ESPANLI: An Adversarial NLI Benchmark for Río de la Plata Spanish via Human–Model Competitions

Mauro Santelli[2,3], Joaquín Toranzo Calderón[1,2,3], Ramiro Caso[2,3], and Bruno Muntaabski[2]

[1] Universidad Tecnológica Nacional, FRBA, Grupo de IA y Robótica
[2] Universidad de Buenos Aires, FFyL
[3] Instituto de Investigaciones Filosóficas (SADAF–CONICET)

Abstract. Natural Language Inference (NLI) benchmarks such as SNLI, MultiNLI and ANLI are standard tools for evaluating language models, but they are overwhelmingly English and increasingly solvable through surface heuristics or prior exposure. Spanish evaluation, and in particular the Río de la Plata variety, remains underrepresented; existing datasets are typically translations that distort lexical overlap, truth conditions and pragmatic context. We describe the design of ESPANLI, an adversarial NLI benchmark for Spanish built from doctoral theses in public repositories of Argentinian universities, with rioplatense phenomena annotated by expert curators. Its core is a human-in-the-loop competition in which trained generators craft instances that defeat reference models and a verification committee retains only majority-validated cases. Expected contributions include a public adversarial corpus, an interactive generation platform, and expert-level NLI sub-tasks native to Spanish.

Keywords: natural language inference, adversarial benchmark, language models, Spanish NLP, Río de la Plata Spanish, human-in-the-loop evaluation

ESPANLI: un benchmark adversarial de NLI para el español rioplatense mediante competencias humano–modelo

Resumen Los benchmarks de Inferencia en Lenguaje Natural (NLI) como SNLI, MultiNLI y ANLI son hoy herramientas estándar para evaluar modelos de lenguaje, pero su cobertura es mayoritariamente inglesa y sus ejemplos resultan crecientemente resolubles por atajos heurísticos o memorización. La evaluación en español, y en particular en la variante rioplatense, está sub-representada; los conjuntos disponibles suelen ser traducciones que distorsionan la superposición léxica, las condiciones de verdad y el contexto pragmático. Describimos el diseño de ESPANLI, un benchmark adversarial de NLI para el español construido a partir de tesis doctorales alojadas en repositorios públicos de universidades argentinas,

con fenómenos rioplatenses anotados por curadores especialistas. Su núcleo es una competencia humano-en-el-bucle en la que generadores con formación específica producen instancias que hacen fallar a modelos de referencia y una comisión de verificación conserva sólo los casos validados por mayoría. Las contribuciones previstas incluyen un corpus adversarial público, una plataforma interactiva de generación y un conjunto de sub-tareas de NLI de nivel experto nativas en castellano.
Palabras clave: inferencia en lenguaje natural, benchmark adversarial, modelos de lenguaje, PLN en español, español rioplatense, evaluación humano-en-el-bucle

1. Introducción

La Inferencia en Lenguaje Natural (NLI) sigue siendo una tarea informativa para evaluar a los modelos de lenguaje actuales (Madaan et al., 2025). Aunque los modelos de lenguaje de gran tamaño (LLMs) superan hoy el 90 % en los principales benchmarks de NLI (Bowman et al., 2015; Devlin et al., 2019), esa solvencia, sin embargo, convive con una paradoja: los mismos modelos fallan de forma sistemática ante instancias adversariales y se apoyan en atajos superficiales —superposición léxica, sesgo de hipótesis aislada, inserción de negaciones— que no requieren comprensión inferencial (McCoy et al., 2019; Nie et al., 2020; Poliak et al., 2018). El problema se agudiza en castellano: los benchmarks de NLI disponibles para el español son en su mayoría traducciones directas del inglés, estrategia que degrada la superposición léxica entre premisas e hipótesis y arrastra un sesgo pragmático anglocéntrico ajeno al hablante hispanohablante.

El proyecto ESPANLI se propone abordar esta doble carencia —fragilidad adversarial y sub-representación del castellano, especialmente de la variante rioplatense— mediante el desarrollo de un benchmark construido desde cero para el español, siguiendo una estrategia adversarial humano-en-el-bucle. En lo que sigue describimos brevemente los antecedentes (§2), la propuesta de curado del corpus (§3), la metodología de competencias adversariales humano–modelo (§4), y los resultados esperados y líneas futuras (§5).

