XI Encuentro de Estudiantes de Doctorado de la UPV

Estimación Lineal de Proyecciones Semánticas mediante Mínimos Cuadrados sobre Embeddings de Palabras

Ana Coronado Ferrer
Programa de Doctorado en Matemática Aplicada
Director: Enrique A. Sánchez Pérez

Introducción

Las proyecciones semánticas son funciones que asignan a cada palabra u_i de un universo semántico U y un término externo t, un valor en $[0, 1]$, indicando la relación entre ellos. Se calculan a partir de coocurrencias en documentos recuperados por un buscador B:

$$P_{u_i}^{B}(t) = \frac{|D(u_i) \cap D(t)|}{|D(t)|}$$

donde $D(s)$ representa el conjunto de documentos que contienen el término s.

En este trabajo se estudia cómo estimar las proyecciones semánticas sobre un universo U^2 cuando solo se conocen en otro universo U^1 con características similares. El ejemplo desarrollado corresponde a las proyecciones de la palabra *"rice"* en contextos relacionados con el agua en la agricultura, como parte de un **proyecto PROMETEO** sobre gestión de información agrícola.

Se destaca que diferentes buscadores (repositorios científicos, motores semánticos, etc.) pueden producir variaciones en los valores obtenidos, lo que hace necesario un enfoque robusto de estimación cuando se cambia de universo semántico.

Objetivos

- **Estimación de proyecciones:** Inferir proyecciones semánticas desconocidas usando universos de referencia, mediante la proximidad vectorial en embeddings.
- **Modelo lineal con mínimos cuadrados:** Representar los términos del universo objetivo como combinaciones lineales de un universo base, resolviendo el sistema por mínimos cuadrados.
- **Evaluación del error:** Validar las estimaciones con métricas cuantitativas y acotar el error bajo hipótesis de continuidad Lipschitz.

Metodología

Sean $U^1 = \{u_1^1, \ldots, u_{n_1}^1\}$ y $U^2 = \{u_1^2, \ldots, u_{n_2}^2\}$ dos universos semánticos, con una función de embedding $I : U^1 \cup U^2 \to \mathbb{R}^d$. Suponemos conocidos los valores $P_{u_i^1}(t)$ para un término t y queremos estimar $P_{u_j^2}(t)$.

- Representamos cada vector $I(u_j^2)$ como combinación lineal de los $I(u_i^1)$:

$$I(u_j^2) \approx \sum_{i=1}^{n_1} a_{i,j} I(u_i^1)$$

- Los coeficientes $a_{i,j}$ se obtienen resolviendo un sistema de mínimos cuadrados:

$$X^\top X a_j = X^\top I(u_j^2),$$

donde X es la matriz de embeddings de U^1.

- Estimamos la proyección semántica como:

$$\widehat{P}_{u_j^2}(t) = \sum_{i=1}^{n_1} a_{i,j}\, P_{u_i^1}(t)$$

Este procedimiento permite transferir proyecciones semánticas de un universo a otro, preservando la estructura geométrica del espacio de representación.

Visualización de Resultados

Presentamos en la Figura 1 los valores de las proyecciones semánticas del término "rice"sobre el universo $U^1 =$

$$\{crop, farming, farmland, harvest, irrigation, orchard, soil, tractor\},$$

usando los buscadores de Arxiv, Google Scholar, y DOAJ. También se representa Google, para mostrar como el resultado de un buscador generalista difiere mucho de los otros.

