UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS

Centro de Investigación en Ciencias

"Propuesta de Marco Metodológico para Modelar la Abstracción Semántica en Inteligencia Artificial: Aplicación a la Tarea del RIT"

TESIS PROFESIONAL PARA OBTENER EL GRADO DE:

Doctor en Ciencias

PRESENTA:

David Torres Moreno

DIRECTOR: Dr. Jorge Hermosillo Valadez

CUERNAVACA, MORELOS Diciembre, 2025

Instituto de Investigación en Ciencias Básicas y Aplicadas

Índice general

Índice de cuadros

Índice de figuras

Resumen

Un área fundamental y de gran oportunidad en la Inteligencia Artificial (IA) es dotar a las máquinas de una comprensión lingüística profunda, que vaya más allá del procesamiento superficial de palabras para abarcar sus relaciones semánticas, el contexto y el significado subyacente. Este objetivo enfrenta importantes desafíos teóricos, técnicos y metodológicos.

Aunque los Modelos de Lenguaje Preentrenados (PLMs, por sus siglas en inglés: Pretrained Language Models) y, especialmente, los Modelos de Lenguaje Grandes (LLMs, por sus siglas en ingles: Large Language Models) como BERT y GPT han logrado avances notables en tareas como traducción y generación de texto, su comprensión es esencialmente limitada. Desde la perspectiva de la semántica psicológica, estos sistemas adolecen de carencias fundamentales ya que su conocimiento se deriva principalmente de patrones estadísticos extraídos de grandes corpus, sin un anclaje conceptual a las creencias, intenciones y al mundo real que caracterizan la cognición humana. Esto se traduce en problemas como la generación de alucinaciones, la perpetuación de sesgos y una notable falta de razonamiento abstracto y profundo. Un problema fundamental es que, al trabajar con relaciones léxicas sin procesar, los LLMs tienden a atascarse en asociaciones literales, sin captar la profundidad semántica. Para superar estas limitaciones, las investigaciones actuales se han enfocado en dos líneas principales: 1) La integración de conocimiento estructurado externo (como bases de datos ontológicas) para enriquecer la representación semántica y 2) La mejora de los mecanismos de razonamiento mediante técnicas como el debate entre múltiples agentes o prompting avanzado.

Sin embargo, estas aproximaciones no han logrado incorporar de manera efectiva la abstracción semántica, que es una habilidad humana clave que permite construir representaciones conceptuales flexibles y jerárquicas, trascendiendo los patrones textuales superficiales. Modelar esta capacidad es, por tanto, la pieza faltante para alcanzar una comprensión del lenguaje más robusta y genuina.

En este contexto, esta investigación propone y evalúa un Marco Metodológico de Abstracción de Conocimiento Semántico (SKA). Su innovación central unifica las líneas de investigación:

- Abstracción Semántica: En lugar de usar relaciones léxicas sin procesar, el SKA estructura el conocimiento semántico en categorías abstractas de compatibilidad e incompatibilidad semántica. Esto guía al LLM a formar nuevos conceptos y conexiones, evitando atajos basados en pistas superficiales y fomentando un razonamiento más profundo.

- Mecanismo de razonamiento (SKA_DT): El SKA emplea un árbol de decisión jerárquico explícito para llegar a una decisión final. Este mecanismo identifica patrones en las líneas de razonamiento propiciadas con diferentes grupos de abstracción, revelando cómo cubren las lagunas de conocimiento del modelo. A diferencia de estrategias basadas en debate entre agentes, el SKA_DT garantiza un mayor control, fiabilidad y consistencia sistemática en las respuestas, fortaleciendo al agente individual sin depender del costo de modelos extremadamente grandes.

La tarea elegida para evaluar la propuesta es el Reconocimiento de la Implicatura Textual (RIT), la cual exige inferir significados no explícitos y relaciones semánticas, constituyendo así una prueba rigurosa de comprensión profunda. Los resultados en benchmarks como *SuperGLUE* revelan que, incluso los modelos de mayor rendimiento, presentan un desempeño alto en esta tarea, pero su rendimiento cae al cambiar de corpus, evidenciando las limitaciones actuales.

La implementación del marco SKA demostró compensar efectivamente las brechas de conocimiento semántico en los LLMs. Los resultados mostraron mejoras sustanciales en exactitud (superiores al 15 % en algunos modelos), particularmente en la clase de no implicatura. Es así que el marco logra:

- Controlar y estructurar el conocimiento a un nivel de abstracción superior.

- Guiar líneas de razonamiento coherentes y sistemáticas.

- Revelar y compensar lagunas en el conocimiento previo del modelo de manera fiable.

En conclusión, la abstracción semántica, implementada a través del marco metodológico SKA con su árbol de decisión, se confirma como un enfoque prometedor y eficiente. Este paradigma dota a los LLMs de una mayor flexibilidad cognitiva y una capacidad de razonamiento más profunda y sistemática en tareas complejas de comprensión lingüística, superando limitaciones de estrategias actuales y sin incurrir necesariamente en el alto costo de modelos de escala industrial.

Capítulo 1

Introducción

1.1. Antecedentes

Un problema fundamental y área de oportunidad para el campo de la Inteligencia Artificial (IA) es modelar la comprensión lingüística. Queremos que las máquinas puedan procesar el lenguaje de manera similar a los humanos, comprendiendo no solo las palabras, sino también su significado, contexto y relaciones entre ellas. Sin embargo, modelar la comprensión lingüística en una computadora implica desafíos; principalmente teóricos, técnicos y metodológicos importantes.

Los dos niveles principales de análisis lingüístico son el sintáctico y el semántico. La sintaxis aborda la estructura de la oración: las relaciones entre palabras, el orden de los constituyentes de la oración y sus funciones sintácticas. Para saber si una frase es gramatical o no, recurrimos a reglas que determinan la función de cada constituyente dentro de la oración. La semántica aborda el significado de las palabras, los constituyentes y cómo se compone el significado de la oración. De esta forma, sintaxis y semántica van de la mano en la descripción de la gramática de una lengua.

Los PLMs y LLMs son una muestra del avance en este campo, permiten la generación automática de texto y son capaces de realizar tareas complejas como traducir entre idiomas, generar texto e incluso responder preguntas con cierta precisión, pero también pueden generar respuestas imprecisas o incluso incorrectas si el contexto es complejo o ambiguo. Los PLMs y LLMs comparten una naturaleza común: ambos se fundamentan en el paradigma del pre-entrenamiento. Los primeros sentaron las bases técnicas para tareas específicas, mientras que los segundos escalaron masivamente estos principios, desarrollando capacidades emergentes como el razonamiento complejo y la generación versátil.

Si bien es cierto que actualmente existen esfuerzos de investigación de punta por alcanzar un

grado de sofisticación muy elevado en el modelado de habilidades lingüísticas generales por parte de una máquina: BERT (Devlin y cols., 2018), XLNet (Z. Yang y cols., 2019), GPT-1 (Radford y Narasimhan, 2018), GPT-2 (Radford y cols., 2019), GPT-3 (Brown y cols., 2020) y GPT-4 (OpenAI y cols., 2024), también es cierto que estos modelos de lenguaje han sido fuertemente criticados. Se ha argumentado en la literatura que distan mucho de "comprender" el lenguaje como lo hacemos los humanos y de ser opacos en términos de dar cuenta de sus decisiones y representaciones (Marcus y cols., 2017).

En este sentido, modelar la comprensión lingüística en una máquina implica tomar postura respecto de lo que significa que una máquina "entienda" y tomar decisiones respecto de cómo representar los distintos elementos que entran en juego y cómo ponerlos en relación para los fines de atender una tarea específica.

Desde el marco teórico de la semántica psicológica, que estudia cómo los humanos representamos y combinamos significados, los sistemas actuales de Procesamiento de Lenguaje Natural (PLN) son modelos bastante exitosos en determinar la similitud de las palabras, desambiguar su sentido y ejecutar tareas complejas. Sin embargo, estas competencias coexisten con carencias fundamentales que limitan su capacidad para emular una comprensión semántica humana. Los modelos actuales están demasiado vinculados a los patrones basados en grandes corpus y demasiado poco vinculados a los deseos, objetivos y creencias que las personas expresan a través de las palabras (Lake y Murphy, 2023).

Sin embargo, las respuestas de los LLMs heredan sentido y referencia (significado) de los datos de entrenamiento y estas no son construidas por accidente. Este significado no depende del lector. Es así que, (Lederman y Mahowald, 2024) sugieren a través del interpretacionismo y el problema de la nueva referencia, que la atribución de creencias, deseos e intenciones, pueden ser más útiles para predecir y explicar el comportamiento de los LLMs que una descripción en términos de predictores estadísticos complejos.

Los LLMs han demostrado algún grado de capacidad en *comprensión, conocimiento, razonamiento y cálculo* (W. Zhong y cols., 2024; Zhao y cols., 2025). A pesar de su excelente rendimiento en muchas tareas de PLN, los LLMs presentan problemas críticos (Figura 1.1). Se ha demostrado, por ejemplo, que el conocimiento deficiente o las lagunas evidentes en los datos de entrenamiento tienen un impacto en el fenómeno de las alucinaciones en los LLMs (Manakul y cols., 2023; Rawte y cols., 2023; Dhuliawala y cols., 2024). Además, si el proceso de entrenamiento se contamina con datos de evaluación de tareas, se crea un espejismo de competencia, evidenciado por modelos que memorizan ejemplos específicos pero fallan en tareas nuevas (McCoy y cols., 2019; Jin y cols., 2020; C. Li y Flanigan, 2024), o reproducen problemas heredados de los datos humanos, como sesgos sociales y artefactos de anotación, es decir, palabras clave que permiten utilizar para la tarea incluso fuera de contexto (Proebsting y Poliak, 2025). Para compensar las lagunas de conocimiento, se ha demostrado que pro-

porcionar grafos de conocimiento externos puede mejorar la calidad de las representaciones incrustadas (D. Yang y cols., 2022), mitigar las alucinaciones (J. Li y cols., 2023) y ayudar a responder mejor a preguntas con grandes cantidades de datos (Dai y cols., 2025).

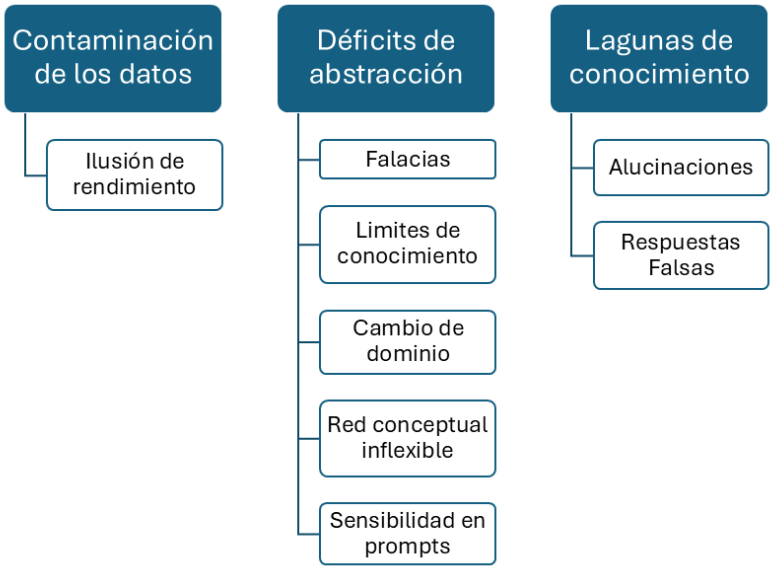


Figura 1.1: Problemas de los LLMs en las tareas del PLN

Sin embargo, estos avances se han visto desafiados por la naturaleza compleja del lenguaje humano, un factor que afecta directamente a las capacidades de los LLMs. El desafío reside en la ambigüedad del lenguaje humano. Un mismo significado puede expresarse mediante múltiples formas (variabilidad), y las frases suelen contener significados implícitos y relaciones lógicas no explícitas. Además, la ambigüedad léxica, estructural y semántica complica aún más la interpretación automática. Frente a esto, la investigación actual cuestiona si la mera exposición a grandes volúmenes de datos textuales y el aumento de parámetros son suficientes para dotar a los LLMs de una comprensión profunda, más allá de su capacidad demostrada en tareas específicas.

Estudios como el de (Cao y cols., 2025) revelan que estos modelos adquieren un conocimiento limitado e insatisfactorio de relaciones semánticas fundamentales (como la hiponimia o la sinonimia), a menudo mostrando sesgos y un desempeño muy inferior al humano. Esta limitación se manifiesta en problemas concretos. Por un lado, los LLMs pueden carecer de conocimiento de fondo, logrando responder correctamente a una pregunta principal sin poseer la información semántica relevante que lo sustente (Sahu y cols., 2022). Por otro lado, su proceso de razonamiento puede ser frágil e inconsistente. Se ha destacado la imposibilidad de estos modelos para realizar razonamientos abstractos (Xiong y cols., 2024; S. Wang y cols., 2024; Q. He y cols., 2025). Si bien hay indicios de que los PLMs reconocen las diferencias entre las relaciones hipónima-hiperónima en los sustantivos (Regneri y cols., 2024), se

han observado limitaciones cognitivas de nivel superior (Peng y cols., 2022). Además, los LLMs carecen de habilidades lingüísticas funcionales y formales (Mahowald y cols., 2024) y muestran limitaciones fundamentales en el razonamiento abstracto y la planificación (Lee y cols., 2025).

Para abordar estas deficiencias, la comunidad científica ha desarrollado estrategias para mejorar las capacidades cognitivas y semánticas de los LLMs. Esta discusión se centrará en estos dos principales caminos de investigación: la integración de conocimiento estructurado externo y mecanismos de razonamiento. La primera busca integrar conocimiento estructurado externo, como bases de conocimiento ontológicas (por ejemplo, ConceptNet), para proporcionar a los modelos información semántica y de sentido común explícita (Agrawal y cols., 2024; Ilievski y cols., 2021). No obstante, esta integración plantea sus propios retos, como la incompletitud y el ruido de los recursos externos, y la sensibilidad de los LLMs a datos corruptos (Dai y cols., 2025). La segunda línea se ha centrado en mejorar los mecanismos de razonamiento de los LLMs mediante técnicas avanzadas de prompting y consolidación de respuestas. Métodos como Chain-of-Thought (CoT) (Wei y cols., 2022), Tree-of-Thought (ToT) (Yao y cols., 2023), Abstraction-of-Thought (AoT) (Hong y cols., 2024), Autoconsistencia (X. Wang y cols., 2023) y estrategias de votación dinámica (Xue, Liu, Lei, Ren, y cols., 2023; Dogan y Birant, 2019a) han demostrado mejorar la robustez y la coherencia del razonamiento paso a paso, aunque a menudo con un costo computacional grande. En este sentido, como señalan (L. Chen y cols., 2024), el rendimiento de un LLM no mejora monotónicamente con realizar más consultas adicionales en tareas complejas, y su fiabilidad para evaluar textos o emitir juicios puede no capturar los matices y criterios múltiples del razonamiento humano experto.

Estas dos áreas de investigación se han convertido en un punto clave para mejorar la comprensión y el procesamiento del lenguaje por parte de los modelos de lenguaje. A pesar de estos avances, los LLMs actuales, en su arquitectura predominante, reducen la semántica a correlaciones estadísticas entre palabras, careciendo de un anclaje conceptual a la experiencia y al mundo real que caracteriza la cognición humana (Lake y Murphy, 2023).

En vista de estas limitaciones, se destaca una necesidad urgente por cambiar la perspectiva sobre cómo abordamos el PLN con estos modelos. Se requiere un nuevo paradigma que no solo genere texto coherente sino también modele y procese lenguaje a nivel conceptual, similar al modo en que lo hace una mente humana (L. Chen y cols., 2024). La capacidad para ir más allá del patrón textual y la coincidencia de hechos, para construir representaciones conceptuales flexibles y jerárquicas del mundo es la abstracción semántica.

1.2. Planteamiento del problema

La abstracción semántica es una habilidad fundamental para la comprensión y el razonamiento sobre texto, ya que permite a los seres humanos identificar y representar conceptos y relaciones abstractas que subyacen a una frase. Sin embargo, esta habilidad aún no se ha incorporado de manera efectiva en los LLMs. Estos modelos tienden a depender de señales superficiales y redes conceptuales rígidas, lo que limita su flexibilidad cognitiva y capacidad para manejar relaciones semánticas complejas. Por lo tanto, es necesario desarrollar un mecanismo que proporcione una Abstracción de Conocimiento Semántico para identificar y compensar estas brechas de conocimiento, mejorando tanto la precisión como la robustez del razonamiento de los LLMs. Para evaluar este marco metodológico, se requiere de una tarea que requiera comprensión lingüística.

La tarea del Reconocimiento de la Implicatura Textual (RIT) exige que un modelo no solo procese literalmente la información, sino que infiera significados no explícitos, relaciones semánticas e intenciones detrás de las oraciones. La implicatura, al requerir la interpretación de lo que se sugiere o da por sentado en un enunciado, pone a prueba la comprensión profunda y el razonamiento del modelo, yendo más allá de la coherencia textual. Esta tarea permite una evaluación más robusta y humana, ya que detecta si el modelo puede aplicar lógica a información incompleta y evitar respuestas sesgadas o literales que fallen en captar el sentido implícito de la tarea, revelando así sus verdaderas capacidades y limitaciones cognitivas.

Desde enero de 2023, el benchmark *Natural Language Inference on RTE*[1] es liderado por *Vega V2* (Q. Zhong y cols., 2022) con un 96 % de precisión, seguido por *PALM* (Chowdhery y cols., 2022) con un 95.7 %. La diferencia en rendimiento es mínima, pero abismal en número de parámetros: 6B[2] frente a 540B. Sin embargo, al evaluar a *Vega V2* en el corpus de diagnóstico de *SuperGLUE* (A. Wang, Pruksachatkun, y cols., 2019b) para la tarea del RIT, su precisión desciende al 43 %. Este contraste subraya que la eficiencia paramétrica debe ir acompañada de robustez semántica, y que las métricas globales pueden ocultar sesgos significativos en contextos más desafiantes

En este contexto, la presente investigación se sitúa en la intersección de estas problemáticas. Es necesario operar a un nivel de abstracción semántica superior, donde el modelo pueda construir y manipular representaciones conceptuales para mejorar su razonamiento. Nuestro objetivo es, por tanto, proponer y evaluar un marco metodológico que explore esta vía.

[1] https://paperswithcode.com/sota/natural-language-inference-on-rte
[2] Miles de millones de parámetros

1.3. Preguntas de investigación

1. ¿En qué medida un marco metodológico basado en la abstracción de relaciones semánticas puede mejorar el proceso de inferencia de un LLM en el RIT?
2. ¿Qué tipos de relaciones de compatibilidad e incompatibilidad semántica son más efectivas para guiar el razonamiento del modelo y corregir inconsistencias y limitaciones de los LLMs?

1.4. Hipótesis

- Los LLM no tienen flexibilidad en su red conceptual, ni reconocen qué relación semántica es útil e ignoran cómo usarlas, para decidir sobre una clase concreta en la tarea del RIT, por lo tanto, la implementación de un marco metodológico de abstracción de conocimiento semántico mejorará significativamente la capacidad de razonamiento de los LLMs en el RIT.

