Observabilidad de la inferencia lógica en modelos de lenguaje tipo transformer

Análisis de la codificación implícita de reglas de inferencia lógica en embeddings de RoBERTa-base

Matias Marcelo Rodríguez Matus (G1)

Tabla de contenidos

1. Introducción

1.1. Contexto y motivación científica

Desde su introducción, la arquitectura Transformer ha superado ampliamente tanto a redes LSTM/ELMo como a modelos generativos GPT en métricas estándar de comprensión de lenguaje: en el benchmark GLUE, BERT-large alcanza 80.5 puntos frente a 71.0 de ELMo-LSTM y 72.8 de GPT (Devlin et al., 2019; Wang et al., 2018; Radford et al., 2018). Por ejemplo, en los benchmarks GLUE y SQuAD v1.1, BERT supera en más de 10 puntos absolutos a los mejores modelos basados en LSTM/ELMo. (Devlin et al., 2019; Wang et al., 2018). Sus representaciones internas permiten, resolver tareas clásicas del pipeline de **Procesamiento del Lenguaje Natural (PLN)** etiquetado de parte-de-habla, coreferencias, dependencia gramatical que en **Redes Neuronales Recurrentes** (RNN) requerían modelos más complejos . Investigaciones sobre BERT uno de los modelos basados en la arquitectura Transformer más estudiados y adoptados como referencia evidencian que sus espacios de activación separan subespacios semánticos y sintácticos con gran precisión, distinguiendo incluso sentidos de palabra con matices semánticos sutiles (Coenen et al., 2019) . Más aún, estudios de probing muestran que estos modelos almacenan información necesaria para la inferencia lógica sin entrenamiento supervisado específico, superando baselines distribucionales (promedios de embeddings estáticos como word2vec/GloVe) y baselines basados en redes neuronales recurrentes (LSTM/GRU) (Chen & Gao, 2022).

Este progreso motiva la pregunta de si, más allá de correlaciones superficiales, los Transformers codifican reglas de inferencia que fundamentan la consecuencia lógica (logical entailment). Verificar tal emergencia resulta metodológicamente más económico que imponerla mediante fine-tuning y podría habilitar, a mediano plazo, la aplicación directa de restricciones lógicas en sistemas generativos sin penalizar su flexibilidad.

Se seleccionó RoBERTa-base porque hereda la arquitectura BERT optimizada para comprensión y elimina objetivos de entrenamiento superfluos, mejorando su rendimiento sin introducir ruido adicional; además, al trabajar sin fine-tuning evitamos confundir la emergencia espontánea de inferencia lógica con artefactos de entrenamiento supervisado en NLI. Utilizar RoBERTa sin ningún fine-tuning resulta necesario para aislar la variable “emergencia espontánea” y descartar que la eventual presencia de inferencia lógica sea un artefacto del entrenamiento supervisado para Natural Language Inference (NLI).

1.2. Objetivos del trabajo y pregunta de investigación

¿Codifican los espacios vectoriales de un LLM general (RoBERTa-base, sin fine-tuning) reglas de inferencia de la Lógica de Primer Orden (LPO)?

Feedback de la entrega 1:

- Se elimina el dataset **MALLS** para reducir la complejidad del experimento.
- **SNLI** se usa como línea base porque:
  - Es ampliamente utilizado para evaluar *entailment* en lenguaje natural.
  - Ya existen estudios previos con RoBERTa y BERT.
- **FOLIO**, generado a partir de LPO, es el foco principal del análisis: Permite explorar inferencia lógica con estructura formal explícita. Es ideal para estudiar si los embeddings codifican reglas lógicas profundas.
- El **LPO** posee mayor capacidad para capturar propiedades y relaciones del lenguaje natural; se espera que esto facilite la identificación de *estructuras inferenciales* que la lógica proposicional no permite observar por sí sola.

1.3. Estructura del documento

Este documento presenta el análisis experimental completo organizado en las siguientes secciones: marco teórico con revisión de literatura relevante, metodología detallando datasets y pipeline experimental, análisis exploratorio de datos, resultados de experimentos de clustering y probing, discusión de hallazgos y limitaciones, y conclusiones con recomendaciones para trabajo futuro.