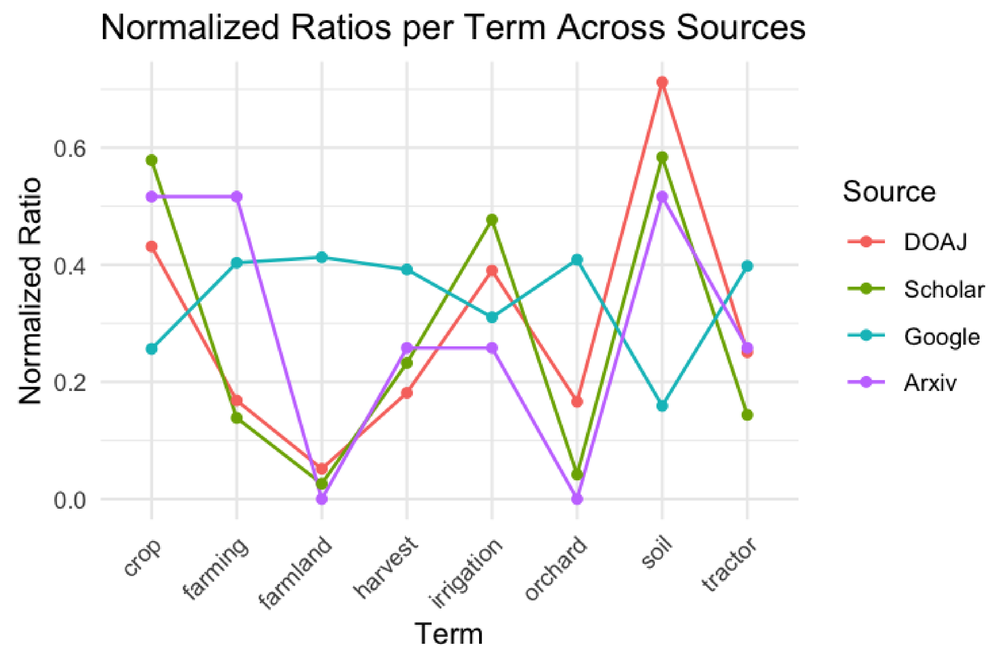

Figura 1. Proyecciones del término 'rice"sobre el universo U^1 mediante cuatro buscadores/repositorios.

La idea es calcular la estimación sobre el universo

$$U^2 = \{farming, waterresources, watering, wetland\}.$$

El resultado puede verse en la Figura 2.

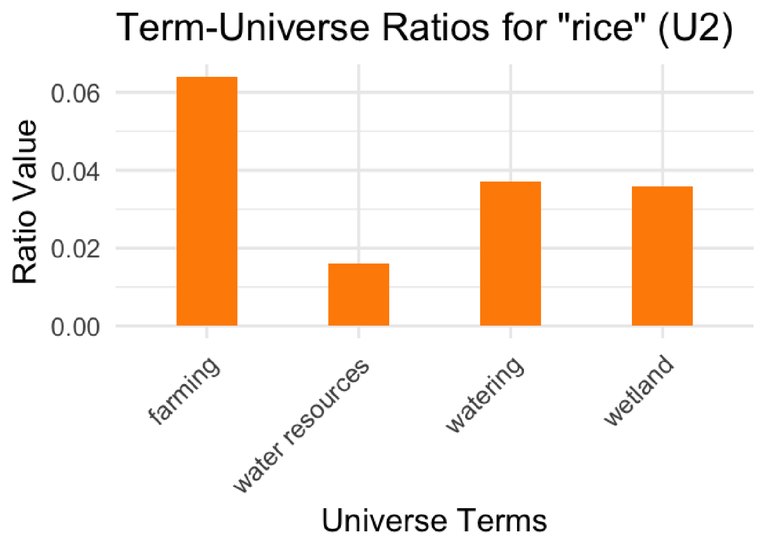

Figura 2. Proyecciones estimadas del término "rice"sobre el universo U^2.

Conclusiones

- La estimación lineal de proyecciones semánticas es viable y eficaz en contextos donde no se dispone directamente del universo de referencia.
- La metodología se basa en fundamentos algebraicos sencillos (mínimos cuadrados) y es aplicable a múltiples entornos vectoriales.
- Puede ser utilizada para estudiar estabilidad semántica entre dominios o entre fuentes de información distintas.

Referencias

Manetti, A., Ferrer-Sapena, A., Sánchez-Pérez, E. A., & Lara-Navarra, P. (2021). Design trend forecasting by combining conceptual analysis and semantic projections: New tools for open innovation. Journal of Open Innovation: Technology, Market, and Complexity, 7(1), 92. https://doi.org/10.3390/joitmc7010092

Fernández de Córdoba, P., Reyes Pérez, C. A., & Sánchez Pérez, E. A. (2025). Mathematical features of semantic projections and word embeddings for automatic linguistic analysis. AIMS Mathematics, 10(2), 3961–3982.

Fernández de Córdoba, P., Reyes Pérez, C. A., Sánchez Arnau, C., & Sánchez Pérez, E. A. (2025). Set-word embeddings and semantic indices: A new contextual model for empirical language analysis. Computers, 14(1), 30.

Agradecimientos

We would like to acknowledge funding from the Generalitat Valenciana (Spain) through the PROMETEO 2024 CIPROM/2023/32 grant.