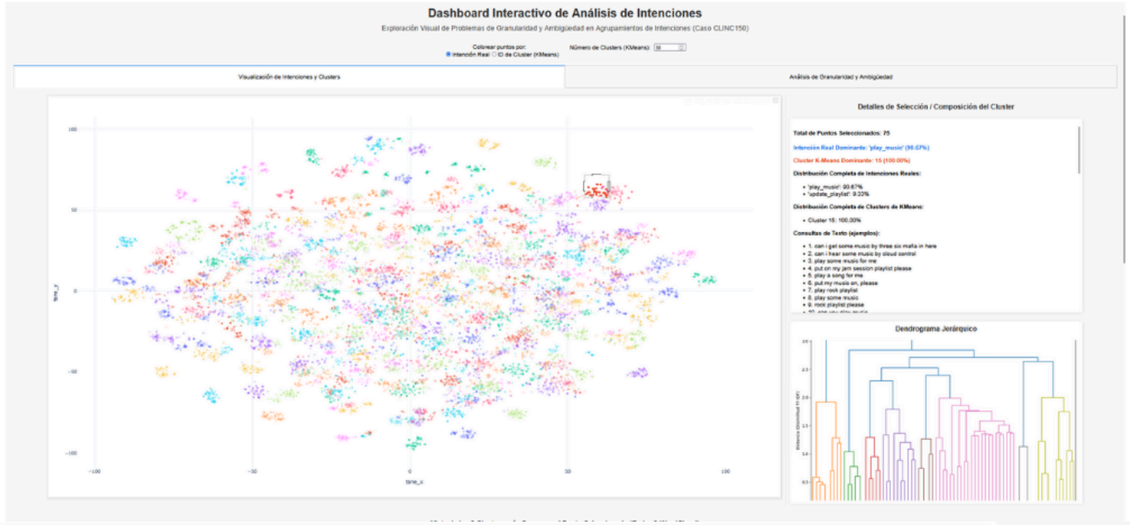
KMeans

Datos Agrupados

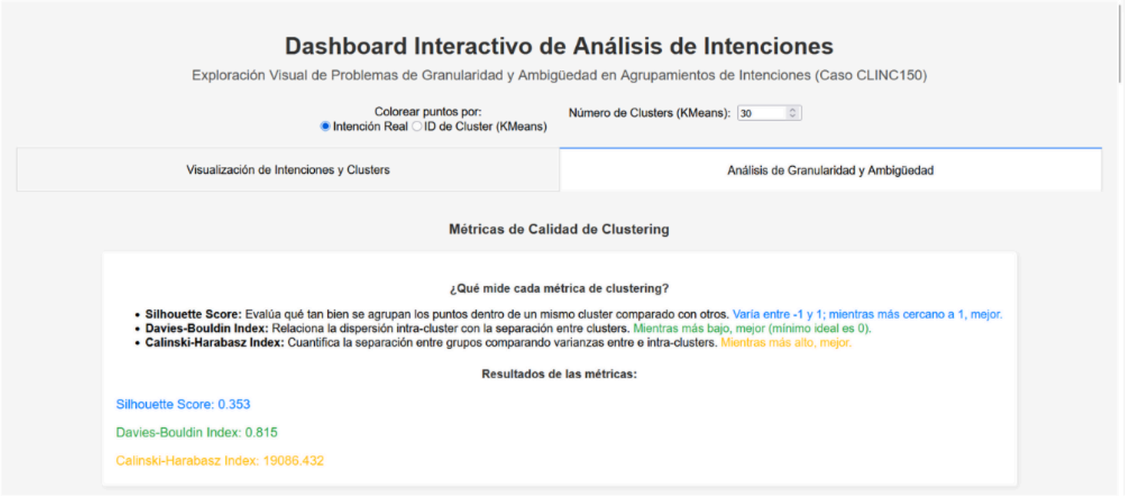
Dashboard

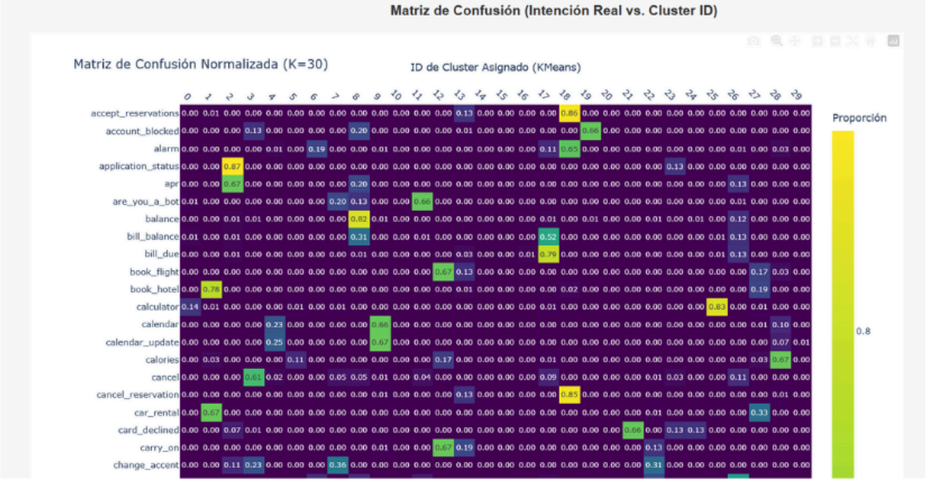


Dashboard


Dashboard Interactivo de Análisis de Intenciones
Exploración Visual de Problemas de Granularidad y Ambigüedad en Agrupamientos de Intenciones (Caso CLINC150)
Colorear puntos por:
Intención Real
ID de Cluster (KMeans)
Número de Clusters (KMeans):
Visualización de Intenciones y Clusters
Análisis de Granularidad y Ambigüedad
Detalles de Selección / Composición del Cluster
Dendrograma Jerárquico


Vista de los 3 Clusters más Cercanos al Punto Seleccionado (Radar & WordCloud)
Cluster 3
Cluster 23
Cluster 26
Radar Chart - Cluster 3
Radar Chart - Cluster 23
Radar Chart - Cluster 26
Tópicos Principales - Cluster 3
Tópicos Principales - Cluster 23
Tópicos Principales - Cluster 26


Dashboard Interactivo de Análisis de Intenciones
Exploración Visual de Problemas de Granularidad y Ambigüedad en Agrupamientos de Intenciones (Caso CLINC150)
Colorear puntos por:
Intención Real
ID de Cluster (KMeans)
Número de Clusters (KMeans): 30
Visualización de Intenciones y Clusters
Análisis de Granularidad y Ambigüedad
Métricas de Calidad de Clustering
¿Qué mide cada métrica de clustering?
Silhouette Score: Evalúa qué tan bien se agrupan los puntos dentro de un mismo cluster comparado con otros. Varía entre -1 y 1; mientras más cercano a 1, mejor.