1.5. Objetivos

- Desarrollar un marco metodológico que modele la Abstracción del Conocimiento Semántico para desarrollar mejores modelos de Inteligencia Artificial para la tarea del RIT.
- Evaluar la aportación del marco metodológico en la tarea del RIT a través de diferentes métricas.

1.6. Estructura de la tesis

En el capítulo 2 se presenta una revisión sobre cómo se aborda el modelado del lenguaje así como sus diferentes problemáticas y cuál es la brecha de conocimiento semántico del estado del arte de modelos en materia de generación de texto y las deficiencias al depender únicamente de la estadística. En el capítulo 3 se aborda el concepto de abstracción, así como las técnicas de *prompting* para guiar el pensamiento estructurado de los LLMs. En el capítulo 4 se aborda la tarea del Reconocimiento de la Implicatura Textual para evaluar nuestra propuesta y el estado de arte. En el capítulo 5 se presenta la propuesta del marco metodológico planteado para la Abstracción del Conocimiento Semántico (SKA). En el capítulo 6 se presentan los diferentes experimentos que se realizaron para evaluarlo y demostrar su validez. Además se realiza un comparativo con métodos similares del estado del arte. En el capítulo 7 se presentan la discusión de los resultados, y por último en el capítulo 8 que aborda las conclusiones obtenidas del modelo propuesto y finalizando con las contribuciones de esta investigación.

Capítulo 2

Conocimiento Semántico

El conocimiento semántico es la base para una comprensión del lenguaje. En el campo del PLN, el conocimiento semántico resulta indispensable para tareas que requieren comprensión profunda, generalización y razonamiento, como la tarea del RIT. Sin el conocimiento semántico, los modelos, se limitarían a identificar patrones formales, sin captar la intencionalidad, las ambigüedades o las relaciones semánticas y lógicas entre conceptos. Por ejemplo, entender que la palabra "banco" puede referirse a una entidad financiera o a un asiento según el contexto, exige un acceso al conocimiento semántico para capturar intenciones y relaciones complejas dentro del lenguaje natural. Es así que es crucial discernir implicaciones, lo cual solo se logra integrando conocimiento sobre el mundo y las relaciones entre significados. Así, la semántica actúa como el puente entre el procesamiento superficial de cadenas de texto y la comprensión genuina, necesaria para modelos que atienden las tareas del PLN.

Existen diferentes niveles de análisis en el lenguaje, a saber: fonológico, morfológico, léxico, sintáctico y semántico. En cada uno de estos niveles existen retos para la creación y desarrollo de modelos que resuelvan de manera individual el problema de variabilidad y ambigüedad lingüística. La variabilidad y ambigüedad del lenguaje dificulta que el ser humano y los modelos de IA puedan obtener el significado de manera confiable sin hacer uso de información adicional. La variabilidad del lenguaje hace referencia a que un mismo significado se puede obtener de distintas frases. Por otro lado, la ambigüedad es el problema donde la información no se puede entender o interpretar de una manera clara. Existen tres principales categorías de ambigüedad:

- Ambigüedad léxica.
- Ambigüedad estructural.
- Ambigüedad semántica.

Ambigüedad léxica

El contexto tiene mucha importancia para identificar y eliminar la ambigüedad de las palabras. Por ejemplo cuando se hace referencia a la palabra "banco", lo que cognitivamente se relaciona con esta palabra se encuentra dentro de dos definiciones[1]:

- Banco: Institución financiera donde se gestiona dinero.
- Banco: Objeto para sentarnos.

Para poder desambiguar el significado de la palabra hacemos uso de su contexto. Es decir, si la palabra se encuentra inmersa en una frase se entiende cuál es su significado.

- "*El banco aumentó sus intereses el año pasado.*"
- "*Me senté en el banco después de una larga caminata.*"

Cuanto más corta es la frase y menor el contexto, mayor es la ambigüedad. Si se cuenta con el conocimiento de los dos significados, es muy fácil seleccionar uno de los sentidos de una palabra ambigua sin ser necesario un procesamiento adicional. Por otro lado, cuando no se tiene el conocimiento del significado de una palabra en una frase, esto puede llevar a una interpretación incorrecta. Aunque el contexto aporta información relevante para el proceso de interpretación también existen otros problemas. Un ejemplo de esto es la siguiente frase:

- "*la cura de su enfermedad*".

¿Cómo entender esto? La primera revisión de la frase podría dar a entender que se refiere a "la cura" como el tratamiento medicinal para la enfermedad. Pero por otro lado, se puede entender a "la cura" como una acción de atender a la enferma, es decir, una persona que está realizando el proceso de curación a otra persona.

Ambigüedad estructural

La estructura es importante para interpretar una frase. Pero una frase puede interpretarse con dos o más significados posibles, debido al orden y agrupamiento o la distinta función gramatical de las palabras.

- "*El pollo está listo para comer*",

El ejemplo anterior puede entenderse de dos formas: La primera, es que un animal está esperando su comida para que se alimente. La segunda, es que el animal está cocinado para que alguien más se lo coma.

[1] Dependiendo del idioma y el conocimiento general se puede tener más definiciones.

Ambigüedad semántica

La ambigüedad semántica ocurre cuando una palabra o concepto tiene un significado de por sí difuso que se basa en el uso informal o generalizado.

- "*Solo pueden opositar personas con ambos títulos*".

Este ejemplo se puede interpretar de dos formas. Primero, se entiende que cualquier persona puede opositar si tiene los dos títulos. Pero también se entiende que con cualquiera de los dos títulos se puede opositar.

La interpretación del lenguaje es un proceso activo que depende de supuestos y conocimiento del mundo compartido, sin los cuales la comunicación sería imposiblemente lenta e ineficiente. Interpretar una frase no es solo entender las palabras, sino activar instantáneamente una red de suposiciones y conocimiento del mundo. Por ejemplo:

- "*El hermano de María tiene tres perros de raza dálmata*".

De la afirmación explícita se activa de inmediato una serie de suposiciones y de conocimiento del mundo que deben ser verdad; información que podemos asumir como verdadera por el hecho de que proviene, o se puede inferir, del texto mismo. Algunos ejemplos de esto se muestran a continuación:

- s_1: "*María tiene un hermano*".
- s_2: "*Su hermano de María es hombre*".
- s_3: "*Su hermano de María tiene tres perros*".
- s_4: "*Los perros que tiene su hermano de María son de raza*".
- s_5: "*Los perros de su hermano de María son dálmatas*".

De la misma forma se generan supuestos relacionados con el texto; estos pueden resultar ser falsos o verdaderos de contar con información adicional, por ejemplo:

- s_6: "*Los perros dálmatas son caros*".
- s_7: "*El hermano de María tiene mucho dinero*".
- s_8: "*Los perros dálmatas son blancos con manchas negras*".

La validez de estos supuestos están relacionados con conocimiento del mundo. Por ejemplo, si se sabe que los perros dálmatas son de color blanco y negro y no son muy comunes como mascotas y además de esto, los perros de raza son caros, se puede asumir la frase s_7 y s_8. Es decir, depende de la información que se tiene y la incorporación de nueva información. Lo que no se sabe pero se puede inferir es la frase s_6, ya que no hay certeza, de cómo es que el hermano de María adquirió los perros. Del mismo modo, se puede discriminar información que contradiga el conocimiento que se tiene. Esta

información adicional es útil para el soporte de las decisiones.

El lenguaje natural presenta una serie de características intrínsecas que lo hacen enormemente expresivo, pero también notoriamente complejo de modelar para una máquina. Para que los sistemas de PLN puedan procesar, generar o manipular texto de manera efectiva, deben estar equipados con técnicas específicas que aborden estas dificultades desde sus niveles más básicos. Las técnicas más utilizadas son: Etiquetador de partes del discurso (por sus siglas en inglés, POS, Part of Speech), análisis sintáctico superficial (Chunking / shallow parsing), eliminación de palabras funcionales (Stop-words), lematización (Stemming), Frases compuestas o estadísticas (Compound or Statistical Phrases) y Desambiguación de la palabra (Word Sense Disambiguation).

La variabilidad y ambigüedad del lenguaje impone retos a la tarea de tratamiento del lenguaje, así como una representación adecuada para procesar el conocimiento semántico. Es decir, generar una representación que capture el significado de las palabras, sub-frases y frases y diseñar medidas que capturen adecuadamente la semántica siguen siendo un campo abierto en la actualidad. Los modelos de IA deben ser capaces de manejar esta complejidad para lograr una comprensión real del lenguaje humano.

Para alcanzar una comprensión profunda del lenguaje, es crucial reconocer que las palabras no existen de manera aislada, sino dentro de una red de relaciones semánticas. Estas relaciones de sinonimia, hiperonimia, hiponimia, co-hipónimos (conceptos distintos que comparten un mismo hiperónimo), antonimia, holonimia (relación todo-parte) y meronimia (parte-todo), entre otras, constituyen el conocimiento semántico. Este conocimiento permite dotar de sentido funcional a las palabras y es indispensable para cualquier análisis lingüístico que abarque los niveles léxico, estructural y semántico.

Sin embargo, la interpretación final depende del contexto, que da significado específico a los conceptos dentro de un texto o enunciado concreto. En el ámbito del PLN, el conocimiento semántico se modela principalmente mediante dos estrategias computacionales complementarias:

1. **Word Embeddings (WE):** Captura el significado de las palabras a partir de sus patrones de co-ocurrencia en grandes volúmenes de texto. Este enfoque es potente para inferir asociaciones y analogías de manera implícita, pero no representa explícitamente relaciones jerárquicas como la hiperonimia.

2. **Recursos Ontológicos:** Se basan en estructuras formales, como taxonomías y ontologías que definen explícitamente las relaciones entre conceptos. Estas bases de conocimiento ofrecen un andamiaje semántico claro y jerárquico, muy valioso para tareas que requieren razonamiento preciso, pero su construcción es costosa y a menudo está limitada en cobertura y escala.

La comprensión del lenguaje más robusta surge, por tanto, de la integración de ambos enfoques: la capacidad para capturar el uso del lenguaje en contexto, y la capacidad estructural y explicativa de las ontologías para representar el conocimiento semántico organizado y las relaciones lógicas entre conceptos. Esta combinación permite a los sistemas no solo "calcular" el significado a partir de datos, sino también "razonar" sobre las relaciones que subyacen a las palabras.

2.1. Semántica distribucional

El PLN combina herramientas lingüísticas y métodos estadísticos para modelar el lenguaje humano y permitir la interacción entre máquinas y personas. Su objetivo principal es que los sistemas computacionales puedan procesar el lenguaje natural y en cierta medida, "comprender" su significado, capturando relaciones a nivel léxico, sintáctico y semántico. Actualmente, existen avances importantes gracias a modelos de inteligencia artificial, como los modelos de lenguaje de gran escala, asistentes virtuales y chatbots, que buscan abordar de manera más profunda el conocimiento semántico.

En este contexto, un enfoque fundamental es la semántica distribucional, que se basa en la idea de que el significado de las palabras se deriva de su uso lingüístico. Esta teoría sostiene que las palabras adquieren sentido a partir de su relación con otras palabras dentro de estructuras de frases y contextos comunicativos. En otras palabras, el significado no reside de manera aislada en cada término, sino que emerge de cómo se combinan y utilizan en la práctica lingüística, permitiendo así representaciones matemáticas que capturan similitudes y asociaciones semánticas a partir de grandes volúmenes de texto.

Las frases conllevan una gramática, estructura y orden para poder crearse correctamente, principalmente por tipos de palabras: artículos, sustantivos, verbos, adjetivos, adverbios, entre otros. Pero el lenguaje y su variabilidad, permite distintas formas para generar frases con el mismo significado. Podríamos analizar esto, con el siguiente ejemplo:

- f_1: Un hombre va a la fiesta.
- f_2: Una mujer va a la fiesta.
- f_3: Una persona va a la fiesta.

La hipótesis distribucional asevera que palabras que se usan y aparecen en los mismos contextos tienden a transmitir y a tener significados similares. En efecto, cuando hablamos de las relaciones $hombre - persona$, $mujer - persona$ preservan el mismo significado, de izquierda a derecha. Pero, ¿$hombre - mujer$ tienen el mismo significado? Es aquí, donde la hipótesis distribucional enfrenta su prueba crítica con pares como $hombre - mujer$.

Los enfoques clásicos en el campo de *PLN* se han beneficiado de las representaciones vectoriales del uso de las palabras. Los primeros modelos de las representaciones vectoriales fueron los *one-hot-encoding* que presentaron serios inconvenientes, por ejemplo, el problema de la dimensionalidad, entre mayor sea el vocabulario, las dimensiones del vector son proporcionales. En este esquema, cada palabra del vocabulario se representaba como un vector disperso (sparse) de longitud igual al tamaño total del vocabulario, con un valor 1 en la posición correspondiente a esa palabra y 0 en todas las demás. Por lo que, con el enfoque de la semántica distribucional se obtiene una representación vectorial a través de redes neuronales. Las redes neuronales mejoran la capacidad de capturar información semántica de las palabras generando *words embeddings* (WE) a través de la co-ocurrencia con otras palabras y la distribución de probabilidad.

Las *WE* son vectores numéricos con un tamaño de dimensiones *d*. Son una representación de valores numéricos calculados sobre la co-ocurrencia de palabras con su contexto, es decir, captura la distribución de las palabras sobre un corpus. Para poder obtener estas representaciones se requiere de grandes cantidades textuales. El contexto de las palabras refiere a las palabras contiguas de la palabra objetivo *w*. Los *WE* capturan algunos aspectos léxico-conceptuales del uso de las palabras y con el uso de medidas de similitud sobre estos han permitido grandes rendimientos en la tareas tareas del PLN.

En este sentido, las redes neuronales típicamente utilizan las *WE* para posteriormente realizar operaciones vectoriales y obtener algún tipo de significado de la frase. De hecho se ha demostrado (Mikolov y cols., 2013) que analogías del tipo *rey - es_a - hombre*, lo que *reina - es_a - mujer*, se capturan en estas representaciones. Por lo que el enfoque de semántica distribucional presenta una estrategia para representar relaciones semánticas parciales; de uso de las palabras. Las *WE* son entonces la traducción del lenguaje humano (en un corpus) a un espacio geométrico donde la semántica se convierte en distancia y dirección. Son el puente fundamental que permite a las máquinas comenzar a interactuar matemáticamente con conceptos lingüísticos, sentando las bases para la comprensión del lenguaje por parte de la IA.

Existen distintos algoritmos que intentan capturar relaciones semánticas sobre el uso de las palabras, asignando un vector numérico que intenta abstraer dicha información. Las limitaciones que existen en este enfoque es que al asignar un vector a cada palabra se acarrea la ambigüedad léxica, es decir, la palabra "banco" tiene varios sentidos pero solo un vector que lo representa. Los vectores reflejan sesgos implícitos en los corpus en los que se entrenan debido a la naturaleza del lenguaje y del uso de las palabras. Actualmente existen diversos algoritmos para la generación de *WE* tales como:

- **BoW:** Bolsa de palabras donde el vector numérico contiene 1s y 0s y la dimensión esta dada por el tamaño del vocabulario. El valor 1 denota la palabra en la posición determinada del vocabulario.
- **TF-IDF:** (frecuencia del término - frecuencia inversa de documento) es un valor matemático

determinado para reflejar la importancia de una palabra (término) en un documento dentro de un conjunto de documentos (corpus).

- **Word2vec:** Basado en redes neuronales, genera vectores (WE) para cada palabra del vocabulario de acuerdo a la aparición en el contexto definido por n palabras contiguas.
- **GloVe:** (Global Vectors) Es un modelo de aprendizaje no supervisado. El entrenamiento se realiza a partir de estadísticas globales agregadas de co-ocurrencia palabra-palabra de un corpus, y las representaciones (WE) resultantes muestran interesantes subestructuras lineales del espacio vectorial de palabras.
- **FasText:** Es una extensión del modelo Word2vec y se basa en la idea de representar palabras como una bolsa de n-gramas de caracteres o unidades de subpalabras, en lugar de palabras individuales.
- **ELMo:** (Embeddings from Language Models) es distinto a las incrustaciones de palabras como las mencionadas antes porque usa una red neuronal profunda que analiza el contexto completo en el que aparece la palabra.

Las redes neuronales que generan *WE* lo hacen sobre una función de costo que minimiza la similitud coseno entre los vectores numéricos. La idea central es que palabras con significados similares tendrán vectores similares (cercanos en el espacio vectorial). Esto es para medir que tan similares son dos palabras. Otra de las medidas usadas para medir similitud de palabras es la distancia euclidiana sobre los puntos que representan sus vectores. La distancia Euclidiana está dada por

$$d(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|.$$

y la similitud coseno está dada por:

$$\cos\theta = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|\|\vec{y}\|}.$$

donde $\vec{x}, \vec{y} \in \mathbb{R}^m$.

En el modelado del lenguaje dichas representaciones se generan con el propósito de predecir la continuación de frases a partir de un contexto dado. Los modelos como BERT (Devlin y cols., 2018), XLNet (Z. Yang y cols., 2019) y GPT-3 (Brown y cols., 2020) son capaces de producir lenguaje natural fluido. Estos modelos de lenguaje profundos tienen varios miles de millones de parámetros debido a la enorme cantidad de capas ocultas de su red neuronal. Para poder obtener representaciones de las palabras, se entrenan con extensos corpus con una gran cantidad de palabras. De esta forma, aprenden a generar vectores de representación densos (con números reales), usualmente de algunos cientos de dimensiones, cuya distribución en el espacio refleja correlaciones de co-ocurrencia entre las palabras. Sin embargo, estos modelos están aislados de otros conocimientos externos que pueden ser útiles para su desempeño. Actualmente, no existe ningún sistema de aprendizaje profundo que pueda realizar inferencias basadas en el conocimiento del mundo real (Marcus y cols., 2017).

Los autores (Westera y Boleda, 2019) argumentan que, aunque la semántica distribucional ha tenido un gran éxito en modelar la similitud semántica y otros fenómenos lingüísticos, esta tiene limitaciones en relación con la inferencia de implicaturas y otros aspectos de la pragmática del lenguaje natural. Para aclarar esto, podemos analizar lo siguiente: las representaciones de *cat* y *dog* son similares y al calcular la similitud coseno de sus *WE* da un resultado cercano a 1. A pesar de esto, los conceptos *dog* y *cat* son distintos en su definición ontológica.

Es así que podemos identificar el problema que para modelar el lenguaje computacionalmente se requiere de "comprender" e identificar el significado de las palabras, a pesar de que se usen en contextos similares no refiere que sean lo mismo. Continuando con el ejemplo, *cat* y *dog*, son los dos un tipo de animal, pero son dos entidades diferentes, por lo que refieren a cosas distintas. Esto por un lado, pero también existe el problema de ambigüedad, ya que la palabra *cat* tiene diferentes sentidos en el idioma inglés. Por otro lado, los *WE* se calculan sobre un corpus (escritos por humanos) donde es posible que acarreen estereotipos de uso de lenguaje y de intenciones comunicativas.

A pesar de esto, la semántica distribucional abona a la teoría de la significación de las expresiones sin tomar en cuenta estos aspectos inferenciales. Los *WE* tienen propiedades importantes para el tratamiento computacional del significado.