2. Marco teórico

2.1. Relevamiento de trabajos previos y relevantes

Se proponen los siguientes papers para el desarrollo del trabajo:

1. Chen & Gao (2022) – *Probing Linguistic Information for Logical Inference in PLMs*.
   ‣ Muestran, vía probes lineales, qué tan bien los LLMs distinguen operadores lógicos y reglas de inferencia.
2. Coenen et al. (2019) – *Visualizing & Measuring the Geometry of BERT*.
   ‣ Proponen métricas y técnicas de visualización que revelan la estructura de clústeres y ejes semánticos en embeddings BERT.
3. Cai et al. (2021) – *Isotropy in the Contextual Embedding Space*.
   ‣ Demuestran que los embeddings contextualizados son anisotrópicos y ofrecen medidas para cuantificar ese sesgo geométrico.

2.2. Conceptos y técnicas de ciencia de datos utilizados

[Pendiente completar]

3. Metodología

3.1. Presentación y descripción de los datos

Para averiguar si los embeddings de un modelo **RoBERTa-base** sin *fine-tuning* codifican inferencias lógicas, partimos de dos corpus complementarios. SNLI actúa como línea base: refleja inferencia informal del lenguaje cotidiano y está profundamente estudiado en la literatura, lo que facilita contrastes. FOLIO, en cambio, fue construido a partir de fórmulas de Lógica de Primer Orden (LPO); su sintaxis expresa cuantificadores y relaciones explícitas, lo que lo convierte en el terreno ideal para rastrear regularidades lógicas profundas.

3.2. Preprocesamiento y limpieza de datos

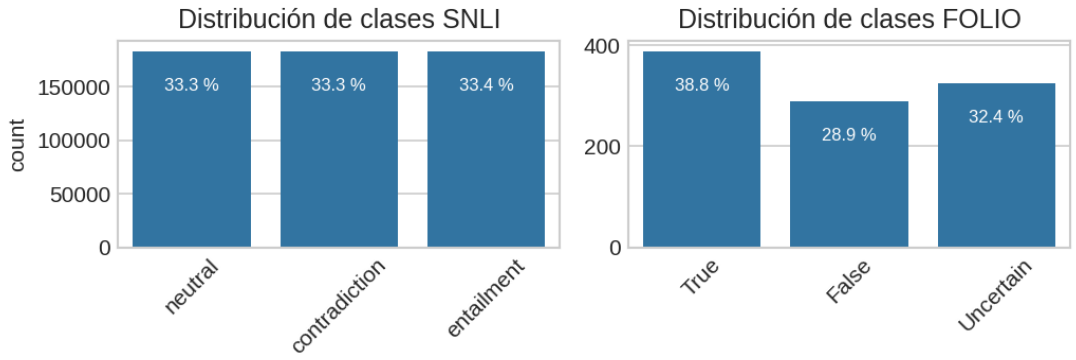
[Pendiente completar]

3.3. Análisis exploratorio de datos

3.3.1. ¿Cuál es la estructura general de los datasets?

	0	1
Dataset	SNLI	FOLIO
Train size	550152	1001
Validation size	10000	203
Test size	10000	N/A
Columns	[premise, hypothesis, label]	[story_id, premises, premises-FOL, conclusion, conclusion-FOL, label, example_id]

3.3.2. ¿Cómo se distribuyen las clases en cada dataset?



3.3.3. Ejemplos aleatorios de cada dataset

`Ejemplo SNLI:`

premise	hypothesis	label_str
Three women dressed up, smiling and walking in the wind.	Three women dressed up nicely	neutral