2. Antecedentes

2.1. La tarea de NLI y su formulación estándar

Siguiendo a Dagan y Glickman (2004), la tarea de NLI consiste en determinar, dadas una premisa P y una hipótesis H, si el significado de H se infiere del de P. La clasificación canónica distingue *implicación* (entailment), *contradicción* y *neutralidad*. Es una tarea asimétrica que opera sobre unidades léxico-sintácticas más que sobre representaciones semánticas formales, y por eso captura patrones probabilísticos de inferencia humana que la lógica formal no alcanza.

2.2. Saturación y atajos heurísticos de los benchmarks estándar

Los conjuntos tradicionales —SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018)— se construyeron por crowdsourcing y exhiben artefactos estadísticos sistemáticos: contradicciones por negación, implicaciones por borrado de palabras y neutralidad por inserción de modificadores irrelevantes (Poliak et al., 2018). McCoy et al. (2019) muestran además que los modelos explotan heurísticas de *lexical overlap*, *subsequence* y *constituent* que correlacionan con la etiqueta en entrenamiento pero se quiebran en casos construidos ad hoc. ANLI (Nie et al., 2020; Williams et al., 2022) y HANS (McCoy et al., 2019) respondieron con estrategias adversariales, pero siguen concentradas en el inglés.

2.3. El problema del castellano

Para el español, la estrategia dominante ha sido traducir benchmarks existentes (Conneau et al., 2018) con la excepción de Kovatchev y Taulé (2022), con dos efectos acoplados: el texto traducido distorsiona la superposición léxica, la polisemia y las condiciones de verdad de los pares originales, y transfiere un sesgo pragmático anglocéntrico que hace fallar al modelo por falta de conocimiento culturalmente situado, no por falta de competencia inferencial. Iniciativas como Somos 600M (Grandury, 2024) subrayan la necesidad de recursos nativos, pero la cobertura adversarial sigue siendo escasa.

3. Propuesta: curado de un corpus adversarial en castellano

ESPANLI toma sus materiales textuales de tesis doctorales alojadas en repositorios públicos de universidades argentinas. La cobertura inicial es Filosofía y Humanidades, con extensiones previstas en versiones sucesivas a disciplinas con fuerte contenido formal (Matemática, Ingeniería) y a Derecho. El corpus se estructura como tuplas *(contexto, hipótesis, etiqueta)*: el contexto se toma de párrafos etiquetados del corpus base y la hipótesis se construye de manera adversarial. La selección de contextos, cuando es posible, prioriza rasgos que explotan sesgos conocidos de los modelos de NLI —vocabulario polisémico, marcas argumentales (conectores lógicos, adversativos, concesivos) y complejidad textual—. La cobertura lingüística apunta a fenómenos del castellano rioplatense que los benchmarks traducidos no capturan, como el léxico o modos verbales locales.

A pesar de que el corpus elegido maneja un registro formal o culto, estos fenómenos pueden encontrarse igualmente. De este modo, partir de materiales nativos y culturalmente anclados, en lugar de traducir SNLI o ANLI, es la principal diferencia metodológica respecto de los conjuntos de NLI existentes para el español. Una estimación preliminar pesimista de las tesis doctorales disponibles en la Facultad de Filosofía y Letras de la Universidad de Buenos Aires sugiere que el corpus base podría alcanzar el millón de párrafos[4] aunque se realizarán rondas de curación de 5000 párrafos por ronda.

[4] Se dispone de cerca de 1500 tesis, 1200 de las cuales están en formato digital nativo, con un promedio de 300 páginas con texto utilizable y 3 párrafos por página.

4. Metodología: competencias adversariales humano–modelo

El núcleo operativo de ESPANLI es un protocolo humano-en-el-bucle inspirado en ANLI (Nie et al., 2020) en el que los anotadores actúan como *adversarios* del modelo: producen hipótesis coherentes con la etiqueta objetivo (I para "implicada", C para "contradictoria", y N para "neutral") pero capaces de confundirlo. El bucle tiene cuatro componentes:

Generadores adversariales. Participantes seleccionados por formación académica que reciben un contexto y una etiqueta objetivo (I, C, N). Deben producir una hipótesis coherente con esa etiqueta bajo juicio humano pero que haga fallar al modelo de referencia. Pueden rechazar instancias justificando la razón tras agotar un número acotado de iteraciones, a definir en pruebas piloto. Se pretende motivar a los generadores con estrategias de gamificación en competencias regulares.

Modelo(s) de base. Cada instancia se evalúa contra uno o más modelos de referencia. Si la clasifican correctamente, el generador reformula la hipótesis sobre las mismas premisas; si fallan, la instancia pasa a verificación. Se pretende emplear modelos de base que se correspondan con el estado del arte para el momento en el que se realice cada competencia.