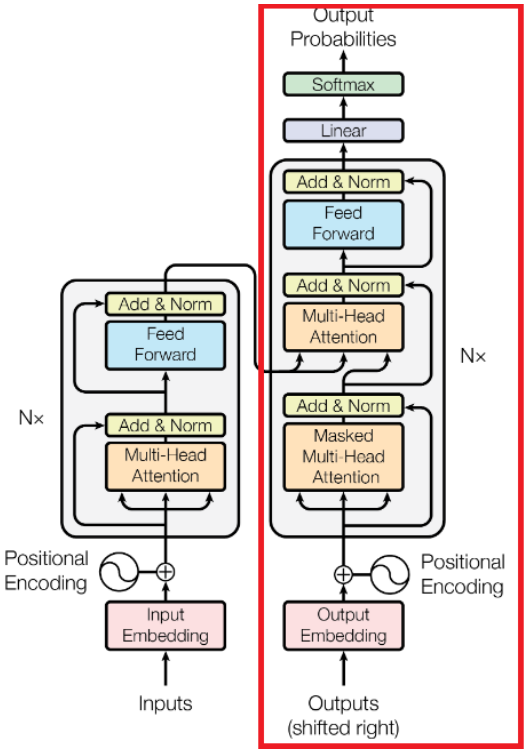


Figura 2.1: Arquitectura *Transformers*.

Nota. Adaptado de *Attention is all you need*, por (Vaswani y cols., 2017).

Actualmente, la arquitectura *Transformer* representa la innovación fundamental detrás de los LLMs. Fue propuesta por (Vaswani y cols., 2017), y sustituyó a las arquitecturas recurrentes y convolucionales, al depender exclusivamente de un mecanismo denominado *Atención*. Una ventaja crucial del *Transformer* es que el mecanismo de atención y las operaciones *feed-forward* pueden calcularse para todas las palabras simultáneamente. Los fundamentos de la arquitectura *Transformer* incluyen:

- **Mecanismo de atención (Attention):** Este es el núcleo del modelo. Calcula para cada token[2] una representación ponderada de su relevancia con todos los demás tokens en la secuencia. Esto permite capturar dependencias de largo alcance, independientemente de la distancia entre palabras.
  - **Codificación posicional (Posicional Encoding):** Dado que la autoatención es paralela, se inyectan señales explícitas sobre la posición de cada token en la secuencia. Esto permite al modelo recuperar la información estructural secuencial. Generalmente se usan funciones seno/coseno para dotar de la noción del orden en la secuencia.
  - **Atención multicabezal (Multi-Head Attention):** En lugar de un único cálculo de atención, se realizan múltiples cálculos en paralelo a través de *heads* ("cabezas"). Se cree que cada cabeza puede especializarse en diferentes tipos de relaciones (sintácticas - semánticas), enriqueciendo la representación final.
- **Arquitectura de capas en cascada (Encoder-Decoder):** El modelo se organiza en pilas de capas idénticas.
  - **Codificador (Encoder):** Transforma la secuencia de entrada en una secuencia enriquecida de representaciones contextuales (WE), capturando las relaciones dentro de la entrada (izquierda de la Figura 2.1)
  - **Decodificador (Decoder):** Genera la secuencia de salida token a token. Utiliza atención sobre las representaciones del codificador y atención enmascarada sobre su propia salida parcial para mantener la coherencia. (derecha de la Figura 2.1)
- **Componentes por capa y normalización:** Cada capa del codificador/decodificador contiene:
  - Una subcapa de autoatención (o atención en el decodificador).
  - Una subcapa de red neuronal feed-forward (aplicada independientemente a cada posición).
  - Conexiones residuales alrededor de cada subcapa, seguidas de normalización de capa (LayerNorm). Esto estabiliza el entrenamiento de redes profundas, mitiga el problema del desvanecimiento/explosión del gradiente y acelera la convergencia.

La evolución de los LLMs; como GPT en sus diferentes versiones (**?**) (Radford y cols., 2019) (Brown y cols., 2020) (OpenAI y cols., 2024), utilizan la arquitectura de Transformers (solo decoder). Estos modelos parten de la estadística como forma de capturar regularidades sobre el uso del lenguaje

[2]Un token es la unidad básica e indivisible que un sistema de PLN utiliza para analizar o procesar un texto

y utilizan las representaciones vectoriales de palabras, que capturan en sus valores, la probabilidad de que ocurra una palabra dado un contexto, la distribución del uso de las palabras y algunas relaciones entre estas. Para generar/aprender estas representaciones se realiza un proceso de entrenamiento (costoso computacionalmente) sobre grandes cantidades de texto. Es así, que grandes empresas como OpenAI, Google, Microsoft y Meta han desarrollado modelos como GPT-4, Gemma, Phi3, Llama, entre otros, proporcionando interfaces para mantener una conversación usuario-máquina y que han sido bien recibidos por la sociedad.

2.2. Bases de conocimiento

Las palabras de por sí muestran una referencia a los conceptos del mundo; un objeto, una acción o un atributo. Por lo que es fundamental obtener más información que tipo de relaciones existen y como interactúan con otros conceptos, y no únicamente que tanto contexto comparten (Speer y cols., 2016). La semántica distribucional nos brinda de información del uso de las palabras, codificando la co-ocurrencia de las palabras, relaciones e información de uso en los *WE*. El conocimiento del mundo no se captura al generar únicamente los *WE*, ya que al momento de identificar qué tipo de relación existe entre dos palabras se vuelve complicado. Por ejemplo, *cat* y *dog* tienen un valor en similitud coseno de 0.8, que indica que comparten contexto. Es decir, es muy común en los textos que se utilicen estos dos conceptos en frases similares. Pero el tipo de relación no se identifica a través de los *WE*. Los humanos pueden saber que estas palabras representan dos entidades diferentes. Es así que se requiere de información adicional para lograr capturar esto en los modelos. El uso de recursos externos en forma de ontología o grafos de conocimiento que intenta cubrir el hueco ontológico, permite dotar de significado a las palabras y las relaciones entre otras palabras.

La cantidad de relaciones semánticas de las palabras y sus sentidos, no se puede encapsular en una base de datos, debido a que existen múltiples usos dependiendo del idioma y el contexto en el que se use. Realizarlo de manera manual es costoso y generarlo de manera automática genera ciertas problemáticas. Por un lado, al realizar el etiquetado manual de las relaciones de las palabras en campos semánticos resulta en un alto costo de recurso humano. Por otro lado aunque las redes neuronales o arquitecturas podrían ayudar a identificar de forma automática estas relaciones, aún hay trabajo por hacer.

Al integrar los dos enfoques (WE y recursos externos) podría ser una manera adecuada para construir modelos que capturen mejor la semántica. En este sentido, dos grandes recursos externos que proveen esta información en forma de grafo son: Wordnet y ConceptNet. Los grafos de conocimiento se construyen principalmente a través de la combinación de conocimiento de sentido común *crowdsourced*, recursos estructurados expertos y PLN para extraer relaciones semánticas entre palabras y frases en múltiples idiomas.

WordNet[3] es una gran base de datos de relaciones conceptuales-semánticas y léxicas. El cual cuenta con conjuntos de sinónimos y sentidos. Además de esto se agrega información por tipo de palabra: sustantivos, verbos, adjetivos y adverbios. ConceptNet[4] es un grafo de conocimiento que conecta palabras y frases del lenguaje natural con aristas etiquetadas. Está diseñado para representar el conocimiento general que sirve para comprender como es que funciona el lenguaje. Contiene un

[3] https://wordnet.princeton.edu/
[4] https://conceptnet.io/

gran numero de relaciones semánticas a partir de muchas fuentes que incluyen recursos creados por expertos.

Existen trabajos que intentan agregar esta información externa a partir de una combinación de recursos externos y semántica distribucional (Speer y Lowry-Duda, 2017). Es así que el aprendizaje automático puede mejorar incluyendo conocimientos y fuentes de información externas para relacionar los significados de palabras y frases. Con el fin de que los modelos "comprendan"mejor los significados que hay detrás de las palabras que utilizamos para comunicarnos. El incluir las relaciones de ConceptNet, puede hacer que *WE* sean más robustas y estén más correlacionadas con los juicios humanos, por lo que unir estas herramientas es un posible camino a explorar.

Continuando con el ejemplo, la relación entre $dog - cat$ no es la misma entre $dog - animal$. El término animal es un concepto más general, que abarca más conceptos. A esto se le conoce como hiperonimia. En el ejemplo, animal es un hiperónimo de dog, y en dirección contraria, se le conoce como hipónimia. Otro tipo de relación que se da entre dog y cat, es la co-hiponimia, ya que los dos pertenecen al mismo concepto general "animal", por su misma naturaleza, son dos entes distintos. No podemos prescindir de este tipo de relaciones entre los diferentes conceptos que existen: hiperonimia, hiponimia, co-hiponimia, holonimia, meronimia, sinonimia y antonimia, debido a que aportan más información para el sentido funcional de las palabras, más aún se deben de considerar los diferentes niveles de análisis del lenguaje, para hablar sobre comprensión del mismo. Sin embargo, en otro nivel de análisis, el contexto da sentido a la interpretación y significado de los conceptos y los textos.

Es así que se debe ir aún más profundo en el uso y análisis de los *words embeddings* para poder entender a través de otras medidas esos fenómenos lingüísticos que se capturan en ellos. Por ejemplo, los hiperónimos e hipónimos son útiles para tareas como la clasificación de textos, la recuperación de información y RIT. Debido a que pueden ayudar a entender las relaciones entre palabras o unidades léxicas, que en efecto los *WE* no dan por si solos. Otro ejemplo son los co-hiponimos que refiere a que las palabras comparten el mismo hiperónimo (Figura 2.2). En PNL, los co-hipónimos pueden ser útiles en tareas para desambiguar el sentido de las palabras y la similitud semántica. Identificar los co-hipónimos puede ayudar a agrupar palabras relacionadas y distinguirlas entre ellas.

Las relaciones semánticas se dan por los vínculos que tienen las palabras y sus categorías gramaticales. Estas constituyen la forma de comprensión del significado. Por ejemplo, el par $red - color$ es una forma de generalizar el concepto red (es decir, rojo es un color) es decir, el concepto color incluye (se encuentra contenido) el concepto red. Esto también sucede con el par $purple - color$. La relación inversa de esos pares hace referencia a algo más específico y acotado (conocido como hiponimia), es decir, del par $color - red$, se entiende que un tipo de color puede ser el concepto red. Lo mismo para el par $color - purple$. En este tipo de mapa conceptual también se puede identificar la relación

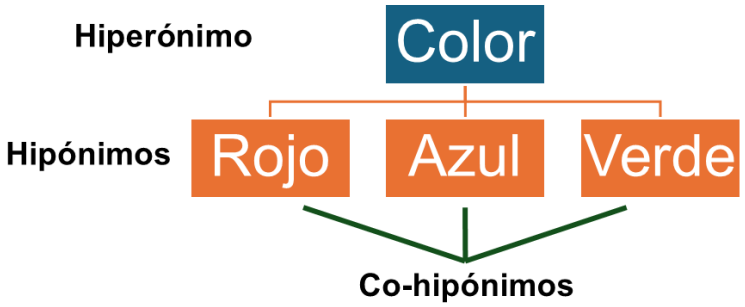


Figura 2.2: Jerarquía conceptual. Relación de co-hiponimia: conceptos relacionados con diferentes significados

$red - blue$. Los dos pertenecen a un mismo hiperónimo, por lo tanto la relación que los une se conoce como co-hiponimos, es decir, son distintos entre sí. A un nivel ontológico, este par, son incompatibles semánticamente. Un ejemplo que puede ser más claro, es el par $dog - cat$, ya que sabemos que son dos entidades diferentes. Esto no solo se da únicamente con sustantivos también con verbos: $walk - run$.

Los recursos más utilizados son WordNet[5], ConceptNet[6], Wikipedia[7], DBpedia[8], entre otros. Debido a la amplia cantidad de recursos externos y que cada uno cuenta con su propia estructura y tipo de relaciones, se ha propuesto (Ilievski y cols., 2021) recategorizar las relaciones existentes de distintos recursos externos para homogenizar en una única estructura. Los autores muestran que dependiendo de la tarea y las categorías, así como de la capacidad del modelo preentrenado, tiene un impacto en su rendimiento.

Por un lado, los recursos externos, aunque extensos, no cubren todas las posibles relaciones semánticas entre conceptos. Aunado a esto, los recursos pueden contener información ruidosa o imprecisa o demasiado general para ser útil en el contexto específico de la tarea. Más aún, la extracción automática de relaciones podría llevar a errores, lo que podría afectar negativamente el rendimiento del modelo. En este sentido, es importante contar con mecanismos y estrategias para filtrar información irrelevante o errónea de los recursos externos para evitar confundir al modelo.

La necesidad de integrar bases de conocimiento estructuradas se vuelve crucial cuando hablamos de la comprensión semántica profunda de los LLMs. La simple correlación de palabras no es suficiente para emular el pensamiento humano, ya que requiere un anclaje conceptual a la experiencia y al mundo real. Estas bases de conocimiento actúan como un "mapa del mundo"que podría proporcionar a los LLMs información semántica y contextualizada sobre conceptos, relaciones entre ideas y estructuras

[5]https://wordnet.princeton.edu/
[6]https://conceptnet.io/
[7]https://es.wikipedia.org/wiki/
[8]https://www.dbpedia.org/

lógicas complejas. En otras palabras, se convierten en un puente entre el lenguaje textual y el mundo real, conectando las frases con significado, contexto y profundidad.

Un modelo que solo procesa información textual de manera aislada; sin una base de conocimiento de sentido común, esta limitado a la "comprensión" del lenguaje. Sin embargo, cuando se integra con una base de conocimiento estructurada, con una metodología de abstracción (ya que los LLMs tienen limitada esta capacidad), es la clave para la coherencia semántica, permitiendo que los modelos "comprendan" la lógica y el contexto del lenguaje de forma más profunda y completa.

2.3. Brecha Semántica de los LLMs

Actualmente, en el área de PLN, se han propuesto modelos que intentan atender las diferentes tareas, por ejemplo, paráfrasis, traducción, generación de texto, razonamiento, resumen, preguntas y respuestas, entre otras. Esto ha iniciado una carrera por obtener el mejor rendimiento optando por la creación de modelos complejos de aprendizaje profundo, con capacidades ocultas y que capturan conocimiento implícito del lenguaje; aunque no se tiene claro cuál es. Los LLMs permiten la generación automática de texto, son una muestra del avance en este campo, aunque aún con problemas que analizaremos en este capítulo.

Más aún y como hemos visto, tratar con el lenguaje es una tarea difícil. Uno de los principales problemas al tratar con el lenguaje natural es su variabilidad, es decir, un mismo significado puede darse con distintas frases. En el proceso de comprensión del significado de una frase, se requiere de diferentes mecanismos para descifrarlo. El lenguaje provee significado incrustado en frases o textos que no son explícitos, por lo que el tratamiento de esto, en los modelos actuales aún no queda claro. Más aún, existe el problema de la ambigüedad: léxica, estructural y semántica. Esto vuelve complicado realmente dar solución a las distintas tareas del PLN.

La semántica psicológica postula que la capacidad humana para el lenguaje es muy productiva a partir de elementos finitos, un principio conocido como composicionalidad. Aun al crear representaciones vectoriales de frases (como "perro de apartamento") a partir de palabras individuales, y logran cierto éxito al predecir juicios de similitud humana, este proceso es fundamentalmente superficial. La composición humana es un acto interpretativo que se fundamenta de manera crítica en el conocimiento del mundo. El desafío central se manifiesta en la incapacidad de los LLMs para capturar las características emergentes: propiedades que son típicas del concepto compuesto, pero no de sus partes constituyentes (Lake y Murphy, 2023).

A pesar de esto, se afirma que la capacidad de los LLMs, al ser entrenados con grandes cantidades de datos son una herramienta útil para apoyar en tareas posteriores (**?**). La ventaja principal es

que al entrenarlos con información real del mundo y de diferentes dominios, permite construir frases correctamente en un nivel gramatical, léxico, sintáctico y en la mayoría de los casos semántico. Es decir, abstrae características implícitas en el lenguaje, en cada uno de esos niveles de análisis, que apoya la predicción de la secuencia de las siguientes palabras dado un contexto. Esto plantea la pregunta ¿Es suficiente para que el modelo pueda comprender estructura, semántica y significado de un texto?

Un primer acercamiento: La paráfrasis

En este orden de ideas, como parte de la propuesta de esta investigación, se analiza la tarea de generación automática de paráfrasis. La paráfrasis se define de manera simple como una "*frase que expresa el mismo contenido que otra pero con diferente estructura sintáctica y léxica*". El enfoque es que la paráfrasis de oraciones se traduce como una transformación de un texto que deberá transmitir el mismo significado pero con palabras, estructura y redacción diferentes. La paráfrasis es un problema en PLN que tiene una amplia gama de aplicaciones, a saber: Aumento de datos, Conservación de información, Mapeo de intenciones y Comprensión semántica.

Los modelos tradicionales de generación de paráfrasis se basan en reglas de sustituciones léxicas a partir de recursos como WordNet. Asimismo, otro enfoque es usar plantillas de expresiones del lenguaje natural. Se ha propuesto el uso de reglas obtenidas (manual o automática) para generar grandes y complejos patrones de paráfrasis en la frase. Por otro lado, se propuso también usar sinónimos de un tesauro para substituir algunas palabras en la frase principal. No obstante, el usar únicamente sustitución de sinónimos limita la diversidad de la paráfrasis (Zhou y Bhat, 2021).

En la actualidad, la generación de paráfrasis se ha beneficiado ampliamente de los recientes avances en el diseño de distintas formas de entrenamiento y arquitecturas de modelos lingüísticos. Sin embargo, las exploraciones se han centrado en gran medida en métodos supervisados, que requieren una gran cantidad de datos etiquetados que son costosos de recopilar. Los investigadores en este campo de la generación de paráfrasis, suelen seguir dos caminos comúnmente:

- Perspectiva estadística: es decir, codificar la frase de origen en una representación vectorial que captura el significado de la oración y la estructura.
- Perspectiva tradicional: separa la planificación del contenido y la generación de la estructura de la salida.

Por ejemplo, en los últimos años, la generación de paráfrasis se formula como un problema de aprendizaje *seq*2*seq* (secuencia a secuencia) o codificador-decodificador. En así que, el enfoque de generación de paráfrasis puede considerarse un caso especial de traducción automática (MT por sus siglas en inglés — Machine Translation) al buscar formas alternativas de expresiones del contenido semántico de una frase. Con este modelo, el codificador representa una frase en un vector de longitud

fija a partir del cual un decodificador genera una traducción.

Por lo regular en la generación de paráfrasis se considera principalmente una única salida para cada entrada. Pero, la naturaleza de la paráfrasis indica que se puede transformar una frase de múltiples formas preservando el significado. (Qian y cols., 2019) proponen un enfoque para atender esto, utilizan dos discriminadores y múltiples generadores para obtener varias paráfrasis de un texto siendo diferentes entre ellas, pero preservando su significado. Para esto, aplican un algoritmo de aprendizaje por refuerzo para entrenar su modelo. El primer discriminador realiza la tarea de examinar si dos frases transmiten el mismo significado; y la tarea del segundo es distinguir qué frase ha sido generada por cuál generador, para identificar de donde proviene una mejor paráfrasis en términos de diversidad.