Davies-Bouldin Index: Relaciona la dispersión intra-cluster con la separación entre clusters. Mientras más bajo, mejor (mínimo ideal es 0).
Calinski-Harabasz Index: Cuantifica la separación entre grupos comparando varianzas entre e intra-clusters. Mientras más alto, mejor.
Resultados de las métricas:
Silhouette Score: 0.353
Davies-Bouldin Index: 0.815
Calinski-Harabasz Index: 19086.432


Matriz de Confusión (Intención Real vs. Cluster ID)
Matriz de Confusión Normalizada (K=30)
ID de Cluster Asignado (KMeans)
Proporción

Referencias

[1] R. Peng, Y. Dong, G. Li, D. Tian, and G. Shan, “TextLens: Large language models-powered visual analytics enhancing text clustering,” *Journal of Intelligent & Fuzzy Systems*, DOI: 10.1007/s12650-025-01043-y, Feb. 2025.

[2] A. Petukhova, J. Carvalho, and N. Fachada, “Text Clustering with Large Language Model Embeddings,” *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 100–108, Dec. 2025.

[3] N. Arias, P. Singh, and A. B. Imbert, “Visual Analytics for Fine-grained Text Classification Models and Datasets,” *arXiv preprint arXiv:2405.02980*, 2024.

[4] L. K. Miller and C. P. Alexander, “Human-interpretable clustering of short text using large language models,” *Royal Society Open Science*, vol. 12, no. 2, pp. 241088, 2025.

[5] S. Hamada, “Processing of Semantic Ambiguity Based on Words Ontology,” *Journal of Computer Science*, vol. 16, no. 1, pp. 1–9, 2020.

Muchas Gracias

Por ver esta presentación