Verificadores. Una comisión de tres personas formadas en análisis de argumentos juzga por mayoría si la etiqueta se corresponde con la relación inferencial pretendida. Sólo los casos con acuerdo mayoritario se incorporan al corpus; los demás se descartan.

Instrucciones. Cada tipo de etiqueta (I, C, N) recibe una consigna específica. La redacción de las consignas es en sí misma un parámetro a controlar: las pruebas piloto variarán formulaciones para identificar las más efectivas y usarlas en las competencias posteriores.

Aunque no buscamos entrenar un modelo, como en ANLI, el benchmark generado involucrará iteraciones periódicas con los fines de actualizarlo y anular la saturación que pueda producirse entre cada par de iteraciones.

5. Resultados esperados y trabajo futuro

Las contribuciones previstas son cuatro: (i) un subcorpus académico adversarial derivado de tesis doctorales en repositorios públicos; (ii) una plataforma web interactiva para la generación abierta de instancias, vinculada al corpus; (iii) un protocolo de gamificación periódica con estudiantes universitarios como sujetos adversariales, supervisada por especialistas y organizada por áreas de formación para mitigar la saturación; y (iv) sub-tareas de NLI de nivel experto nativas en castellano, co-diseñadas por curadores y anotadores disciplinares (filosófica, formal, jurídica) para evaluar razonamiento experto fuera del alcance de los benchmarks traducidos de propósito general. La evaluación de modelos de base y populares sobre el corpus, así como la incorporación prevista de fuentes no académicas, serán objeto de publicaciones separadas.

Las pruebas piloto permitirán refinar los parámetros operativos: iteraciones por instancia, selección de modelo(s) de base, formulación de consignas por etiqueta y ponderación de rasgos en la selección de contextos.

6. Conclusión

ESPANLI busca cubrir una doble brecha de la evaluación actual: la fragilidad adversarial de los benchmarks saturados y la sub-representación del castellano rioplatense y experto. Al combinar fundamentación conceptual, curado nativo de corpus y competencias humano–modelo con verificación, apunta a un recurso que no sólo mida el desempeño de los modelos sino que permita caracterizar *qué* resuelven cuando aciertan, qué "habilidades" poseen y *cómo* fallan cuando no.

Declaración sobre el uso de herramientas de IA. Durante la preparación de este trabajo, los autores utilizaron Claude Opus 4.8 (Anthropic), a través de Claude Code, con el fin de mejorar el lenguaje y la legibilidad del manuscrito, así como para asistir en su composición y formateo en LaTeX. Tras utilizar esta herramienta, los autores revisaron y editaron el contenido según fuera necesario y asumen plena responsabilidad por el contenido de la publicación.

Bibliografía

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632-642. https://doi.org/10.18653/v1/D15-1075

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating Cross-lingual Sentence Representations. En E. Riloff, D. Chiang, J. Hockenmaier & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2475-2485). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1269

Dagan, I., & Glickman, O. (2004). Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. *Learning Methods for Text Understanding and Mining*, *2004*(26-29), 2-5.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Grandury, M. (2024). The #Somos600M Project: Generating NLP Resources That Represent the Diversity of the Languages from LATAM, the Caribbean, and Spain. https://doi.org/10.48550/arXiv.2407.17479

Kovatchev, V., & Taulé, M. (2022). InferES : A Natural Language Inference Corpus for Spanish Featuring Negation-Based Contrastive and Adversarial Examples. En N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond & S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 3873-3884). International Committee on Computational Linguistics.

Madaan, L., Esiobu, D., Stenetorp, P., Plank, B., & Hupkes, D. (2025). Lost in Inference: Rediscovering the Role of Natural Language Inference for Large Language Models. En L. Chiruzzo, A. Ritter & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 9229-9242). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.466

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. En A. Korhonen, D. Traum & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428-3448). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1334

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020). Adversarial NLI: A New Benchmark for Natural Language Understanding. En D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4885-4901). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.441

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis Only Baselines in Natural Language Inference. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 180-191. https://doi.org/10.18653/v1/S18-2023

Williams, A., Nangia, N., & Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. En M. Walker, H. Ji & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1112-1122). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Williams, A., Thrush, T., & Kiela, D. (2022). ANLIzing the Adversarial Natural Language Inference Dataset. En A. Ettinger, T. Hunter & B. Prickett (Eds.), *Proceedings of the Society for Computation in Linguistics 2022* (pp. 23-54). Association for Computational Linguistics.