(Fu y cols., 2020) proponen un modelo que cuenta con dos etapas: 1) planificación de contenidos y 2) generación de la estructura de salida. Para la primera, utilizan el método Bag of Words (BoW) para la generación de paráfrasis. Es decir, utilizan las palabras de origen (frase original) para obtener palabras contiguas, con esto, obtienen la BoW. Esta es usada por la segunda etapa, donde se válida qué puede ser útil para la generación de una frase destino (frase parafraseada) limitando que se repita la misma estructura de la frase original y sus palabras.

(Niu y cols., 2020) proponen un enfoque similar con transferencia de aprendizaje (*Fine-Tunning*); permite que un modelo preentrenado se adapte a una nueva tarea. Este enfoque ajusta los pesos de las capas del modelo para mejorar su rendimiento en la nueva tarea y evita entrenar un modelo desde cero. Los autores lo utilizan para generar paráfrasis de alta calidad en un entorno no supervisado. Además, añaden un algoritmo denominado *Bloqueo Dinámico* (DB). El objetivo del algoritmo es imponer una forma superficial distinta a la de entrada. Es decir, cada vez que el modelo emite un *token* contenido en la secuencia de origen, el DB impide que se emita el siguiente que se encuentra en el origen, para la generación de la frase de salida. Esto permite que el modelo tenga que buscar un nuevo *token* que conserve las palabras importantes pero no la secuencia de la frase de origen.

Otra propuesta para esta tarea es el modelo *Round- trip MT*, que refiere a una trasformación de una fase de entrada en algún idioma, para generar la traducción de la frase en otro idioma, para posteriormente regresar una frase en el idioma origen. Para poder validar el parafraseo correcto utilizan una función de similitud, que requiere un *match* de la distribución de la salida y la entrada. La idea central es que al pasar de un idioma a otro, se modifica la estructura y las palabras usadas por sinónimos. (Ormazabal y cols., 2022) cuestionan que el uso del modelo *Round- trip MT* para la generación de parafraseo, es susceptible a cometer errores y no obtener parafraseo de calidad. Esto es debido a que se generan traducciones ambiguas, es decir, específicamente, el modelo no aborda el problema de marcado de género[9]. Para evitar esto, los autores proponen una nueva alternativa a la

[9]En algunos idiomas solo se cuenta con un artículo para describir varios géneros, por ejemplo el inglés.

función de similitud. Su modelo combina un *encoder* que remueve la información que no es relevante para generar la traducción, y un *decoder* que reconstruye un parafraseo a partir de la codificación.

Para ilustrar las deficiencias de los modelos anteriores, se profundiza en el modelo de (P. He y cols., 2020). Aunque se ha mostrado generar buenos resultados en la paráfrasis de frases con las siguientes características: calidad, diversidad y similitud, aún existen diferentes problemáticas para la validación de estos conceptos. El enfoque propuesto es aprovechar la gran capacidad de los LLMs en la tarea de parafraseo. Es decir, se aborda este problema a través de la transformación del texto inicial y generando una neuva frase con la ayuda de un LLM, manteniendo el mismo significado de la frase original. A continuación se describe la propuesta de los autores (P. He y cols., 2020), y los resultados al replicarlo.

El *LLM* que utilizaron es GPT-2 (*Generative Pre-trained Transformer 2*) (Radford y cols., 2019), el cual se basa en la arquitectura de los *Transformers* (solo decoder). *GPT-2* es un modelo generativo de texto y es conocido por su capacidad para producir respuestas en tareas de generación de lenguaje.

Estos modelos se pueden ajustar para generar paráfrasis a través de la técnica de *Fine Tuning*, la idea es que el modelo ya ha aprendido características del lenguaje que son útiles para la nueva tarea. Existen diferentes versiones del GPT-2 con diferencias de tamaño y del número de parámetros (Figura 2.3). La versión utilizada por (P. He y cols., 2020) es la *medium* que cuenta con 345 millones de parámetros y se encuentra disponible en la librería *huggingface*.

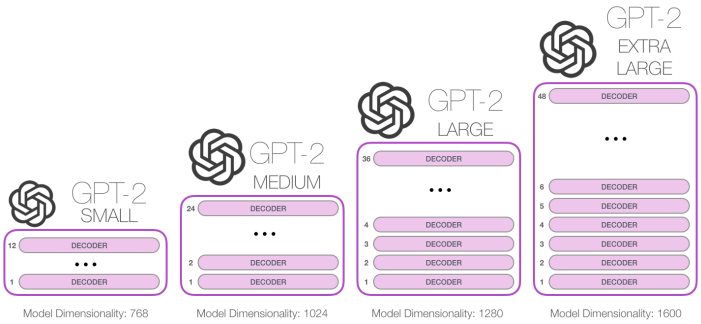


Figura 2.3: Diferentes tamaños de modelos de *GPT-2*.

Nota. Adaptado de *The Illustrated Transformer [Blog post]*, por (Alammar, 2018), https://jalammar.github.io/illustrated-gpt2.

El objetivo es la reconstrucción de la frase original f_o, para lo cuál se realiza un procesamiento de la siguiente forma: 1) eliminación de *stopwords* (palabras funcionales: artículos, preposiciones, conjunciones, pronombres y determinantes), 2) sustitución de sinónimos (Wordnet) y 3) barajeo de palabras restantes para destruir el orden de la frase original y obtener una frase corrupta f_c. Con este

procedimiento se busca reconstruir la frase original f_o a partir de la frase corrupta f_c manteniendo el significado de la frase original.

Los autores (P. He y cols., 2020) generan máximo 10 paráfrasis a partir de una frase, las cuáles son candidatas a paráfrasis. En cada una de estas frases candidatas se debe de considerar que se preserve el significado, que contenga diversidad y que cuente con una estructura diferente a la original. La evaluación de esto se realiza a través de las métricas; calidad: *ROUGE-L* y *METEOR*, diversidad: *BLEU* y *SELF-BLEU* y similitud: similitud coseno y distancia de Levenshtein:

A continuación se muestra como funciona y que captura cada una de las métricas:

- ROUGE-L: la secuencia común más larga entre la hipótesis candidata y la original.
- METEOR: Metric for Evaluation of Translation with Explicit ORdering mide la cantidad de *un-gram* que coinciden (toma en cuenta los sinónimos) entre la hipótesis candidata y la original.
- BLEU: Bilingual Evaluation Understudy que mide el sobrelapamiento de los tokens entre la hipótesis candidata y la original.
- SELF-BLEU: mide el promedio del sobrelapamiento de los tokens entre la hipótesis candidata y la original.
- Similitud coseno > 0.75: mide si la hipótesis candidata y la original son similares.
- Distancian de Levehisten > 6 caracteres: mide que la hipótesis candidata y la original sean diferentes en caracteres.

Este modelo, aprovecha los *WE* de las palabras para ir construyendo las nuevas frases. Es aquí el interés de analizar su funcionamiento que da una idea clara de que estas representaciones abstraen propiedades léxicas, sintácticas y semánticas (parcialmente) que ayudan en la construcción correcta de frases. Una vez replicado el modelo, se generan nuevas frases a partir de una frase original. A continuación se muestra un ejemplo de la generación de paráfrasis para la frase F:

F: **A choir singing at a baseball game**.

- The baseball team is singing at a game. (0.91 similitud)
- Their singing at the baseball game. (0.91 similitud)
- The home team is singing at a baseball game. (0.88 similitud)
- Theirs is singing at a baseball game. (0.88 similitud)
- The home team is singing at the baseball game. (0.88 similitud)
- A pitcher is singing at a baseball game. (0.86 similitud)
- The home is singing at the baseball game. (0.83 similitud)
- Anvir is singing at a baseball game. (0.82 similitud)
- The umpire is singing at a baseball game. (0.76 similitud)

Uno de los problemas que se identifica es que las medidas de similitud que utilizan se quedan cortas al tratar con relaciones semánticas y significado. Debido a que trabajan sobre el plano matemático, sin una relevancia en la parte ontológica de las palabras.

A pesar de que los "parafraseos" de la frase inicial F están escritas correctamente, pierden el sentido y la intención comunicativa de la frase original. Esto es un gran problema, debido a que al depender únicamente de los *WE* (semántica distribucional) se recuperan palabras que comparten y coinciden en contexto: **team**, **home team**, **pitcher**, **Anvir** y **umpire**; pero esto no es suficiente para conservar el significado y lo que se quiere transmitir. En este sentido, la validación del significado compartido de las paráfrasis solo se intenta capturar sobre la similitud coseno.

El objetivo de los modelos de generación automática de paráfrasis producen/extraen frases en lenguaje natural que deben transmitir la misma información semántica. Los LLMs, al centrarse en la estructura gramatical de las oraciones a través de patrones estadísticos, pueden generar reescrituras que carecen del verdadero significado del texto original. La similitud coseno no considera la profundidad y el detalle semántico.

La paráfrasis puede considerarse como una implicatura textual bidireccional y las estrategias para enfrentar las tareas de los dos ámbitos suelen ser similares. De hecho diversos autores mencionan que detectar algún tipo de parafraseo es útil y reduce la complejidad para tomar la decisión de RIT. A pesar de esto, la tarea de generación de parafraseo es tan difícil como la implicatura textual. Aunque la semántica distributiva puede ser útil, la información que nos provee en conservación y composición del significado es mínima. Es decir, las palabras pueden estar relacionadas entre sí de muchas maneras, por lo que una única puntuación será una mezcla de diferentes relaciones o fenómenos lingüísticos que ocurren en el uso de las palabras. El problema es que no es claro cuáles. Los *WE* entrenados con diferentes datos pueden producir resultados diferentes que pueden no ser útiles para el propósito de esta investigación a un nivel semántico, pero si en el uso del lenguaje.

Capítulo 3

Abstracción

Los PLMs y los LLMs comparten una naturaleza común: ambos se fundamentan en el paradigma del preentrenamiento sobre grandes cantidades de textos. Los primeros sentaron las bases técnicas para tareas específicas, mientras que los segundos escalaron masivamente estos principios, desarrollando capacidades emergentes como el razonamiento y la generación versátil de texto. Sin embargo, ambos tipos de modelos siguen enfrentándose a retos fundamentales en sus capacidades lingüísticas, mostrando limitaciones en la comprensión profunda, el razonamiento abstracto consistente y el manejo de contextos complejos o ambiguos.

En su aprendizaje requieren grandes cantidades de textos, que permiten capturar la estructura, características y patrones importantes del lenguaje natural; sin embargo, se ha demostrado (Brown y cols., 2020) que no solo es necesaria la estadística, ni tampoco es suficiente aumentar la cantidad de los datos, ni aumentar el tamaño de dichos modelos, puesto que siguen generando resultados falsos o simplemente inútiles para el usuario. La pregunta que se desencadena es qué, si a pesar de esto ¿Se pueden aprovechar estos modelos de IA para poder abordar el problema descrito en esta investigación?

Aunque los LLM demuestran capacidades sofisticadas de generalización, su forma de abstraer dista de ser análoga a la humana. Estas habilidades se queda corta ante una verdadera comprensión conceptual, lo que resulta en fallos notorios como las alucinaciones o la falta de razonamiento de sentido común (causalidad). Para cerrar estas brechas, se requiere la Abstracción Semántica: la capacidad de trascender la mera estadística textual y construir un modelo interno y estructurado del conocimiento. Esta abstracción permitiría a los LLMs manipular conceptos de forma flexible, verificar la lógica detrás de los hechos y realizar inferencias complejas de alto nivel, elevando su función de un sistema de predicción de texto a un sistema de cognición y razonamiento robusto.

Desde la psicología, la perspectiva predominante concibe el significado de las palabras como un mapeo entre unidades léxicas y estructuras conceptuales. Se postula que los seres humanos poseen un sistema de conceptos que funciona como base de su conocimiento del mundo, y que el significado de una palabra opera esencialmente como un referente o indicador que activa una porción específica de dicho conocimiento conceptual (Lake y Murphy, 2023). Esta aproximación presenta dos ventajas fundamentales como marco explicativo de la cognición humana:

- Las personas organizan su experiencia en conceptos y conocimientos sobre el mundo.
- Proporciona un vínculo directo entre el lenguaje y la realidad: las palabras se conectan al mundo exterior a través de su asociación con otros conceptos.

3.1. Abstracción en LLMs

Más allá de sus capacidades superficiales, los LLMs exhiben limitaciones fundamentales que los separa de una semántica humana genuina. Dos deficiencias críticas son su composición conceptual superficial y su dificultad con la abstracción (Lake y Murphy, 2023). Respecto a la primera, los modelos fallan en capturar las características emergentes que surgen de combinar conceptos, ya que carecen del conocimiento de fondo necesario para esa interpretación. Respecto a la segunda, incluso los sistemas multimodales entrenados para seguir instrucciones no logran internalizar conceptos abstractos y componibles, como la negación, las relaciones de sentido común o la generalización flexible de acciones, lo que revela un aprendizaje ligado a ejemplos específicos y no nuevas representaciones.

Investigaciones recientes han comenzado a desentrañar cómo los modelos desarrollan formas básicas de procesamiento abstracto, aunque con limitaciones claras. El benchmark COPEN (Peng y cols., 2022) muestra que fallan sistemáticamente en la abstracción conceptual: son incapaces de organizar entidades por similitudes semánticas profundas, de inferir propiedades abstractas de forma consistente, de contextualizar conceptos entre distintos dominios del conocimiento y, tienden a confundir relaciones semánticas con co-ocurrencias estadísticas superficiales.

Además, (Elazar y cols., 2021) muestra que la consistencia (la invariancia que preserva el significado con los cambios en su entrada de los PLMs) es deficiente, aunque con una alta varianza entre las relaciones. Este problema se extiende a las comparaciones entre idiomas, en las que proporcionan respuestas desiguales o incorrectas dependiendo del idioma en el que se consulta el mismo hecho (Qi y cols., 2023).

(Beloucif y Biemann, 2021) analizan en qué medida los PLMs codifican relaciones semánticas. Los autores revelan que los PLMs no son capaces de capturar la similitud semántica entre diferentes palabras que se refieren a los mismos conceptos y siguen siendo mucho peores que los humanos en

esta tarea. En este mismo sentido, (Cao y cols., 2025) analizan si los modelos de lenguaje adquieren conocimiento en su proceso de entrenamiento sobre las siguientes relaciones semánticas: hiperonimia, hiponimia, holonimia, meronimia, sinonimia y antonimia, comparándolos con el conocimiento humano. Los resultados revelan que los conocimientos adquiridos por los LLMs examinados son limitados e insatisfactorios. En particular, los resultados indican que los modelos están sesgados hacia la antonimia, ya que les cuesta discernir con relaciones distintas de la antonimia.

Los LLMs encuentran patrones y relaciones de los conceptos de las frases procesadas en el entrenamiento y aprovechan estos conocimientos para la tarea de generación de texto. En este mismo sentido, se pueden evaluar a los LLMs con respecto a que cantidad de información relevante contienen para atender a una tarea principal. Los autores (Sahu y cols., 2022) introducen la medida de consistencia conceptual de los *LLM* de acuerdo a una tarea principal de preguntas y respuestas de elección múltiple. Para esto se extrae conocimiento de una base de conocimiento (ConceptNet), y se transforma a una serie de preguntas, para cuestionar al modelo. Por ejemplo: ¿la flor es roja? con opciones de sí o no. Esta pregunta de fondo se considera relevante debido a que esta asociada a una tarea principal: ¿De qué color es la rosa?. Si el modelo responde que no a la primera, pero a la principal responde "roja" se habla de una carencia de consistencia conceptual. Los autores concluyen de sus experimentos que la consistencia conceptual aumenta con el tamaño del modelo, pero también existen modelos más grandes que muestran un bajo nivel de consistencia conceptual. Es decir, aunque los LLMs responden correctamente a la tarea principal no necesariamente poseen información de fondo. ¿Cómo es posible esto? La intuición de la respuesta se podría deber al corpus de entrenamiento con los que fue entrenados o la falta de parámetros sobre la cantidad de datos procesados.

Más aún, los autores (Sahu y cols., 2022) identifican que los LLMs tienen sesgos al responder, ya que los modelos al preguntarle información de fondo (¿la flor es roja?), la respuesta que dan en su mayoría es sí, aunque en realidad es no. Los autores sugieren dos situaciones a esta problemática, que los corpus de entrenamiento tienden a tener afirmaciones y que las preguntas de fondo son realmente sencillas comparadas con las preguntas de la tarea. Sin embargo, sugieren que los LLM todavía fallan en razonamiento y el conocimiento del mundo.

Del mismo modo, estudios como el de (Dutt y cols., 2024) muestran que, incluso cuando se evalúan sistemáticamente múltiples dimensiones de la generalización (dominancia, robustez y composición), los modelos tienen más dificultades con la generalización robusta, es decir, la capacidad de superar atajos y perturbaciones superficiales, lo que sigue siendo un reto crítico en PLN independientemente del tamaño o la arquitectura del modelo (McCoy y cols., 2019; Jin y cols., 2020). Si bien la capacidad de generalización es inherente a cada modelo, está significativamente influenciada por su tamaño, arquitectura y estrategia de entrenamiento, lo que pone de relieve la necesidad de innovaciones que promuevan un razonamiento verdaderamente abstracto más allá de los enfoques actuales.

(Regneri y cols., 2024) exploran un enfoque innovador para detectar la abstracción nominal en modelos de lenguaje como BERT, centrándose en cómo estos sistemas procesan y representan relaciones jerárquicas entre conceptos, enfocandose en la hiperonimia como fenómeno clave para analizar la abstracción lingüística del modelo. Aunque su investigación se restringe únicamente a sustantivos y no explica cómo emergen estas representaciones en el mecanismo de atención, queda abierta la pregunta sobre si los LLMs desarrollan alguna forma de abstracción o simplemente replican patrones estadísticos. Futuras investigaciones podrían explorar otros tipos de relaciones y profundizar en el vínculo entre atención y conceptualización.

Por su parte, (Al-Saeedi y Harma, 2025) profundizan en la capacidad de los LLMs (arquitecturas pequeñas, específicamente modelos de dos o tres capas) para reconocer, aprender y generalizar patrones simbólicos secuenciales abstractos o plantillas. La tarea es completar una secuencia de símbolos: "ABCABCAB" que después son sustituidos por valores: "12312312". A través del estudio demuestran que estos sistemas no solo pueden internalizar estructuras abstractas, sino también aplicarlas para generar respuestas correctas. Este proceso revela un mecanismo emergente llamado "cabeza de abstracción", que opera identificando relaciones estructurales entre marcadores de posición simbólicos ("ABCABCAB"), independientemente de su contenido concreto ("12312312").Concluyen que la capacidad para manejar patrones abstractos no es inherente a cualquier configuración, sino que emerge a partir de una mínima complejidad arquitectónica. Finalmente, su productividad, como la definen los autores: habilidad de generalizar y aplicar reglas abstractas en nuevos contextos, resulta limitada, especialmente frente a tareas que requieren adaptación a estructuras no vistas previamente.