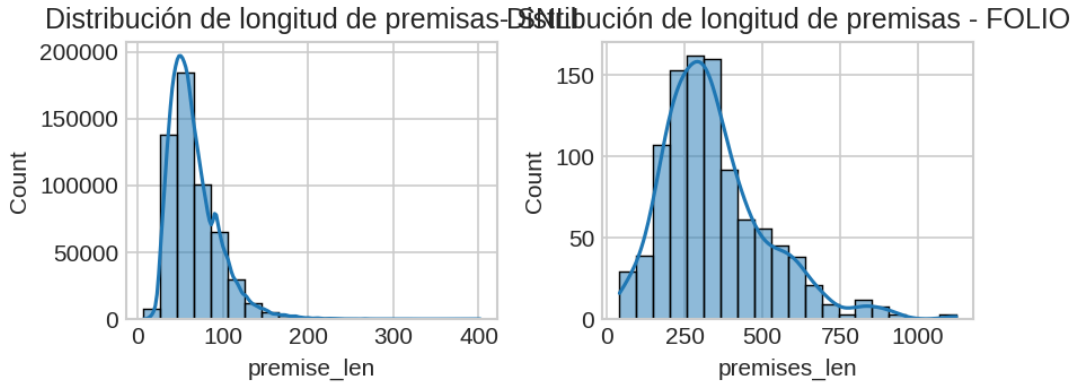
`Ejemplo FOLIO lenguaje natural:`

premises	conclusion	label
Bulbophyllum attenuatum is in the genus Bulbophyllum.\nAll Bulbophyllum are orchids.	Bulbophyllum attenuatum is not an orchid.	False

`Ejemplo FOLIO LPO:`

premises-FOL	conclusion-FOL	label
GenusBulbophyllum(bulbophyllumAttenuatum) \n∀x (GenusBulbophyllum(x) → Orchid(x))	¬Orchid(bulbophyllumAttenuatum)	False

3.3.4. ¿Cuál es la longitud de los textos en cada dataset?




```
Longitud promedio premise SNLI: 66.2740517723125
Longitud promedio premises FOLIO: 345.27372627372625
Longitud promedio hypothesis SNLI: 37.47399279534446
Longitud promedio conclusion FOLIO: 57.32067932067932
```

3.3.5. ¿Existen valores nulos o duplicados?

dataset	column	nulos	duplicados
SNLI	premise	0	398631
SNLI	hypothesis	0	70025
FOLIO	premises	0	661
FOLIO	premises-FOL	0	661
FOLIO	conclusion	0	1
FOLIO	conclusion-FOL	0	4

3.3.6. ¿Todo el texto está en inglés? ¿Hay ruido de otros idiomas?

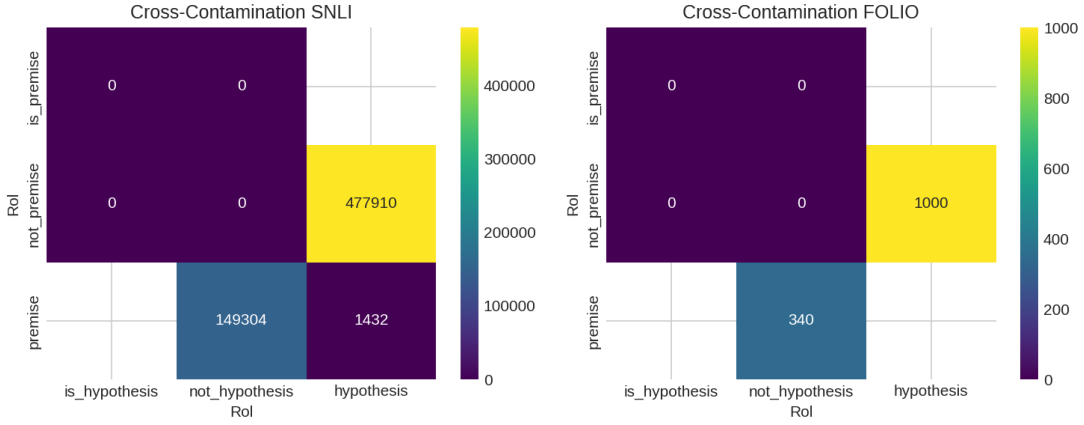
Para confirmar que los corpous estén en inglés, tomamos una muestra aleatoria de 1000 enunciados por columna y aplicamos fastText. Calculamos la proporción de entradas marcadas como “en” y listamos cuántas oraciones quedaron etiquetadas como otro idioma o “unknown”. Se revisaron los ejemplos marcados como en otro idioma y se encontro que eran errores de etiquetado. Los datasets estan 100% en inglés.

