Evaluación de modelos de lenguaje mediante valoración de argumentos en lenguaje natural

Simposio: Modelos de lenguaje y evaluación adversarial de argumentos

Mauro Santelli

En colaboración con Joaquín Toranzo Calderón y Ramiro Caso

REFERENTIA — IIF-SADAF-CONICET & FFyL-UBA & UTN-FRBA

III Congreso Iberoamericano de Argumentación (CIbA)
La Plata, 13 de marzo de 2026

Estructura de la charla

1. ¿Qué es NLI? Definición, conceptos clave, un poco de historia.
2. Ejemplos con un modelo BERT Razonamiento deductivo, inductivo, pragmática; qué funciona y qué no.
3. Problemas y limitaciones Sesgos superficiales, diagnósticos, limitaciones filosóficas; escenarios reales donde NLI importa.
4. Hacia adelante NLI como desafío normativo y descriptivo del razonamiento; benchmark en castellano.

NLI interpela tanto la evaluación de modelos como nuestras propias teorías sobre razonamiento, comprensión lingüística y racionalidad.

¿Qué es la Inferencia en Lenguaje Natural (NLI)?

Definición (Dagan y Glickman, 2004)

Dado un **texto** T y una **hipótesis** H, determinar si el significado de H puede inferirse del significado de T.

Clasificación tripartita:

- **Implicación** (*entailment*): T implica H
- **Contradicción**: T contradice H
- **Neutralidad**: T no dice nada determinante sobre H

La relación es **asimétrica** y opera sobre unidades léxico-sintácticas, sin representaciones semánticas formales. A diferencia de la implicación lógica clásica, NLI captura patrones de inferencia probabilística.

Conceptos clave de la formulación original

Patrones de implicación (entailment patterns)

Plantillas con variables que capturan relaciones inferenciales:

$$X \xleftarrow{\text{subj}} \text{buy} \xrightarrow{\text{obj}} Y \;\Rightarrow\; X \xleftarrow{\text{subj}} \text{own} \xrightarrow{\text{obj}} Y$$

Enfoque probabilístico

- Definición abstracta: determinista
- Modelo operativo: **fuzzy**
- Se asigna P de que T implique H

Objetivo general

- Semántica "superficial"
- Sin representaciones de significado
- ¿Cuánto se logra sin operar a nivel proposicional?

De la lógica natural a los transformers

1. **2004–2008**: Sistemas basados en reglas y lógica natural
2. **2015**: Dataset SNLI (570k ejemplos) — primera evaluación masiva
3. **2018**: Modelos BERT en SNLI/MultiNLI — >90 % de precisión
4. **2019–hoy**: Datasets adversariales (ANLI, HANS) revelan fragilidades

BERT (*Bidirectional Encoder Representations from Transformers*, Devlin et al. 2019): modelo de lenguaje pre-entrenado que genera representaciones contextuales bidireccionales. Se puede *afinar* (*fine-tune*) en tareas específicas como NLI, lo que lo convierte en un banco de pruebas accesible para evaluar capacidades inferenciales *tomadas* como correspondiendo a NLI por la comunidad de investigadores.

La paradoja del rendimiento

Precisión >90 % en benchmarks estándar, pero vulnerabilidades sistemáticas ante casos adversariales. ¿Razonamiento genuino o reconocimiento de patrones?

BERT fine-tuneado en MNLI: razonamiento deductivo

Veamos cómo se desempeña un modelo BERT[1] fine-tuneado en MNLI. Es un modelo débil comparado con ChatGPT o Claude, pero permite experimentar con la idea detrás de NLI.