Por otro lado, (Lee y cols., 2025) examinan las capacidades de razonamiento de los LLMs a través del corpus ARC (Abstraction and Reasoning Corpus), un benchmark diseñado para evaluar habilidades de inferencia abstracta. A diferencia de enfoques tradicionales centrados únicamente en la precisión de las respuestas, la investigación adopta una perspectiva orientada al proceso, utilizando la Hipótesis del Lenguaje del Pensamiento (LoTH) como marco teórico. Este enfoque permite analizar tres dimensiones clave: la coherencia lógica, la composicionalidad y la productividad en el razonamiento de los LLMs. Los resultados revelan una brecha significativa entre el desempeño de estos modelos y el razonamiento humano. Aunque los LLMs demuestran cierta capacidad inferencial, presentan fallas fundamentales en coherencia semántica y lógica, a menudo llegando a conclusiones correctas mediante procesos de razonamiento inconsistentes o arbitrarios. Además, muestran dificultades en el razonamiento composicional, es decir, en la capacidad de integrar operaciones paso a paso para resolver problemas complejos. Estos hallazgos sugieren que, para alcanzar un nivel de abstracción y coherencia comparable al humano, los modelos deben superar limitaciones fundamentales en su arquitectura, posiblemente integrando mecanismos más explícitos de representación simbólica y validación lógica.

Estas investigaciones abren perspectivas innovadoras para el procesamiento neuro-simbólico,

destacando cómo los LLMs pueden realizar inferencias que trascienden lo puramente estadístico. La identificación de la "cabeza de abstracción" no solo explica parte de la eficacia del aprendizaje en contexto, sino que también plantea interrogantes sobre cómo integrar mecanismos similares en paradigmas híbridos, donde la flexibilidad simbólica y la eficiencia neuronal converjan. El estudio, así, sienta bases para explorar modelos capaces de razonamiento abstracto más allá de los límites actuales.

La preocupación es latente en el marco de la explicabilidad de lo que ocurre cuando utilizamos LLM en tareas de PLN. (Hendel y cols., 2023; Z. Liu y cols., 2024; Al-Saeedi y Harma, 2025) han estudiado los procesos internos de los LLM utilizando ejemplos concretos de tareas, analizando la transparencia del aprendizaje contextual (CLI) para dilucidar los mecanismos que permiten a estos modelos aprender y realizar tareas basadas en nueva información o pocos ejemplos, aunque reconocen que una comprensión completa y teóricamente fundamentada de los complejos mecanismos internos del CLI sigue siendo un área de investigación activa con importantes retos, lo que podría conducir a mejoras en la arquitectura de los LLM y en las estrategias de preentrenamiento.

Por último, los enfoques actuales para evaluar estos modelos carecen de una metodología que permita medir las habilidades cognitivas necesarias para realizar las tareas. Para abordar esta cuestión, (Huber y Niklaus, 2025) propone la taxonomía de Bloom como marco jerárquico, y demuestran que los LLMs obtienen mejores resultados en los niveles cognitivos básicos y que los parámetros de referencia actuales presentan importantes lagunas. Se argumenta que una evaluación más precisa requiere un enfoque estructurado que mida sistemáticamente las habilidades cognitivas subyacentes.

La cognición humana adquiere relaciones abstractas con pocos ejemplos, capacidad que el aprendizaje profundo puro no posee. (Marcus y cols., 2017) argumenta que la comprensión requiere superar este límite mediante modelos híbridos que integren el aprendizaje profundo con sistemas simbólicos. Esta hibridación emularía la flexibilidad del cerebro, que surge de tratar con múltiples tipos de cómputo. Las críticas del autor se centran en las deficiencias en generalización y razonamiento en los LLMs:

- **Estructura jerárquica**: Los LLMs, al usar representaciones secuenciales planas, no capturan explícitamente la jerarquía del lenguaje, limitando su generalización estructural.
- **Inferencia Abierta y Razonamiento Multi-etapa**: Su rendimiento decae en tareas que requieren combinar múltiples premisas o integrar conocimiento pragmático e implícito.
- **Causalidad y sentido común**: Operan sobre correlaciones estadísticas, careciendo de modelos causales o una base de sentido común para su razonamiento.

Todo esto sugiere que el reto no solo reside en la escala, sino también en la arquitectura cognitiva. Aunque el enfoque actual de predicción de palabras permite que surja cierta abstracción, parece

insuficiente para replicar los mecanismos composicionales y abstractivos del pensamiento humano. La verdadera prueba será si pueden trascender la imitación estadística y lograr una comprensión estructural genuina. Nuestra propuesta de investigación no es abandonar el aprendizaje profundo, sino integrarlo en arquitecturas más amplias que combinen lo neuronal y lo simbólico, permitiendo operar en distintos niveles de abstracción como lo hace la inteligencia humana.

3.2. Más allá de la superficie

Frente a estas limitaciones, una línea de trabajo recurrente consiste en enriquecer los modelos con recursos externos estructurados (p. ej., bases de conocimiento ontológicas) para proporcionarles información semántica y conceptual (Agrawal y cols., 2024). No obstante, esta estrategia presenta sus propios retos: los recursos externos suelen ser incompletos, pueden contener información ruidosa o imprecisa (que una relación entre conceptos pueda pertenecer a distintas categorías, que semánticamente son incompatibles), y su extracción e integración automática puede introducir errores que perjudiquen al modelo (Dai y cols., 2025).

Ante estos retos, se ha desarrollado técnicas de *prompting* cuyo objetivo fundamental es, precisamente, utilizar el conocimiento que los LLMs, adquieren durante su entrenamiento. Desde la creación manual de instrucciones, un proceso que requería gran intuición y experimentación, hasta estrategias mucho más complejas y automatizadas. El fin es diseñar un proceso de interacción que permita acceder al conocimiento interno del modelo útil para alguna tarea específica.

Investigaciones recientes han desarrollado sofisticadas estrategias de *prompting* para intentar mejorar la fiabilidad y la capacidad de razonamiento de los LLMs. Por ejemplo: *0-shots* o *few-shot*. La primera alude a que podemos realizar preguntas a los LLMs, solicitando claramente una respuesta de una tarea objetivo. Por ejemplo *Question Answering*, donde solo se requiere la predicción de una respuesta. *few-shots*, a diferencia, hace referencia a que le podemos proporcionar, además de lo anterior, ejemplos de la tarea con su solución. Otro ejemplo, más comúnmente utilizado es la cadena de pensamiento (CoT) (Wei y cols., 2022), que mejora el rendimiento de los LLMs al desglosar la tarea en procesos de razonamiento explícitos solicitando que se realicen paso a paso. Por ejemplo, ante una pregunta: ¿Cuántas pelotas tiene Roger si empezó con 5 y compró 2 latas de 3 pelotas cada una?, el modelo debe responder: Roger empezó con 5 pelotas. 2 latas de 3 pelotas cada una son 6 pelotas. 5 + 6 = 11. La respuesta es 11". Sin embargo, CoT a menudo conduce a la repetición y la optimización local. Para solucionar esto, (Lin y cols., 2024) proponen un nuevo método que primero toma las respuestas del LLM, generadas sobre una diversidad de rutas de razonamiento y, a continuación, se le pide que determine la respuesta más coherente entre las rutas generadas.

Otras mejoras, como la autoconsistencia con CoT (CoT-SC) (X. Wang y cols., 2023; W. Chen y

cols., 2024), apoyan la robustez mediante el muestreo de múltiples rutas de razonamiento y la selección de la respuesta más coherente, es decir, la respuesta más común. De manera similar, Tree of Thoughts (ToT) (Yao y cols., 2023) amplía este paradigma explorando y evaluando dinámicamente diversas trayectorias de razonamiento, lo que permite el retroceso, la detención y la autocorrección de rutas de razonamiento, aunque con una mayor sobrecarga computacional. En la siguiente Figura (Figura 3.1) se ilustran las diferentes técnicas mencionadas:

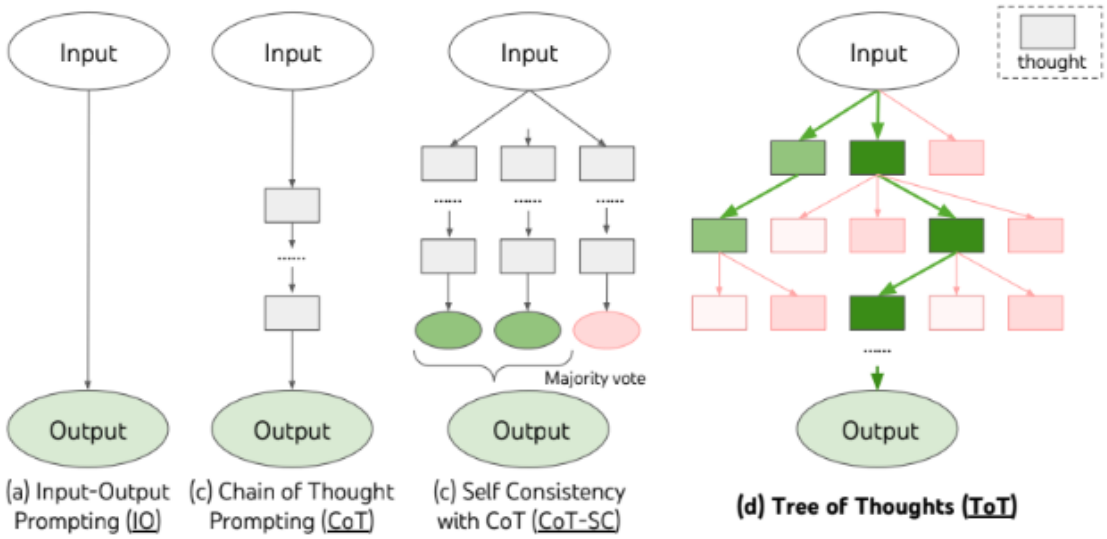


Figura 3.1: Enfoques para resolver problemas con LLMs

Nota. Adaptado de *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*, por (Yao y cols., 2023).

Otras técnicas se han desarrollado, por ejemplo, *prompting* iterativo (DIVERSE) por (Y. Li y cols., 2023). DIVERSE genera múltiples prompts diferentes para una misma pregunta (sobre el corpus GSM8K[1]). Además integran verificadores paso a paso (usando el modelo *deberta-v3-large*) y mecanismos de votación para mejorar la fidelidad del razonamiento, lo que reduce sustancialmente los errores en tareas complejas. Sin embargo, con métodos avanzados como la abstracción del pensamiento (AoT) (Hong y cols., 2024), que obliga al razonamiento jerárquico (de lo abstracto a lo concreto). Por ejemplo, para resolver una ecuación cuadrática, el modelo primero debe identificar la fórmula cuadrática general $x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$, como un paso abstracto superior, y luego sustituye los coeficientes específicos para obtener el resultado. Ante esto, los LLMs siguen mostrando limitaciones en tareas que requieren una generalización profunda, ya que tienen una rigidez interpretativa (Sedova y cols., 2024) y un sesgo cognitivo (Echterhoff y cols., 2024).

Más aún, estas mejoras en el razonamiento de los LLMs son especialmente relevantes cuando se

[1]GSM8K fue diseñado específicamente para evaluar las capacidades de razonamiento aritmético en LLMs.

trabaja con conocimiento estructurado, ya que exige no solo recuperar hechos aislados, sino también comprender las relaciones, dependencias y jerarquías que existen entre ellos. El estudio (Dai y cols., 2025) muestra que los LLM procesan los grafos de conocimiento (KG) de forma más eficaz cuando se representan como tripletas (c_1, r, c_2) en comparación con lenguaje natural. Es importante destacar que, aunque los modelos más grandes son más precisos, son más sensibles a los datos incorrectos que a la información que falta. Esto pone de relieve la importancia de minimizar el ruido en los KG, especialmente cuando se trabaja con LLMs avanzados, y proporciona una valiosa orientación para optimizar las consultas de conocimiento estructurado.

Aunque pareciera que el tratamiento de las tareas de PLN se pueden ver mejoradas en rendimiento por incorporar conocimiento del mundo, la pregunta que surge es si eso es suficiente o hay que tomar otros parámetros para realmente aprovechar las habilidades de los LLMs. Por ejemplo (L. Chen y cols., 2024) demuestran que las llamadas adicionales al modelo de lenguaje mejoran el rendimiento en consultas "fáciles"pero lo degradan en consultas "difíciles". Es decir, si hay una cantidad suficiente de solicitudes a un LLM y este responde al final correctamente, entonces es una pregunta fácil, y análogamente para las difíciles donde nunca se produce la respuesta correcta. Realizar más llamadas no necesariamente es mejor, entonces es fundamental comprender cómo la dificultad de las consultas afecta el rendimiento. Esto da pie a que se debe de considerar la respuesta guiando al modelo sobre diferentes formas de razonar al adicionar a las solicitudes información diferente para poder permitir nuevas "líneas de razonamiento".

Investigaciones recientes, como (S. Feng y cols., 2024), proponen soluciones innovadoras que utilizan enfoques colaborativos multimodelo: COOPERATE y COMPETE. Es decir, utilizan múltiples modelos que colaboran para identificar cuándo abstenerse de responder y mejorar la capacidad de los LLMs para reconocer sus limitaciones y evitar errores, especialmente en tareas complejas que requieren un razonamiento avanzado. Un problema clave es el sesgo del *prompt*, ya que basarse en una sola solicitud para juzgar un texto introduce sesgos y no capta la complejidad multidimensional del problema. Además, los LLMs a menudo no logran emular el proceso jerárquico y exhaustivo de los expertos humanos, que tienen en cuenta múltiples perspectivas y subcriterios en sus juicios. Esto ha suscitado dudas sobre la fiabilidad de las evaluaciones automatizadas basadas en LLMs, ya que sus juicios no siempre reflejan los matices o las prioridades humanas.

En lugar de simplemente predecir una única etiqueta, los LLMs pueden proporcionar una representación más completa del panorama de acuerdo a la integración de la información externa, lo que es esencial para atender una comprensión matizada del lenguaje. La idea es considerar diferentes perspectivas de acuerdo a la información proporcionada, ya que normalmente se usa una serie de *prompts* con la misma información presentada de forma distinta. Utilizar ToT con agentes (Haji y cols., 2024), más aún asumiendo roles especializados o proporcionando la salida de un LLM a otro o

utilizando diferentes agentes para posteriormente unificar su respuesta final. En contraste, el método Soft Self-Consistency (SOFT-SC) constituye una sofisticada extensión de la técnica Self-Consistency (SC) , diseñada específicamente para superar sus limitaciones en entornos interactivos. Mientras que SC, efectiva en tareas de razonamiento, emplea un voto por mayoría basado en coincidencia exacta entre múltiples cadenas de pensamiento, su eficacia decae en dominios con espacios de acción amplios y respuestas válidas diversas (H. Wang y cols., 2024).

En este sentido, para consolidar una respuesta de múltiples consultas, existen estrategias donde no intervienen agentes, como *Majority Vote* (MV) que aumenta la confiabilidad de la solución final, ya que se basa en el acuerdo de las predicciones de múltiples modelos funcionando de forma independiente. (Dogan y Birant, 2019b) propone un suavizado a un común MV, al introducir el aporte de la solución final en cada respuesta por modelo. Esto permite dar más importancia a las respuestas de los modelos que han demostrado ser más precisos, es decir, recompensa a los modelos que predicen correctamente los ejemplos donde los otros modelos fallan. Por otro lado, los autores (Xue, Liu, Lei, Xingzhang, y cols., 2023) proponen Dynamic Voting (DV) con el objetivo de tener rendimientos equiparables utilizando significativamente menos caminos de razonamiento que el ToT. El DV introduce la idea de "salida temprana", donde el proceso puede terminar antes si se alcanza un cierto umbral de consistencia, es decir, si las respuestas a las preguntas por varias rutas de razonamiento dan una misma respuesta.

En un esquema tradicional de votación mayoritaria, se parte del supuesto de que, cuando se formula la misma pregunta, independientemente de cómo se plantee al modelo, este debe llegar a la misma respuesta y ser coherente en su razonamiento. En otras palabras, el modelo tendría que seguir la misma línea de razonamiento independientemente de cómo se le plantee la pregunta. Sin embargo, se ha demostrado que los LLMs pueden simular patrones de razonamiento no monótonos en escenarios controlados, gestionando excepciones sin ignorar las reglas generales. Sin embargo, esta capacidad sigue siendo superficial y frágil (Leidinger y cols., 2024; Z. Li y cols., 2025). Cuando se exponen a información irrelevante o contradictoria, los LLMs tienden a abandonar las creencias previamente establecidas. No logran demostrar de manera consistente el dominio de esta capacidad de una manera robusta, generalizable y lógicamente coherente, tanto en contextos de lenguaje natural como en lógica formal (Xiu y cols., 2022; Leidinger y cols., 2024; Z. Li y cols., 2025).

Las técnicas más avanzadas de *prompting* buscan superar el procesamiento estadístico superficial y no solo dotar a los LLMs de información contextual relevante, sino también explotar y reorganizar su conocimiento, guiándolo hacia formas de pensamiento estructurado. Los LLMs por una parte poseen un límite en sus capacidades. El objetivo de esta investigación es superarlo mediante la integración estructurada de bases de conocimiento explícitas y mecanismos de razonamiento, permitiendo que el LLM, más que simular procesos cognitivos, pueda acceder, manipular y fundamentar conceptos de manera abstracta y lógica, cerrando así la brecha la coherencia y comprensión lingüística.

Capítulo 4

Reconocimiento de la Implicatura Textual (RIT)

En el presente estudio, para validar la propuesta se utiliza la tarea: el Reconocimiento de la Implicatura (o entrañamiento) Textual (RIT). La tarea de RIT es muy importante en el marco del PLN ya que captura la principal necesidad de inferencia semántica de muchas otras tareas, tales como: la respuesta a preguntas; el resumen automático de textos; la recuperación de información; la extracción de información y la traducción automática, entre otras. Es así que el desempeño de estas aplicaciones depende en gran medida de resolver la implicatura textual.

Desde que (Dagan y cols., 2006) promueven los retos de RIT, el interés por resolver la tarea de implicatura textual ha crecido hasta la fecha. Esta tarea requiere reconocer cuándo el significado de un texto está contenido en el significado de otro. Es decir, dadas las siguientes frases:

P: El perro café y blanco está jugando en el pasto.
H: El perro juega en el pasto verde.

La tarea consiste en determinar si una pieza de texto se sigue lógicamente[1] (se infiere) de otra pieza de texto. Entonces, podemos decir que P implica (o entraña) a H ($P \rightarrow H$) si H se sigue a partir de saber P. Analizando el ejemplo anterior, se sabe que el pasto por lo general es verde (dado por H), y tenemos el antecedente que un perro esta jugando en él (dado por P), entonces podemos inferir que H es verdadera siempre y cuando P sea verdadera. Para los humanos resulta sencillo hacer este tipo de inferencias, pero, ¿Cómo se puede acercar de manera computacional?

[1] A lo largo de esta tarea, utilizamos el término "se sigue lógicamente" o "implica" en un sentido amplio, propio del razonamiento y la comprensión del discurso natural.

Para abordar la implicatura textual, se han propuesto a través de los años diferentes enfoques y modelos. Los enfoques propuestos de RIT pueden clasificarse en función de los fenómenos lingüísticos involucrados, el tipo de representación de P y H y como se decide si existe implicatura o no. Este análisis nos lleva a identificar dos mecanismos principales en RIT:

- La representación.
- La inferencia.

Por un lado, la representación se refiere a la forma de abstraer el contenido léxico-semántico tanto de la Premisa como de la Hipótesis. En el estado del arte, podemos encontrar enfoques que van desde la construcción de conjuntos de palabras de P y H con el fin de enriquecerlos, ya sea utilizando el mismo corpus, o bien mediante bases de conocimiento externas. Al utilizar el mismo corpus, las redes neuronales profundas son, de lejos, las mejores opciones para capturar la semántica distribucional (co-ocurrencias frecuentes con potencial de predicción). Los modelos de redes neuronales profundas han dado un gran salto al enfrentarse a diversas tareas planteadas en PLN.