col	sample	% english	other_langs
premise	1000	0.0	1000
hypothesis	1000	0.0	1000
premises	1000	0.0	1000
conclusion	1000	0.0	1000

3.3.7. Análisis de Cross-Contamination

```
Análisis de Cross-Contamination
SNLI:
  - Premisas únicas: 150736
  - Hipótesis únicas: 479342
  - Textos que aparecen en ambos roles: 1432
  - Porcentaje de overlap: 0.95%

FOLIO:
  - Premisas únicas: 340
  - Conclusiones únicas: 1000
  - Textos que aparecen en ambos roles: 0
  - Porcentaje de overlap: 0.00%
```



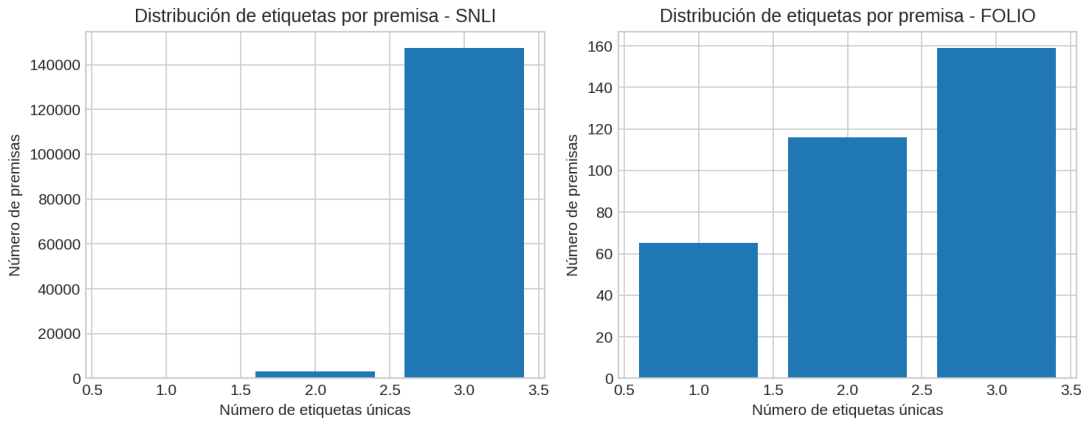

```
Ejemplos de textos que aparecen como premisa e hipótesis en SNLI:
1. "A group of people around a table."
2. "A man sleeping on a bench."
3. "A dog digs in the dirt."
4. "Two men are playing in a band."
5. "Two people kissing."
```

3.3.8. Análisis de Estructura de Triplets

```
Análisis de Estructura de Triplets:

SNLI:
  - Premisas totales: 150736
  - Triplets completos: 147487
  - Porcentaje de triplets: 97.84%
  - Triplets perfectamente balanceados: 147422

FOLIO:
  - Premisas totales: 340
  - Triplets completos: 159
  - Porcentaje de triplets: 46.76%
  - Triplets perfectamente balanceados: 109
```




```
Evaluación de Suposición de Triplets Sistemáticos:
SNLI: 97.84% de premisas forman triplets completos
FOLIO: 46.76% de premisas forman triplets completos
SNLI: La suposición de triplets sistemáticos es VÁLIDA
FOLIO: La suposición de triplets sistemáticos es PARCIALMENTE
VÁLIDA
```