P: *"If cats are nice, then the economy will do good. Cats are indeed nice."*
H: *"The economy will do good."*

BERT: **Entailment** — *86.59 %*

P: *"If cats are nice, then the economy will do good. The economy will not do good, after all."*
H: *"Cats are not nice."*

BERT: **Contradiction** — *80.18 % (debería ser implicación)*

P: *"Cats are nice and the economy is doing good."*
H: *"Cats are nice."*

BERT: **Entailment** — *97.98 %*

[1] Notebook de Colab

Inferencias materiales no monotónicas (I)

Las inferencias materiales son típicamente derrotables. El ejemplo de la torre de celular:

P: *"The cellphone tower emitted a calling signal to my phone."*
H: *"My cellphone will ring."*

BERT: **Entailment** — *73.30 %*

P: *"The cellphone tower emitted a calling signal to my phone. My phone is off."*
H: *"My cellphone will ring."*

BERT: **Neutral** — *47.95 %*

P: *"The cellphone tower emitted a calling signal to my phone,* **but** *my phone is turned off."*
H: *"My cellphone will ring."*

BERT: **Contradiction** — *55.19 %*

La conjunción adversativa "but" es leída como un marcador de disrupción inferencial, más allá del contenido proposicional.

Inferencias materiales no monotónicas (II)

Base

P: "John is a bird."
H: "John can fly."
BERT: Entailment — 85.50 %

+ Excepción con "but"

P: "John is a bird, **but** John is a penguin."
H: "John can fly."
BERT: Neutral — 60.32 %

+ Excepción con "and"

P: "John is a bird, **and** John is a penguin."
H: "John can fly."
BERT: Neutral — 68.33 %

+ Más información contextual

P: "John is a bird. We might add that John is a penguin and he's hurt."
H: "John can fly."
BERT: Entailment — 51.34 % (error)

"But" baja la confianza un 8 % más que "and". Pero con información contextual compleja, el modelo se confunde y vuelve a clasificar como implicación.

Razonamiento no deductivo: inducción y abducción peirceana

NLI no se limita a la deducción. Tres formas de razonamiento más débil:

P: *"John is a white swan. Peter is a white swan, also Mary is a white swan."*
H: *"Every swan is white."*

BERT: **Entailment** — *55.65 % (confianza baja, razonable)*

P: *"A swatch is a complex mechanism that implies it was designed by an intelligent watchmaker; life on earth is more complex than that."*
H: *"Life on earth was designed by an intelligence."*

BERT: **Entailment** — *57.80 %*

P: *"There was a loud noise on the kitchen, that's surprising. But if my cat would have thrown one of the glasses to the ground, the noise would be a matter of course."*
H: *"My cat might have thrown a glass in the kitchen."*

BERT: **Entailment** — *87.20 %*

Presuposición, verbos factivos e implicatura escalar

Presuposición

P: "The king of France stopped smoking."
H: "There is a king of France."
BERT: Entailment — 97.95 %

Verbo factivo

P: "John regrets buying the car."
H: "John bought a car."
BERT: Entailment — 99.08 %

Implicatura escalar (bien)

P: "Some students passed the exam."
H: "Not all students passed."
BERT: Entailment — 88.38 %

Implicatura escalar (mal)

P: "Mary can speak three languages."
H: "Mary can speak two languages."
BERT: Contradiction — 99.60 %
(debería ser *implicación*)

Quien habla tres lenguas habla dos: el modelo clasifica esto como contradicción. Necesitamos buenos criterios para identificar la naturaleza de estos errores.

Más fenómenos inferenciales ligados a la comprensión: Contrafácticos, temporalidad, cuantificadores y modalidad

Contrafáctico

P: "If Napoleon had won at Waterloo, European history would be different."
H: "Napoleon did not win at Waterloo."
BERT: Entailment — 98.11 %

Inferencia temporal (falla)

P: "The meeting starts at 3pm and lasts two hours."
H: "The meeting ends at 5pm."
BERT: Neutral — 98.24 %
(debería ser *implicación*)

Alcance de cuantificadores

P: "Every student read a different book."
H: "No two students read the same book."
BERT: Entailment — 98.68 %

Modalidad epistémica

P: "John must be home by now."
H: "The speaker believes John is home."
BERT: Entailment — 83.87 %

El modelo falla en aritmética temporal: no puede inferir que "empieza a las 3 y dura 2 horas" implica "termina a las 5".

Contextos opacos

Ejemplo Superman/Clark Kent: sensibilidad a la opacidad referencial en contextos de creencia.