Por otro lado, el mecanismo de inferencia se refiere la forma en que se decide si la implicatura ocurre; es decir, si podemos establecer que $P \rightarrow H$ se cumple. Al respecto, los enfoques van desde ver el problema como una tarea de clasificación, utilizando herramientas del aprendizaje automático o redes neuronales, hasta la construcción de grafos de inferencia o de medidas de similitud para tomar la decisión.

Para poder evaluar el rendimiento de los modelos de aprendizaje automático se utilizan corpus etiquetados para la tarea de RIT. Específicamente, se han desarrollado los corpus RTE-1 a RTE-7. Sin embargo, la investigación sobre aprendizaje automático se ha visto muy limitada por la falta de recursos a gran escala, es así que se ha realizado propuestas para extender los corpus de RIT usando recursos como Wikipedia (Zanzotto y Pennacchiotti, 2010), esto con el fin de hacerlos homogéneos en los fenómenos lingüísticos involucrados y con más ejemplos. En este sentido, (Bowman y cols., 2015) presentan el corpus SNLI y MultiNLI (Williams y cols., 2018), que introdujeron la categoría *neutral*, ampliando el espectro del razonamiento requerido. Para evaluar específicamente la semántica composicional, se creó el corpus SICK (*Sentences Involving Compositional Knowledge*) (Marelli y cols., 2014). Otros corpus especializados, como SciTail (Khot y cols., 2018), derivado de preguntas de ciencia, y el corpus de diagnóstico de SuperGLUE (A. Wang, Pruksachatkun, y cols., 2019a), han seguido impulsando la complejidad y diversidad de la tarea.

(Poliak, 2020) por su parte presenta un estudio para evaluar y comprender las capacidades de implicatura de los sistemas de PNL. La investigación se centra en el debate de los corpus dedicados para RIT, así como los avances en otros corpus que se centran en fenómenos lingüísticos específicos que

pueden utilizarse para evaluar los sistemas de PNL a un nivel de detalle. Existen diferentes fenómenos que ocurren en la decisión de la implicatura por lo que se debe contar con corpus dedicados a tratarlos. The *General Language Understanding Evaluation* (GLUE) [2] proporciona una colección de recursos para evaluar diferentes tipos de fenómenos lingüísticos que ocurren en el RIT, a saber: Semántica Léxica, Estructura Predicado-Argumento, Lógica, y Conocimiento y Sentido Común.

En la tarea de RIT, hoy en día existe un gran avance en cuestión del rendimiento de los modelos de redes profundas de aprendizaje[3]. No obstante su alto rendimiento, estos modelos distan de ser claros y transparentes en la decisión de implicatura. Consideramos que para ahondar en la comprensión lingüística, no solo es suficiente contar con buenos rendimientos, sino que es importante desarrollar modelos que permitan entender la toma de decisiones que realiza una arquitectura para resolver la implicatura textual. En contra parte, el rendimiento (bajo) de muchos algoritmos de aprendizaje automático depende críticamente de la pertinencia de las características extraídas y representadas, aunque con un poco más de transparencia.

Desde enero 2023 a la actualidad, el benchmark *Natural Language Inference on RTE*[4] coloca en primer lugar al modelo *Vega V2* (Q. Zhong y cols., 2022) con un rendimiento en *accuracy* del 96 %. El segundo lugar se encuentra *PALM* (Chowdhery y cols., 2022) con un *accuracy* del 95.7 %. La diferencia en el rendimiento es mínima, pero abismal en el número de parámetros de cada modelo: 6B vs 540B, respectivamente. A pesar de que los PLM logran tener un alto rendimiento, siguen siendo *cajas negras* puesto que no se sabe que ocurre o qué es lo que se toma en cuenta para tomar una decisión en el RIT.

En el corpus de diagnóstico los resultados generales del modelo *Vega V2* (que ocupa el primer lugar) de su matriz de confusión son la mayor parte prediciones de "Entailment", aunque pertenezcan a la clase "Not entailment", obteniendo un *accuracy* de 43 % de un total de 1,104 ejemplos. Un modelo *Majority class* que predice todos los ejemplos como "Not entailment" tendría un *accuracy* del 58 %. En cambio *PALM* (segundo lugar) obtiene un mejor equilibrio sobre el mismo corpus, en las dos clases predichas, con un 87 % de *accuracy*.

Esto da cuenta de que, los modelos se entrenan para obtener rendimientos superiores en la tarea sobre un corpus específico (el que se entrena) y no atender verdaderamente la tarea del RIT. Existe un gran número de fenómenos lingüísticos que ocurren en la implicatura textual (Poliak y cols., 2018), por lo cual es la necesario una evaluación de las capacidades lingüísticas de los modelos, como propone *Super GLUE Benchmark*[5], para identificar limitaciones y avanzar en el desarrollo de sistemas más robustos y consistentes.

[2] https://gluebenchmark.com/diagnostics
[3] https://paperswithcode.com/sota/natural-language-inference-on-rte
[4] https://paperswithcode.com/sota/natural-language-inference-on-rte
[5] https://super.gluebenchmark.com/diagnostics

El mayor rendimiento se obtienen de los modelos de aprendizaje profundos mostrando un comportamiento aparentemente inteligente, pero crea una problemática entre la exactitud e interpretabilidad (Arrieta y cols., 2020). Se han propuesto distintos métodos para ayudar a interpretar las las respuestas de modelos complejos. Sin embargo, todavía no se sabe cuándo es preferible uno que otro. (Lundberg y cols., 2017) proponen un marco unificado de interpretación de predicciones: SHAP (SHapley Additive exPlanations). SHAP asigna a cada característica un valor de importancia (o contribución) para las prediciones realizadas por el modelo, que muestran un mejor rendimiento computacional y/o una mayor coherencia de interpretabilidad con la intuición humana.

Al tener un modelo no transparente con respecto a la explicación y el razonamiento siguen siendo las *black boxes*, es decir, no sabemos si el aprendizaje es válido y generalizable o si el modelo basa su decisión en una correlación falsa de los datos de entrenamiento. Entender como el modelo realiza sus predicciones, lejos de la parte técnica, es importante para obtener un mejor rendimiento, y es necesario ir más allá de las métricas de evaluación para lograr la comprensión en la toma de decisiones y acercarnos a modelos de IA que sean transparentes. (Lapuschkin y cols., 2019) proponen su modelo de Análisis de Relevancia Espectral semiautomatizado, que proporciona una forma de caracterizar y validar el comportamiento de los modelos. Su afirmación es que para generar una explicación no es necesario entender las neuronas individualmente, si no centrarse en explicaciones de ejemplos individuales de la predicción; características o píxeles de la entrada que fueron más importantes para una predicción.

Dentro del marco de la Comprensión del Lenguaje Natural (CLN), (A. Wang, Singh, y cols., 2019; Nangia y Bowman, 2019) muestran que más allá generar modelos que capturen relaciones superficiales de la entrada y salida, es necesario que atiendan los diferentes fenómenos lingüísticos de la tarea en diferentes dominios. La tarea del RIT aún sigue siendo un desafío para los enfoques de redes neuronales en la comprensión del significado. Es así, que para poder abordar la tarea del RIT es necesario analizar los diferentes aspectos lingüísticos involucrados en los dos mecanismos de la implicatura. Para lograrlo, no basta con comparar palabras superficialmente (similitud coseno de WE); es necesario captar el significado subyacente, las implicaciones contextuales y, en muchos casos, realizar saltos inferencias que no solo sean explicitas. Aquí es donde la abstracción se vuelve crucial. Un mecanismo de abstracción semántica puede reconocer frases que expresan ideas con el mismo significado, a pesar de las diferencias en sus unidades léxicas.

A continuación se presentaran algunos trabajos en materia del RIT y como es que abordan las diferentes fenómenos lingüísticos que están involucrados en esta tarea, permitiendo así, generar explicaciones transparentes que justifiquen la decisión de la implicatura textual. De la misma forma existen algunos trabajos que utilizan una arquitectura neuronal que mejoran el rendimiento en la tarea del RIT pero la decisión del modelo no es explícita.

4.1. RIT explícita

La generación de algún tipo de justificación en lenguaje natural o algunas medidas de similitud que sean comprendidas intuitivamente por el humano es una característica importante para aumentar la interpretabilidad de un sistema, y para estas explicaciones se han aprovechado fuentes externas de conocimiento del mundo, que ya se encuentran estructuradas y de fácil acceso.

Los enfoques clásicos generalmente realizan un análisis del par $< P, H >$ sobre dependencias de las palabras, identificación de roles semánticos, análisis a nivel léxico, sintáctico y semántico, léxico-semántico. De la misma forma, se utilizan recursos externos que permiten identificar relaciones entre unidades léxicas del par $< P, H >$, que se encuentran implícitas en la decisión de implicatura textual, a saber: sinonimia, antonimia, hiponimia, hiperonimia, entre otras.

Cuando se trabaja con estructuras sintácticas se intenta desambiguar la estructura de la oración. (Kouylekov y Magnini, 2006) adoptaron un algoritmo de distancia de edición aplicado a los árboles de dependencia de P y de H. Se intenta transformar el árbol sintáctico de la premisa en el árbol sintáctico de la hipótesis a través de operaciones de edición. Las operaciones que se pueden realizar sobre los árboles sintácticos son eliminación, substitución y adición de elementos. Estas operaciones tienen un costo por lo que la implicatura se decidirá si se cumple lo siguiente:

$$ed(P, H) < \gamma.$$

donde $ed(P, H)$ es la suma de los costos de las transformaciones aplicadas. Si este valor es menor a un umbral γ entonces la implicatura se cumple.

Para enfrentar el problema de la ambigüedad semántica, un método tradicional es el del uso de plantillas. Las plantillas capturan relaciones semánticas que no se encuentran presentes. Para resolver la implicatura a través de plantillas se debe contar con relaciones equivalentes en el par $< P, H >$. Un ejemplo de plantilla que captura una relación equivalente es el siguiente:

- X *excavó* Y
- Y *excavadoPor* X

Existen enfoques para encontrar relaciones semánticas en corpus que son importantes para capturar el significado. Para atender la tarea Open Information Extraction (OIE), (Bovi y cols., 2015) realizaron un análisis sintáctico-semántico de las definiciones textuales para generar representaciones de información textual estructuradas y legibles por las máquinas. Esta tarea se encuentra inmersa en el ámbito de Extracción de información (EI).

Los autores generan un grafo de definiciones con el objetivo final de representar las relaciones

semánticas entre unidades léxicas. La arquitectura propuesta por los autores DEFIE permite extraer información basada en el análisis sintáctico-semántico y desambigüación semántica a partir de definiciones textuales. Un ejemplo de relaciones extraídas es la siguiente: $X \rightarrow is \rightarrow album_1 \rightarrow by \rightarrow Y$, donde podemos modificar las variables X y Y por unidad léxica[6]. Este tipo de enfoques permite encontrar una gran cantidad de relaciones semánticas que además también permite desambigüar las palabras en el contexto que se desenvuelven.

(Silva y cols., 2020), por otro lado, propone la construcción de un grafo de conocimientos[7] a partir de definiciones textuales, para apoyar las justificaciones en la decisión de implicatura textual. El grafo de conocimientos se construye de las definiciones de diccionario a partir del análisis del glosario de *WordNet*. La base fundamental del análisis son los conceptos de genus-differentia de la Teoría de la definición de Aristóteles y son lo suficientemente generales para definir conceptos de cualquier campo del conocimiento. Un ejemplo de definición es el siguiente: el ser humano es un animal racional. El animal es el "genus" y racional es la "differentia" que distingue al hombre de los demás animales. Cuando una implicatura se cumple, las palabras presentes en el par $< P, H >$ tienen una fuerte relación semántica, y entonces es posible encontrar un camino en el grafo de definiciones que vincule a cada una y justifique la implicatura. Siendo el camino en el grafo la representación de la justificación de la decisión.

4.2. RIT neuronal

Como hemos mencionado anteriormente, hay un predomino en usar redes neuronales que abstraen en vectores de representación la información de P y H para medir las diferencias en un plano matemático. Algunas propuestas intentan capturar las relaciones semánticas que ocurren entre el par y codificarlos en un vector numérico para poder medir diferencias.

Por ejemplo, para resolver esta tarea, (Ríos y cols., 2010b) reducen la ambigüedad léxica identificando relaciones causales, del tipo $causa \rightarrow efecto$; donde el efecto es una consecuencia directa de la causa. El enfoque de los autores es que la relación de implicatura textual $P \rightarrow H$, es una relación causal: la causa es la premisa y la hipótesis, el efecto. Para poder realizar su modelado, primero obtienen un conjunto de relaciones causales que utilizan el conectivo discursivo "because" (X *because* Y). Este conjunto se extrae a partir de un corpus (ukWAC[8] del motor Sketch) y sirve para representar las relaciones de los elementos del par $< P, H >$ en un contexto. Este mismo enfoque les funciona para enfrentar la ambigüedad léxica y estructural que hemos mencionado.

[6] El subíndice de "album" hace referencia a un sentido en particular, pudiendo contar con varios sentidos.
[7] Estructura de nodos y aristas, donde los nodos son los conceptos y las aristas son las relaciones que los une
[8] https://www.sketchengine.eu/ukwac-british-english-corpus/

En materia de desambigüación semántica, los modelos de aprendizaje profundo sobresalen (Bengio y cols., 2003; Mikolov y cols., 2013; Devlin y cols., 2018). Estos modelos generan representaciones de las palabras que encapsulan similitud en términos de su co-ocurrencia con otras palabras en contextos similares. Para calcular la similitud semántica entre las palabras se utiliza comúnmente la distancia Euclidiana o la distancia coseno sobre los vectores de representación.

Este enfoque de representar palabras, que se extiende a frases completas, se considera uno de los avances más importantes del aprendizaje profundo en la semántica. La característica más interesante de estos vectores es la preservación del "significado"(contextual), ya que logran capturar relaciones semánticas en distintos grados de similitud. Los elementos o frases del texto que sean similares tendrán distancia coseno cercana a 1 o distancia Euclidiana cercana a 0.

En el enfoque de relación causal, (Ríos y cols., 2010b) abordaron la tarea de RIT a través del aprendizaje automático como un problema de clasificación. El vector de características propuesto por los autores que representa cada par $< P, H >$ esta dado por dos valores: una medida de similitud coseno y una medida de similitud no simétrica. Los autores proponen una medida de similitud semántica no simétrica la cual cumple lo siguiente: $sym_{ns}(P, H) \neq sym_{ns}(H, P)$. La base del modelo de los autores es que H es menos informativa que P. Es así, que para poder tomar una decisión de implicatura debe ocurrir la siguiente desigualdad: $sym_{ns}(P, H) > sym_{ns}(H, P)$. Para calcular la medida de similitud semántica no simétrica se usa la siguiente fórmula:

$$sym_{ns}(P, H) = \sum_{j}^{n} max_{p_i}(\frac{ce_{h_j}}{c_{h_j}}),$$

donde ce_{h_j} es la ocurrencia causal de cada termino de la Premisa p_i con la de la Hipótesis h_j en el conjunto de relaciones causales $\{ C \text{ because } E \}$ tales que $p_i \in C, h_j \in E$ y c_{h_j} es la ocurrencia de h_j en las relaciones causales. Para poder medir el rendimiento de su modelo utilizaron el corpus *RTE-1* obteniendo en exactitud el 51 % de aciertos en la tarea.

(Ríos y cols., 2010a) evaluaron el desempeño de su modelo utilizando el rendimiento en diferentes algoritmos, tales como: SVM, Naive Bayes, AdaBoost, BayesNet, LogitBoost y Perceptron. Al comparar los resultados obtenidos con otros sistemas de aprendizaje automático, su modelo con Naive Bayes obtuvo un 63.5 % de exactitud. La conclusión fue que el rendimiento de muchos algoritmos del aprendizaje automático depende de la pertinencia de las características extraídas y representadas.

Para 2015, la tendencia dominante era construir modelos complejos y profundos con redes neuronales para poder representar textos con el objetivo de profundizar en la comprensión del lenguaje y resolver la implicatura textual. (Rocktäschel y cols., 2015) proponen un modelo de red recurrente neuronal con dos *Long short-term memory* (LSTM) (Hochreiter y Schmidhuber, 1997) que recibe como entrada un par de elementos $< P, H >$ y genera representaciones de cada frase. También cuenta con

un mecanismo de atención neural (Bahdanau y cols., 2014) sobre las palabras. La evaluación de la red neuronal propuesta se realizó con el corpus SNLI obteniendo en exactitud el 83.5 %.

(Parikh y cols., 2016) proponen un modelo con el fundamento de que no se requiere un modelado profundo de la estructura de la oración. Tan solo se requiere comparar la estructura local de la premisa y la hipótesis para realizar inferencias globales. La motivación de los autores se basa en la alineación de palabras/subfrases como en (Bahdanau y cols., 2014). La intención de la alineación es descomponer el problema de la implicatura en subproblemas. El mecanismo funciona adecuadamente cuando las palabras de la premisa y la hipótesis están conectadas a través de algún tipo de relación semántica, por ejemplo: alinea las palabras "park" con "outside", "alice" con "someone" y " flute+solo" con "music" (Figura 4.1). El rendimiento obtenido es de 86.8 % en exactitud, sobre el corpus SNLI. Los autores concluyen que las comparaciones por pares son relativamente más importantes que las representaciones globales a nivel de frase y mostraron que el costo computacional se puede reducir.

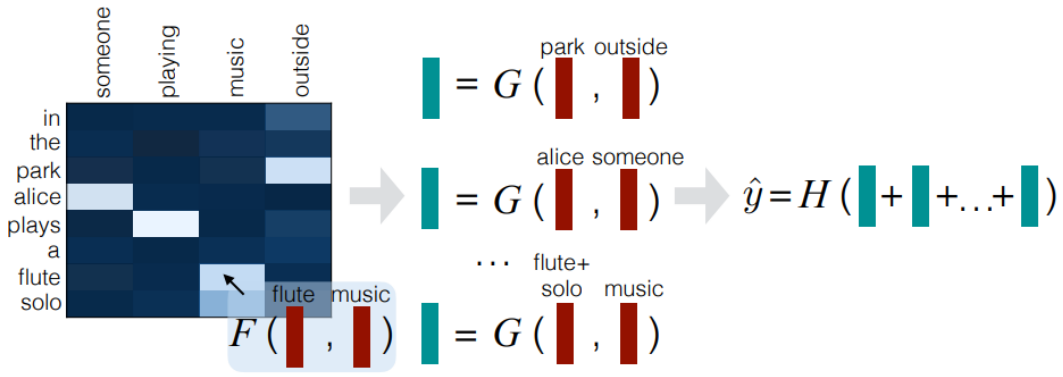


Figura 4.1: Resumen gráfico del enfoque de (Parikh y cols., 2016)

Nota. Adaptado de *A Decomposable Attention Model for Natural Language Inference* , por (Parikh y cols., 2016).