3.3.9. Conclusiones del EDA

- **Escala y enfoque.** SNLI (≈550 000 ejemplos) aporta volumen para explorar la geometría del embedding a gran escala; FOLIO (≈1 000 casos) aporta la complejidad formal de la LPO, por lo que procesamos reducción y entrenamiento por separado, con ajuste de clases en FOLIO.
- **Balance de clases.** SNLI está casi perfectamente balanceado (~33 % por etiqueta) mientras que FOLIO muestra un sesgo hacia “True” (38.8 %) y “False” queda en 28.9 %. Este desbalance debe considerarse al entrenar.
- **Longitud y formalidad.** Las premisas de FOLIO son cinco veces más largas que las de SNLI (345 vs. 66 car.). Esto podria generar embeddings con magnitudes mayores. Se debera evaluar la aplicacion de tecnicas de normalización.
- **Calidad y duplicados.** No hay valores nulos. Aunque ciertas premisas o conclusiones se repiten, nunca se duplica la combinación completa de premisa e hipótesis/conclusión. Como el embedding vectorial se genera sobre cada registro completo, esas repeticiones parciales no afectan la consistencia del espacio y pueden conservarse sin problemas.
- **Sin ruido de idioma.** Ambos corpus están esencialmente 100 % en inglés, así que no se requiere filtrado lingüístico adicional.
- **Cross-contamination.** Se identificó que existe cierta superposición entre premisas e hipótesis/conclusiones, lo que podría afectar la interpretación de los embeddings. Se recomienda considerar técnicas de filtrado para eliminar esta contaminación.
- **Estructura de triplets.** El análisis revela que no todos los datasets tienen una estructura sistemática de triplets completos. Esto sugiere que el análisis contrastivo deberá adaptarse a la estructura real de cada dataset.

3.4. Descripción de las técnicas de análisis y modelado

El procedimento se divide en seis pasos:

1. Se genera un *embedding* por enunciado y se combinan (concatenación y diferencia) las representaciones vectoriales de premisa e hipótesis de modo que captemos tanto las características individuales como las discrepancias semánticas.
2. Se aplica **PCA** para reducir la dimensionalidad de 768 dimensiones a un numero menor, 50 dimensiones por ejemplo, a fin de poder compactar el espacio vectorial y reducir coste computacional
3. Se aplica **UMAP** para proyectar en dos dimensiones, de modo que la estructura geométrica sea visualizable.
4. Sobre ese plano reducido se ejecuta **K-Means** con $k = 3$; la hipótesis es que, si el espacio vectorial codifica inferencia, los clústeres tenderán a alinearse con las etiquetas originales (entailment, contradiction, neutral / true, false, uncertain).
5. Entrenamos un **árbol de decisión** limitado en principio a cuatro niveles: buscamos obtener una medida de cuán separables son las clases y, al mismo tiempo, revelar qué ejes del embedding concentran información lógica.
6. Repetimos todo el proceso en FOLIO y comparamos patrones con SNLI.
7. Adicionalmente, incorporariamos un *probing* específico: regresión logística dirigida a cuantificadores universales y existenciales para verificar si esos rasgos son recuperables directamente de los vectores.

3.5. Descripción de la selección de características

[Pendiente completar]

3.6. Descripción de las métricas de evaluación

Para los clústeres, calculamos **purity** y **NMI** frente a las etiquetas; para el árbol, **accuracy** y analizamos las reglas de decisión con mayor información. Este enfoque pretende medir cuánto de la estructura lógica subyace en la geometría aprendida.

3.7. Descripción de los métodos estadísticos utilizados

[Pendiente completar]

4. Resultados y discusión

4.1. Presentación y análisis de resultados

[Pendiente completar con los resultados experimentales específicos]

4.2. Discusión de los resultados y su relevancia

[Pendiente completar]

4.3. Limitaciones y posibles mejoras

[Pendiente completar]

5. Conclusión

5.1. Resumen de los hallazgos principales

[Pendiente completar]

5.2. Conclusiones generales y relación con los objetivos

[Pendiente completar]

5.3. Recomendaciones para futuros trabajos

[Pendiente completar]

6. Bibliografía

6.1. Referencias bibliográficas citadas

1. Chen, X., & Gao, T. (2022). *Probing Linguistic Information for Logical Inference in Pre-trained Language Models*.
2. Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). *Visualizing and Measuring the Geometry of BERT*.
3. Cai, X., Wang, J., Peng, N., & Wang, X. (2021). *Isotropy in the Contextual Embedding Space: Clusters and Manifolds*.

6.2. Otras fuentes consultadas

[Pendiente completar]

7. Anexos

7.1. Código fuente utilizado en el análisis

El código completo del análisis está disponible en: [Repositorio GitHub - Enlace pendiente]

7.2. Tablas y gráficos adicionales

[Pendiente completar]

7.3. Otros materiales relevantes

[Pendiente completar]