P: *"John believes that Superman can fly."*
H: *"Clark Kent can fly according to John."*

BERT: **Entailment** — *97.72 %*

P: *"John believes that Superman can fly. John does not know that Superman is Clark Kent."*
H: *"Clark Kent can fly according to John."*

BERT: **Entailment** — *95.73 % (debería ser neutral o contradiction)*

P: *"John believes that Superman can fly.* **Although**, *John does not know that Superman is Clark Kent."*
H: *"Clark Kent can fly according to John."*

BERT: **Entailment** — *95.93 % (no cambia)*

El modelo ignora la opacidad de los contextos de creencia. "Although" no produce efecto — a diferencia de "but" en las inferencias no monotónicas.

Sesgos de aprendizaje superficial reconocidos en la literatura

Sesgo de superposición léxica (McCoy et al., 2019): el modelo confunde implicación con superposición de palabras entre T y H. "Juan ama a María" / "María es amada por Juan" → 98.83 % implicación (correcto). Pero: "El juez condenó al asesino" / "El asesino condenó al juez" — superposición léxica idéntica, relación semántica opuesta.

Sesgo de hipótesis aislada (Poliak et al., 2018): los modelos aprenden que "no" y "nunca" en H indican contradicción, **ignorando completamente** el texto T.

Metodología HANS: McCoy et al. identifican tres heurísticas explotadas — **Lexical Overlap**, **Subsequence** y **Constituent**. Los modelos obtienen resultados cercanos al azar cuando estas heurísticas no aplican (lo que permite suponer que no *razonan* sino que identifican otros patrones que coinciden con los resultados de estos razonamientos).

Datasets de diagnóstico

HELP (Yanaka et al., 2019)

- Foco en **monotonicidad** lingüística
- Evalúa inferencias monotónicas y antimonotónicas
- Revela fallas sistemáticas

ANLI (Nie et al., 2020)

- Dataset **adversarial**
- Human-and-model-in-the-loop
- Tres rondas de dificultad creciente

Diagnóstico general

Los datasets estándar (SNLI, MNLI) **fomentan** el aprendizaje de patrones superficiales. Los modelos no desarrollan capacidades genuinas de razonamiento semántico. Se necesitan benchmarks que evalúen fenómenos lingüísticos específicos.

¿Qué nos dicen estos resultados?

NLI, tal como lo capturan MNLI y SNLI al entrenar modelos como BERT, cubre *todas* las formas de inferencia tradicionales:

- **Deducciones**: Modus Ponens, eliminación de conjunción, contrafácticos
- **Inducciones**: enumeración incompleta, analogía
- **Abducciones**: inferencia a la mejor explicación

Paradoja: las que **peor maneja** son precisamente las inferencias **deductivas** — Modus Tollens se clasifica como contradicción en vez de implicación.

¿Está bien o mal que NLI cubra todo tipo de inferencia? NLI **debería** incluir cualquier tipo de inferencia *buena* (en algún sentido teórico razonable). El problema no es el alcance — es que los modelos no lo hacen de forma confiable. Lo que no tenemos, y aquí los filósofos y otros especialistas en argumentación deberíamos intentar ofrecer, son definiciones precisas de esa bondad inferencial general ligada a la **comprensión lingüística**.

Tres limitaciones fundamentales

1. Simplificación de la noción de consecuencia

La formulación estándar ignora relaciones de implicación **derrotables**, **graduales** e **intensionales**. La consecuencia se reduce a un "sí/no" que no captura la riqueza de las relaciones inferenciales reales.

2. Desatención a fenómenos pragmáticos

Implicaturas conversacionales y escalares, presuposiciones y acomodación, aserciones precavidas (*hedged assertions*), y la distinción entre implicación semántica e inferencia pragmática son aspectos poco tematizados o directamente ignorados.

3. Falta de integración filosófica

Benchmarks robustos requieren integrar mejores modelos de semántica y su interacción con la dimensión inferencial del significado, pragmática filosófica y lógica no monotónica—entre otros. Sin esta integración, los benchmarks seguirán teniendo problemas para evaluar *verdadera* comprensión lingüística.