Los diferentes trabajos en este ámbito han intentado capturar relaciones entre los elementos de P y de H en vectores de *features* que por lo general son métricas estadísticas, por ejemplo, el solapamiento de palabras, relación de longitud, similitud coseno, similitud de subcadenas, puntuación Jaccard, BoW, similitud de representación de oraciones, distancia Manhattan, entre otras, delegando la toma de decisiones del RIT a las redes neuronales. De esta forma se han tenido buenos resultados pero con una poca explicabilidad.

(Shajalal y cols., 2023) aprovechan los WE para obtener información que distingue entre la relación de implicatura. El método parte de la representación semántica vectorial de las palabras, donde cada elemento del vector es un número real que captura su significado contextual. Si bien los enfoques clásicos obtienen la representación de una oración calculando simplemente la media aritmética de los vectores de sus palabras, esta técnica tiene una limitación fundamental: no todos los valores en los

vectores de palabras son relevantes para expresar el significado específico de un par $< P, H >$. Para superar esto, el enfoque propuesto implementa una representación semántica empírica basada en un umbral, que filtra sistemáticamente aquellos elementos vectoriales que no contribuyen significativamente o distorsionan la información contextual, permitiendo así una representación más precisa y eficiente para la tarea de reconocimiento de implicatura.

Para abordar la semántica se puede recurrir a recursos externos. El problema de los recursos externos radica en como vincularlos con la información aprendida de los modelos de redes neuronales. Investigaciones han intentado usar esta información sobre las relaciones de las partes textuales de la premisa y de la hipótesis (Lauscher y cols., 2020). (X. Wang y cols., 2018; Kapanipathi y cols., 2020) proponen distintos métodos para extraer el conocimiento relevante en la tarea del RIT y demuestran que el rendimiento de los modelo es comparable al estado del arte al combinar representaciones de los textos y sus grafos obtenidos de recursos externos. (Jiang y cols., 2019; Guo y cols., 2022; Q. Chen y cols., 2018) proponen integrar dinámicamente el conocimiento externo de ConceptNet, WordNet o Wikipedia en modelos de redes neuronales (BiLSTMs) para facilitar la comprensión de las frases. La propuesta se basa en el refinamiento incremental de las representaciones de palabras, es decir, integrando información externa en la representación de la frase, explotando conocimientos previos de un modo semánticamente apropiado.

Los PLM han obtenido buenos rendimientos en diversas tareas del PLN, aunque aún siguen teniendo incoherencias al tratar con el texto. Para abordar esta problemática (Zi y cols., 2023a) (X. Yang y cols., 2019) buscan combinar el conocimiento de ConceptNet con las representaciones semánticas de BERT, logrando resultados equiparables de su modelo en las tareas del *GLUE benchmark*. Esto da pie, que al proporcionar información de relaciones semánticas apoya al tratamiento de la tarea, a la decisión y a la interpretabilidad de los modelos, debido a que permite distinguir cuando se usa la semántica distribucional de los WE y cuando se usa la información de ConceptNet.

En la era de los LLMs, (Madaan y cols., 2024) investigan si los benchmark del RIT todavía pueden desempeñar un papel en la evaluación de estos con una única clase objetivo. Los autores descubren que los LLMs de diferente tamaño y calidad entrenados con los corpus del RIT, brindan señales adecuadas para discriminar entre clases. Por otro lado, en la distribución de opiniones humanas, por ejemplo, el corpus *ChaosNLI*, los autores (Nie y cols., 2020b) muestran que existe similitud de las distribuciones de predicciones de modelos con las distribuciones humanas y además ésta aumenta con la escala de los modelos. A continuación ahondaremos en las problemáticas actuales de los LLMs, para analizar la relevancia de la tarea del RIT con los LLMs.

Como hemos mencionado, los LLMs carecen de abstracción semántica, aunque pueden generar respuestas coherentes y realizar algunas formas de razonamiento. Pueden imitar el conocimiento, pero

no lo internalizan como conceptos independientes del lenguaje mismo. Esta limitación se hace evidente en tareas donde se requiere el significado implícito más allá de lo explícito, por ejemplo en la tarea del RIT. Por ejemplo, en la frase "Luis compró flores para su esposa por su aniversario", algo implico es "Luis está casado". Este tipo de relaciones se puede obtener de recursos externos. En este sentido, existen enfoques como el de (Y. Wang y cols., 2024) que utilizan el RIT como estrategia para retroalimentar a los LLMs y mejorar sus respuestas. Las ventajas de usar LLMs en la tarea, vienen en parte de su capacidad para procesar grandes cantidades de texto, relacionar consecuencias lógicas de las frases, obtener información implícita del contexto, identificar patrones complejos y generar respuestas coherentes y contextualmente relevantes.

(Zi y cols., 2023b) por su parte intentan mitigar estas problemáticas en los resultados de los LLM. Presentan su propuesta IERL (Interpretable Ensemble Representation Learning) el cuál integra las representaciones de BERT y los grafos de conocimiento de un recurso externo, en este caso ConceptNet. Por un lado abonan a la interpretabilidad del modelo, la cual es guía en nuestra investigación, debido a que identifican cuál es el uso del contexto del LLM y también del grafo de conocimiento proporcionado para la toma de la decisión de la tarea. Por otro lado, impulsan la implementación de estrategias para incluir información de recursos externos y utilizar las capacidades de BERT. Su modelo logra obtener en el RTEGLUE el 92.3 % contra BERT base con 88.3 %.

Otra estrategia propuesta por (Lauscher y cols., 2020) utiliza adaptadores entrenados con conocimiento externo y permite conservar el rico conocimiento distributivo adquirido en el preentrenamiento. (X. Wang y cols., 2018; X. Yang y cols., 2019; Kapanipathi y cols., 2020) proponen diferentes métodos para extraer el conocimiento relevante en la tarea del RIT y demuestran que el rendimiento del modelo es comparable al estado del arte cuando se combinan las representaciones de los textos y sus grafos obtenidos de recursos externos.

(K. Yang y cols., 2022) proponen NLProofS que se enfoca en la generación de pruebas por pasos condicionados a la hipótesis y hechos de apoyo, limitando al modelo a generar sólo pasos relevantes y evitar alucinaciones con ayuda de un verificador basado en RoBERTa. Los autores mencionan que podría tener problemas con frases más largas o a un mayor número de pruebas de hechos y que es difícil evaluar los pasos de pruebas válidos.

Para validar razonamientos sobre enunciados escritos en lenguaje natural, estudios como (Sanyal y cols., 2022) proponen un enfoque modular usando transformadores llamado Faithful and Robust Reasoner (FAIRR). Los modulos funcionan de forma independiente: selección de reglas, selección de hechos y composición de conocimientos. Los autores ahondan en la interpretabilidad debido al enfoque modular y al generar una cadena de pruebas que son más fáciles de entender para el usuario. Sin embargo, los métodos tienen un problema en la generación de pasos válidos y pasos relevantes lo cuál

influye en la decisión final. Por otro lado, al trabajar con lenguaje natural los LLMs podrían generar diversas rutas de razonamiento o alucinaciones que no sustente la decisión. Por último, (Kim y cols., 2024) proponen usar enlaces virtuales (conectivos semánticos), emulando la estructura sintáctica al unir la premisa y la hipótesis. Sus resultados muestran que su propuesta mitiga eficazmente usar solo una heurística simple, de suporsición léxica y presencia de negación.

Más aún, el estudio de (Nie y cols., 2020a) revela que existe un alto grado de desacuerdo entre los humanos en una cantidad considerable de ejemplos de corpus del RIT. Este desacuerdo cuestiona la práctica común de usar la clase mayoritaria en la evaluación de RIT. (Zhou y Bhat, 2021) por su cuenta, destaca la necesidad de diseñar modelos que puedan modelar explícitamente las opiniones humanas colectivas y comprender las fuentes del desacuerdo humano en relación con los diferentes fenómenos lingüísticos de las tareas de PLN.

Dentro de la tarea del RIT se han abordado dichos problemas del lenguaje para mejorar el rendimiento de sus modelos (Khot y cols., 2018). Estas redes neuronales obtienen buenos rendimientos en corpus grandes pero en contraste, su rendimiento cae sobre corpus con menos datos, por ejemplo, el modelo de (Parikh y cols., 2016) tiene un rendimiento[9] de 86.8 % (accuracy) sobre el corpus SNLI y su rendimiento en el Scitail es de 72.3 % (Khot y cols., 2018). Esto deja entrever que su gran capacidad para encontrar relaciones esta directamente relacionado con la cantidad de información que procesa, sin embargo, con pequeñas cantidades de entrenamiento encuentra correlaciones espurias entre los datos. Todo indica que los modelos solo aprenden patrones superficiales y muestran una ilusión del rendimiento real.

El análisis de la literatura revela cuatro problemas fundamentales en los abordajes actuales para el RIT: 1) dependencia de correlaciones estadísticas superficiales que queda evidenciada cuando modelos con alto rendimiento caen drásticamente con otros corpus, demostrando que aprenden patrones espurios del corpus en lugar de razonamiento genuino; 2) estrategias de generación de pruebas como NLProofS y FAIRR padecen de alucinaciones y pasos inválidos, generando cadenas de razonamiento cuya fidelidad es difícil de verificar; 3) incapacidad para manejar la direccionalidad jerárquica, modelos que no distinguen que “perro” implica “animal” pero no a la inversa, y 4), la limitada capacidad para integrar conocimiento externo de manera estructural, pues enfoques como IERL combinan representaciones vectoriales sin trascender la lógica estadística, mientras que otros se limitan a inyectar hechos sin reorganizar el proceso inferencial.

Nuestra propuesta aborda precisamente estas limitaciones al abstraer el conocimiento en categorías que imponen restricciones direccionales, sustituir la dependencia estadística por un andamiaje semántico estructurado y ofrecer un mecanismo de consenso interpretable para el razonamiento.

[9] SNLI cuenta con 550 mil ejemplos (pares $< P, H >$), mientras que le Scitail tiene 26 mil.

Capítulo 5

Marco Metodológico: Abstracción de Conocimiento Semántico (SKA)

La motivación cognitiva constituye una de las cuatro categorías superiores en la taxonomía que estudia la generalización dentro del PLN: *Practical*, *Cognitive*, *Intrinsic* y *Fairness and inclusivity* (Hupkes y cols., 2023). Esta motivación orienta el diseño de los experimentos y condiciona la interpretación de los resultados acerca de la capacidad generalizadora de un modelo. El enfoque fundamental es el del "Comportamiento del Modelo", que examina si la generalización de los sistemas artificiales se asemeja a la humana, utilizando esta última como medida de inteligencia (Hupkes y cols., 2023).

Es así que, a partir de la forma en que los seres humanos resuelven la tarea del RIT, se propone un marco metodológico que incorpora procesos cognitivos clave, como la composición del significado, la abstracción de conocimiento semántico, la interpretación semántica y la inferencia. La abstracción en sí, es un mecanismo en la cognición humana que consiste en aislar y/o extraer características o propiedades esenciales de un objeto, para operar con representaciones simplificadas pero significativas y aplicarlas a distintos objetos similares.

La propuesta planteada en esta tesis es establecer un mecanismo de abstracción de conocimiento semántico de forma jerárquica que ayuda a identificar y cubrir lagunas de conocimiento de los LLMs en relación con el RIT. La clave de nuestra propuesta es modelar los aspectos cognitivos que están implícitos en el NLP, como una forma de tratar los distintos fenómenos lingüísticos implicados en la tarea. Como parte del modelo propuesto para atender la tarea del RIT se han identificado dos grandes mecanismos: a) La representación y b) La Inferencia.

La base para el mecanismo de representación se resume en la Figura 5.1. Se comienza definiendo dos categorías abstractas: **Compatibilidad semántica** e **Incompatibilidad semántica**. El siguiente paso es el de composicionalidad, que implica el análisis del árbol de dependencias de P y H para identificar entidades y sus atributos correspondientes. Con este proceso se obtienen entidades y atributos de cada P y H. Entonces, se aprovechan las capacidades de ConceptNet para identificar las relaciones semánticas entre estas entidades y agruparlas según la definición de los grupos abstractos. Finalmente, se alinean los grupos abstractos de acuerdo a las clases de la tarea del RIT.

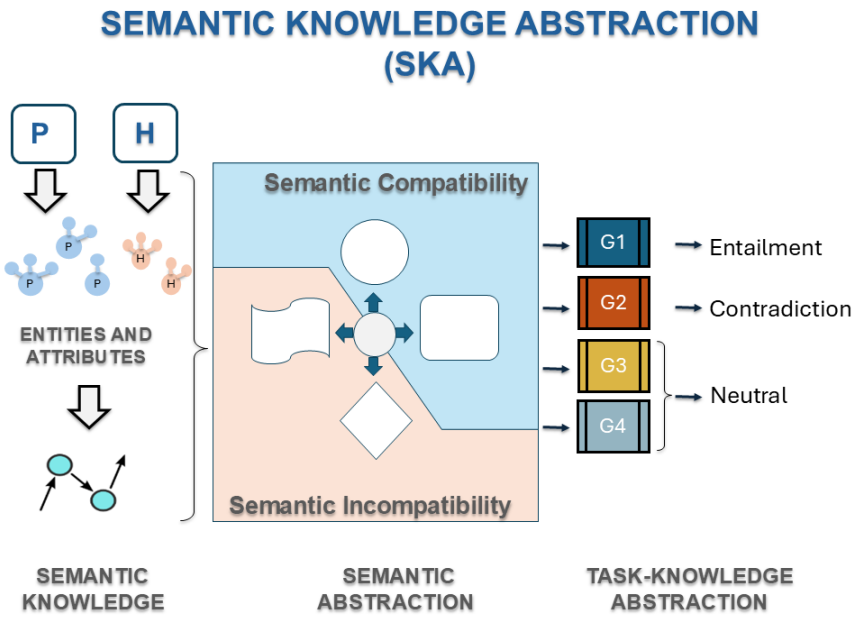


Figura 5.1: Marco metodológico SKA.

Nota. Adaptado de *Semantic knowledge abstraction: Consistent reasoning in large language models for natural language inference*, por (Torres-Moreno y Hermosillo-Valadez, 2026).

La Figura 5.1 muestra el marco metodológico propuesto: A partir de la premisa y la hipótesis, se identifican las entidades y sus atributos correspondientes, que se utilizan para extraer conocimiento semántico de ConceptNet, el cual se clasifica en grupos abstractos. El conocimiento de la tarea se representa como una alineación entre los grupos abstractos y las clases de tareas objetivo. En la siguiente subsección se ahonda en cada una de las etapas del marco metodológico.

El mecanismo de representación genera tripletas de relaciones sobre los grupos abstractos. Para el mecanismo de inferencia, se optó por probar el alcance de la representación a través de las siguientes técnicas de prompting: A partir de las respuestas de los LLMs, es posible analizar las rutas de razonamiento y, finalmente, proponer un modelo de decisión a través del razonamiento o debate.

El LLM genera cuatro respuestas diferentes para cada par P-H, por lo que es esencial contar con un mecanismo de unificación robusto. El enfoque para consolidar las respuestas finales del LLM utiliza dos mecanismos para la toma de decisiones finales, evaluando el mejor proceso de razonamiento según la información abstracta proporcionada:

- **Debate**: Votación por mayoría. Proporciona un enfoque sencillo para seleccionar la respuesta más coherente. Este enfoque tiene por nombre SKA_MV.
- **Razonamiento**: Algoritmo de árbol de decisión. Aprende las reglas sopesando las líneas de razonamiento correctas para la respuesta final. Este enfoque tiene por nombre SKA_DT.

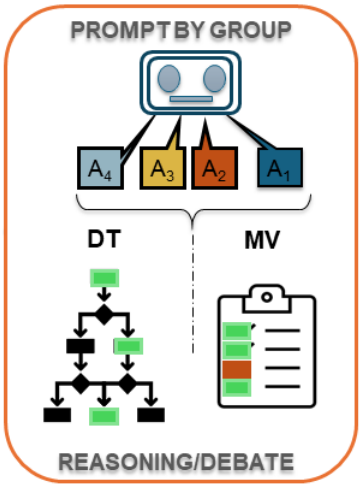


Figura 5.2: Mecanismos de inferencia. Razonamiento con un árbol de decisión y debate con voto mayoritario.

Nota. Adaptado de *Semantic knowledge abstraction: Consistent reasoning in large language models for natural language inference*, por (Torres-Moreno y Hermosillo-Valadez, 2026).

Las categorías abstractas son intuitivas para el ser humano y puedan proporcionar información de lo que ocurre en el par de acuerdo a la implicatura textual. Dejar a un lado la comprensión, de como un modelo trata el problema del RIT abona que se sigan generando modelos *black box* que no permiten desarrollar modelos explicables. Al modelar en términos interpretables lo que ocurre en la decisión del RIT, abona a entender como es que funcionan los procesos cognitivos implicados en esta tarea. Para continuar con el marco metodológico, se requiere sentar las bases y definiciones de a que se establece con el concepto de **Compatibilidad Semántica** e **Incompatibilidad Semántica**. En la siguiente sección se ahonda en la explicación. El código del marco metodológico (SKA) se encuentra disponible en github[1].

[1]https://github.com/labsemco/Semantic-Abstraction-Knowledge-SKAAnapplicationin-NLI-task

5.1. Compatibilidad semántica en el RIT

En la tarea del RIT se busca clasificar una clase objetivo, con base a la representación de lo que ocurre entre la premisa y la hipótesis. Para lograr esto, primero se requiere identificar las relaciones que existen entre $< P, H >$. Por lo que, de primera instancia se deben de identificar relaciones desde la premisa a la hipótesis de **Compatibilidad semántica**; generales o de equivalentes que describen Entailment, o **Incompatibilidad semántica**; específicas y contradictorias que describen Neutralidad y Contradiction, respectivamente. Esta compatibilidad semántica definen las relaciones permitidas entre la premisa y la hipótesis para validar la implicatura textual. A continuación se definen formalmente estos conceptos.

Definición 5.1.1: Compatibilidad Semántica

Decimos que c_2 tiene una relación de Compatibilidad Semántica con c_1 si existe una relación de c_1 hacia c_2 y c_2 esta en un nivel más **alto** en la jerarquía de la red conceptual (**regla de la generalización**) o c_2 está al mismo nivel con un significado **equivalente** (**regla de equivalencia**). Nosotros escribimos $c_1 \underset{\text{rel\_SCG}}{\uparrow} c_2$ para la regla de generalización y $c_1 \underset{\text{rel\_SCE}}{\equiv} c_2$ para la regla de equivalencia. Alternativamente, escribimos $c_1 \xrightarrow{\text{rel\_SC}} c_2$ o $(c_1, \text{rel\_SC}, c_2)$, donde rel_SC es cualquiera de las reglas.

Definición 5.1.2: Incompatibilidad Semántica

Decimos que c_2 tiene una relación de Incompatibilidad Semántica con c_1 si existe una relación de c_1 a c_2 ($c_1 \xrightarrow{\text{rel.}} c_2$) y c_2 está en un nivel más **bajo** en la jerarquía de la red conceptual (**regla de concretización**) o c_2 está en el mismo nivel con un significado **diferente** (**regla de la oposición**). Nosotros escribimos $c_1 \underset{\text{rel\_SIC}}{\downarrow} c_2$ para la regla de concretización y $c_1 \underset{\text{rel\_SID}}{\nsim} c_2$ para la regla de oposición. Alternativamente, escribimos $c_1 \xrightarrow{\text{rel\_SI}} c_2$ o $(c_1, \text{rel\_SI}, c_2)$, donde rel_SI es cualquiera de las reglas.