Sugerencias para anclar NLI en el habla cotidiana

Conversaciones sobre significado

Clara le dice X a Ana y Bruno. Ana le cuenta X a Daniel. Bruno dice: "Clara no dijo eso", o "dijo lo contrario", o "sí lo dijo". ¿Quién tiene razón? Determinar si el reporte de Ana es fiel al enunciado de Clara es una tarea de NLI.

Resúmenes

Ana confecciona un resumen del texto de Bruno. Ana dice que el texto dice "T", pero Bruno desacuerda: "T no representa bien lo que escribí". Evaluar la fidelidad del resumen es NLI.

Discusiones interpretativas

Ana lee "T", extrae la hipótesis "H". Clara lee tanto "H" como "T" y juzga si "H" está implicado por T. Esto es **exactamente** NLI, y lo hacemos constantemente — en la lectura, en la conversación, en la argumentación.

Hacia un benchmark para el castellano

Nuestro proyecto

1. **Fundamentación conceptual** de NLI desde la filosofía (esta charla)
2. **Curado de un corpus** de contextos en castellano (Ramiro Caso)
3. **Diseño de competencias adversariales** humano-modelo (Joaquín Toranzo Calderón)

¿Por qué el castellano? El campo tiene un sesgo masivo hacia el inglés. El castellano tiene particularidades propias: modo subjuntivo, orden flexible de palabras, implicaturas culturalmente situadas.

Objetivo: benchmarks que evalúen fenómenos lingüísticos específicos y requieran comprensión genuina — no atajos heurísticos.

Conclusiones

- NLI es una tarea paradigmática para evaluar comprensión lingüística
- Los modelos actuales explotan patrones superficiales — HANS y ANLI lo demuestran
- NLI cubre deducciones, inducciones y abducciones, pero no de forma confiable
- Tres limitaciones señalan un camino: consecuencia simplificada, pragmática ignorada, falta de integración filosófica

Estas observaciones sugieren que el diseño de benchmarks robustos para NLI se beneficiaría de integrar perspectivas de la semántica inferencial, la pragmática filosófica y la lógica no monotónica. Las charlas de Caso y Toranzo Calderón abordan los próximos pasos: curado de corpus y diseño de competencias adversariales.

Bibliografía

- Bowman, S. et al. (2015). "A large annotated corpus for learning NLI." *EMNLP*.
- Dagan, I. & Glickman, O. (2004). "Probabilistic Textual Entailment." *PASCAL Workshop*.
- McCoy, R. T., Pavlick, E. & Linzen, T. (2019). "Right for the Wrong Reasons." *arXiv:1902.01007*.
- Nie, Y. et al. (2020). "Adversarial NLI." *Proc. ACL 2020*.
- Poliak, A. et al. (2018). "Hypothesis Only Baselines in NLI." *arXiv:1805.01042*.
- Williams, A., Thrush, T. & Kiela, D. (2022). "ANLIzing the ANLI Dataset." *Proc. SCiL 2022*.
- Yanaka, H. et al. (2019). "HELP: Identifying Shortcomings in Monotonicity Reasoning." *arXiv:1904.12166*.
- Yanaka, H. et al. (2019). "Can Neural Networks Understand Monotonicity Reasoning?" *arXiv:1906.06448*.
- Grice, H. P. (1975). "Logic and Conversation." In *Syntax and Semantics*, vol. 3, pp. 41–58.
- Stalnaker, R. (1978). "Assertion." In *Syntax and Semantics*, vol. 9. / Stalnaker, R. (2002). "Common Ground." *Linguistics and Philosophy*, 25(5-6), 701–721.
- Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press.

Discusión

- ¿Qué fenómenos del castellano priorizar en un benchmark NLI?
- ¿Debería NLI incluir toda inferencia "buena", o solo implicación estricta?
- ¿Cómo distinguir comprensión genuina de reconocimiento de patrones?
- ¿Qué otras habilidades deberíamos evaluar en relación con la capacidad de razonar en lenguaje natural?

Para quienes quieran jugar con el modelo de ejemplo:

Gracias

Mauro Santelli · mesantelli@uba.ar

REFERENTIA — IIF-SADAF-CONICET & FFyL-UBA

En colaboración con Joaquín Toranzo Calderón y Ramiro Caso