Si dos palabras son sinónimos es claro entender que tienen un significado equivalente, pero si hablamos de conceptos con distinto nivel semántico, en la jerarquía conceptual , se debe identificar cuál incluye el significado del otro. Por ejemplo, el significado de “man”se encuentra contenido en el concepto “person”. Ya que “person” es un concepto que generaliza al concepto “man”. Si esta generalización dada en la hipótesis se encuentra a partir de la premisa, entonces es una consecuencia lógica y se cumple la implicatura. A continuación se muestra un ejemplo:

- Premise: A **man** is running.
- Hypothesis: A **person** is running.
- Gold label: Entailment

Si se invierte la dirección en la que se da esta relación, no se podría afirmar que la hipótesis es una consecuencia lógica de la premisa, debido a que "person"no necesariamente es un "man". En este sentido hablamos de especificidad o concretes, cuando en la hipótesis ocurre algo más específico que en la premisa. Es decir, a partir de un concepto general para llegar a otro más específico faltaría información explícita (o implícita) en la premisa, por lo que no se cumple la implicatura. En este caso, existe una clase de neutralidad, donde no se puede llegar a una conclusión final a partir de la premisa. Veamos el siguiente ejemplo.

- Premise: A **person** is running.
- Hypothesis: A **man** is running.
- Etiqueta: Not entailment (Neutral)

Por último, las relaciones contradictorias. Es fácil identificar que las relaciones de antonimia promueven un significado distinto. Pero también existen formas de relaciones que marcan una diferencia en contenido semántico y significado: co-hiponimia. Veamos el siguiente ejemplo.

- Premise: A man is **running**.
- Hypothesis: A man is **walking**.
- Etiqueta: Not entailment (Contradiction)

A partir de este conocimiento, sabiendo que el significado se contrapone (o es diferente) entre la premisa y la hipótesis, se concluye que no se cumple la implicatura, más aún podemos decir que se contradicen.

La suposición de que las relaciones semánticas abstractas (por ejemplo, compatibilidad e incompatibilidad) se alinean con las relaciones de implicatura, contradicción o neutralidad se deriva de la observación teórica general de que existen conexiones sistemáticas entre las relaciones entre clases de palabras (como hiponimia, hiperonimia, sinonimia y oposición) y las relaciones proposicionales, en particular la implicación (Cruse, 2004, p. 33),(Hurford y cols., 2007),(Jeffries, 1998, p. 187). De hecho, la implicación entre proposiciones ha servido durante mucho tiempo como herramienta central para definir las relaciones de sentido en semántica (Lyons, 1977). Partiendo de estos vínculos teóricos establecidos entre las relaciones de sentido y las implicaciones proposicionales, nuestro razonamiento es el siguiente.

La compatibilidad semántica (que abarca tanto las relaciones generales como las equivalentes) debería corresponder a la implicación. Las relaciones clasificadas como generalidad (por ejemplo, hiperonimia, meronimia), así como las clasificadas como equivalentes (por ejemplo, sinonimia) establecen una conexión entre oraciones tal que la verdad de P (por ejemplo, hay un hombre) implica necesariamente la verdad de H (por ejemplo, hay una persona) (Cruse, 2004; Hurford y cols., 2007).

La incompatibilidad semántica (incluidas las relaciones de oposición, diferencia y concreción) debe corresponder a la contradicción. Las relaciones de oposición y diferencia —como los antónimos, los opuestos direccionales, los co-hipónimos y los co-meronimos— incluyen elementos léxicos que no pueden ser simultáneamente verdaderos de la misma entidad. Estas relaciones léxicas imponen una relación lógica contraria entre las oraciones: la verdad de una implica la falsedad de la otra (Cruse, 2004; Jeffries, 1998). En consecuencia, cuando P y H difieren en términos de conceptos opuestos o distintos, P no implica H; la verdad de P no permite inferir la verdad de H. Además, en el caso de las relaciones concretas, la verdad de un término más general (superior en la jerarquía léxica) no implica, por definición, la verdad de un término menos general subordinado en la misma escala (Cruse, 2004; Jeffries, 1998).

Estas abstracciones conceptuales son cruciales para el análisis semántico, especialmente a la hora de evaluar la implicatura textual. El marco metodológico categoriza las relaciones predefinidas de *ConceptNet* en **Generales**, **Equivalentes**, **Concretas** y **Opuestas**, de acuerdo a la definiciones 5.1.1 y 5.1.2. En total *ConceptNet5* cuenta con 34 relaciones existentes[2] de las cuales no todas cumplen con las definiciones de compatibilidad e incompatibilidad semántica. A continuación se muestran algunos ejemplos de relaciones categorizadas e introducimos un símbolo que representa a cada categoría (Figura 5.3) para futuras referencias.

Simbolo de Abstracción	Significado	Ejemplos de relaciones
	General	• is_a • manner_of • part_of
	Equivalente	• synonym • form_of
	Concreto	• used_for • related_to
	Opuesto	• antonym • distinct_from

Figura 5.3: Relaciones de ConceptNet que cumplen con las definiciones 5.1.1 y 5.1.2, preservando la direccionalidad en la que se dan ($P \rightarrow H$). Es posible el uso de relaciones en otras categorías, por ejemplo, *is_a*, la diferencias es que se da con la direccionalidad ($P \leftarrow H$), para este caso es una relación concreta.

[2]https://github.com/commonsense/conceptnet5/wiki/Relations

Por lo tanto, es posible identificar el tipo, la dirección y la interacción de las conexiones entre unidades léxicas y subfrases de P y H. En la jerarquía conceptual, los enlaces verticales ascendentes permiten establecer una implicación lógica. Estos pueden combinarse con enlaces horizontales que representan relaciones de equivalencia (sinónimos) pertenecientes a la compatibilidad semántica. Los enlaces horizontales de relaciones opuestas (antónimos y cohipónimos) y los enlaces verticales descendentes pertenecen a la incompatibilidad semántica, lo que indica una no implicación. Estas abstracciones conceptuales se resumen en la Figura 5.4.

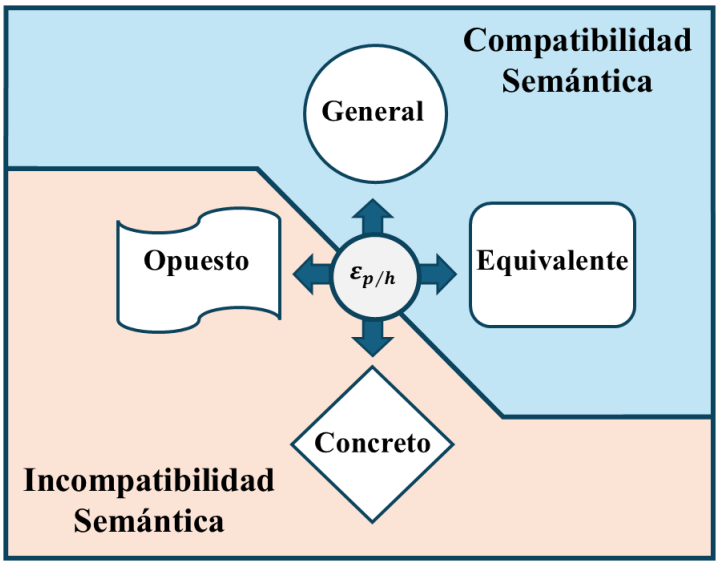


Figura 5.4: Marco de abstracción semántica según jerarquía de conceptos. Para algún elemento $\varepsilon_{p/h}$ de la premisa o hipótesis podemos identificar conceptos que cumplen con relaciones generales, equivalentes, concretas u opuestas.

Nota. Adaptado de *Semantic knowledge abstraction: Consistent reasoning in large language models for natural language inference*, por (Torres-Moreno y Hermosillo-Valadez, 2026).

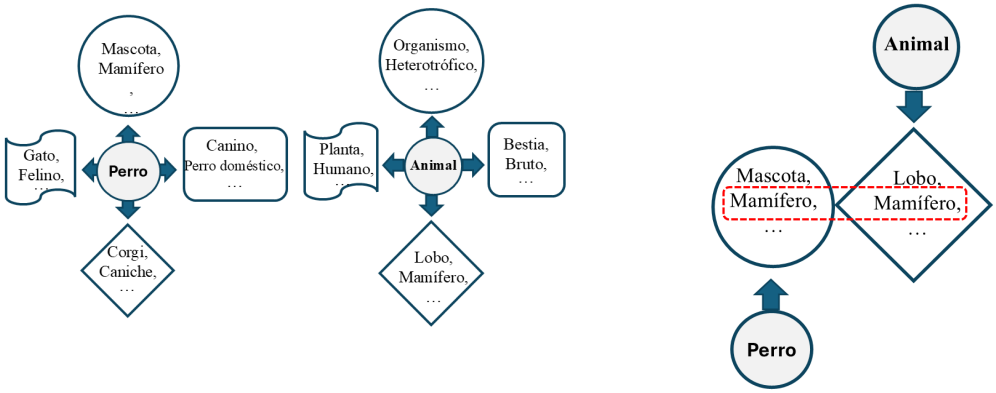
Es así que, se agrupan las relaciones de elementos del par $< P, H >$, de acuerdo al tipo y a la dirección en la que se dan $(P \rightarrow H)$, en grupos abstractos: G_1, G_2, G_3, G_4. Estos grupos abstractos integran tripletas de conexión de conceptos con su respectiva relación para usarlas en la tarea. Las relaciones relevantes de generalidad y equivalencia (G_1) permiten validar si un texto es consecuencia de otro. Es decir, si hay relaciones que contradicen (G_2) el significado de la oración principal, entonces no podemos decir que se sigue lógicamente. Pero también se debe prestar atención a las relaciones que no indican necesariamente una contradicción, por ejemplo, las relaciones concretas (G_3) y la falta de relaciones; puesto que elementos de P y H no se vinculan debido a la ausencia de relación o que no existe forma de vincular (G_4) en ConceptNet. Dado que se buscan relaciones se considera el tipo de relación que existe y preservan el significado o no entre el par analizado, cumpliendo la direccionalidad de la tarea.

A continuación se describen los grupos que cumplen con las definiciones establecidas.

- G_1: **Generalidad y equivalencia** - Compatibilidad semántica. Estas relaciones se dan entre elementos de los pares de P y H que se encuentran etiquetados como Entailment.
- G_2: **Opuestas** - Incompatibilidad semántica. Estas relaciones opuestas se dan entre elementos de los pares de P y H que se encuentran etiquetados como Contradiction.
- G_3: **Concretas** - Incompatibilidad semántica. Estas relaciones se dan entre elementos de los pares de P y H que se encuentran etiquetados como Neutral.
- G_4: Falta de relación entre elementos de P y H que no es posible identificar con ConceptNet.

Los grupos anteriores definen listas de nodos (relaciones) o subgrafos que conectan directamente entidades de P y H, basándose en las definiciones de las categorías abstractas. Con el fin de enriquecer los grupos con relaciones semánticas no directas entre elementos de P y H, se siguen las directrices de las reglas de compatibilidad semántica (SC) e incompatibilidad semántica (SI) utilizando la propiedad de transitividad. Es decir, se quiere saber si hay otros conceptos con relaciones SC de elementos de H que, a su vez, se conectan con elementos de P bajo este mismo principio.

Para ello, se desarrollan los subgrafos de cada elemento de P y de H siguiendo las reglas definidas (véase la Figura 5.4) para averiguar si existen conexiones transitivas que conservan la SC o la SI. Estos nuevos conceptos definen conjuntos de relaciones, o bolsas de conceptos bajo relaciones (generales, equivalentes, concretas y opuestas.), para cada uno de los elementos de P y H. Si hay una intersección entre estos conjuntos, significa que hay conceptos (intermedios) que vinculan elementos de P y H (Figura 5.5).



(a) (b)

Figura 5.5: Extensión de relaciones semánticas según las definiciones 5.1.1 y 5.1.2. a) Relaciones semánticas de cada elemento de P y H y b) Intersección de conjuntos de relaciones entre elementos de P y H preservando la jerarquía.

En la Figura 5.5a, se elige un elemento "perro" de P y "animal" de H y se encuentran todas las relaciones de generalidad, equivalencia, concretas y opuestas que tiene cada elemento a través de ConceptNet. Una vez que se tienen estas relaciones se procede a identificar posibles conexiones que no son directas entre los conjuntos de los elementos obtenidos. En la Figura 5.5b se muestra que el conjunto de relaciones generales de "perro" y el conjunto de relaciones concretas de "animal", comparten un concepto que los une "mamífero", y que además preserva la direccionalidad de la jerarquía. Es así que podemos decir que un "perro" es un "mamífero" y un "mamífero" es un "animal", por lo tanto un "perro" es un "animal" por la regla de la generalización en 5.1.1.

Sin embargo, no todas las intersecciones son válidas, ya que debe respetarse las reglas de transitividad para que se mantenga la SC. Así, de las 16 intersecciones posibles entre los 4 conjuntos (Generalidad, Equivalencia, Concretas y Opuestas), solo 9 son válidas[3] (ver Figura 5.6). De este modo, a partir de estas intersecciones surgen nuevas conexiones válidas con las que es posible enriquecer los grupos abstractos siguiendo su definición. A continuación se muestran las intersecciones válidas representadas por los símbolos de relaciones de Generalidad, Equivalencia, Concretas y Opuestas.

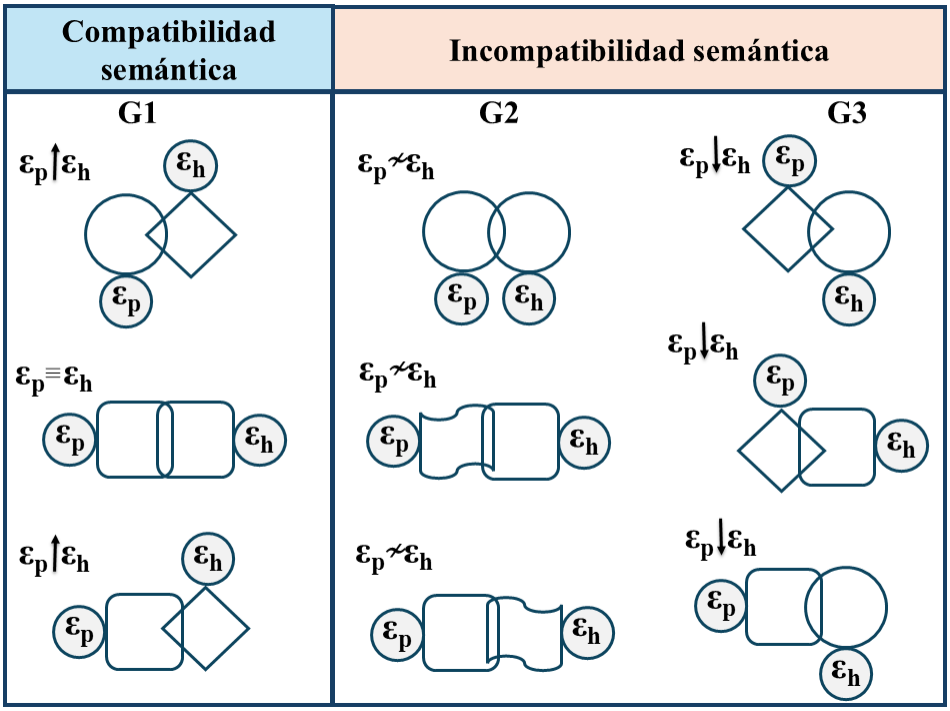


Figura 5.6: Ampliación de las relaciones. Para los elementos ε_p y ε_h se requiere encontrar relaciones que los vincule de acuerdo a las relaciones semánticas (ver Figura 5.5).

Nota. Adaptado de *Semantic knowledge abstraction: Consistent reasoning in large language models for natural language inference*, por (Torres-Moreno y Hermosillo-Valadez, 2026).

[3]Una intersección inválida es: opuestas de elementos de P con opuestas de elementos de H, ya que semánticamente no refiere a algo válido; es decir, no podemos asegurar que los elementos de P y H sean sinónimos por compartir antónimos.

A continuación se describe lo que representa la Figura 5.6. G_1 amplia las relaciones de generalidad (o equivalencia) desde el elemento ε_p utilizando las relaciones concretas o equivalencia del elemento ε_h (ver Figura 5.5). G_2 captura antonimia (sobre la intersección de relaciones de sinonimia y antonimia de elementos de P y de H) y los cohipónimos en la intersección de las relaciones generales de ambos (por ejemplo, "correr" y "caminar" comparten una relación con el concepto "mover", pero entre ellos es una relación de cohiponimia). G_3 amplia relaciones concretas a partir de las relaciones concretas de ε_p en las relaciones generales o de equivalencia de ε_h utilizando la regla de concretización en 5.1.2. Un ejemplo de este último, si el elemento "animal" se encuentra en P y el elemento "perro" en H, existe una relación concreta entre "animal" y "perro".

Estas intersecciones revelan conexiones más profundas que informan sobre la clasificación de implicatura, lo que nos permite identificar relaciones semánticas directas e indirectas que podrían no ser evidentes a simple vista en ConceptNet y que no están presentes en la red conceptual del LLM, como se muestra en la Figura 5.7. Esto permitirá al LLM mejorar su razonamiento al proporcionarle la categoría abstracta y la relación entre conceptos, y al darle una mayor flexibilidad en su red conceptual, lo cual es relevante para la tarea del RIT.

Para ilustrar lo que refiere la Figura 5.6 a continuación mostramos un ejemplo. Si no se encuentra una relación directa, la propuesta es intentar establecer un vínculo entre los elementos de P y H identificando posibles intersecciones entre sus subgrafos, con base a la compatibilidad semántica. La Figura 5.7 ilustra este propósito utilizando el ejemplo sobre "Stockholm".

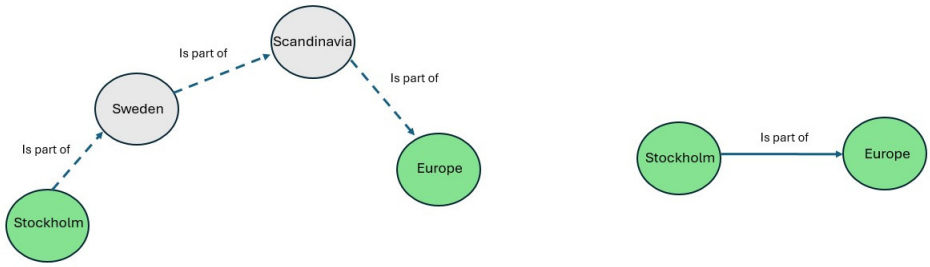


Figura 5.7: Flexibilización de la red conceptual mediante la construcción de nuevas conexiones entre conceptos. La regla de generalización permite llegar al nodo "Europa"desde "Estocolmo", manteniendo la compatibilidad semántica (con la relación "is_part_of"de ConceptNet que pertenece a la relación General) a lo largo del camino. La transitividad de la regla permite establecer el enlace directo entre estos dos conceptos.

Nota. Adaptado de *Semantic knowledge abstraction: Consistent reasoning in large language models for natural language inference*, por (Torres-Moreno y Hermosillo-Valadez, 2026).

A continuación se muestran ejemplos de intersección de caminos entre conceptos de cada uno de las intersecciones válidas que se proponen (Figura 5.8):

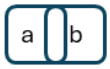
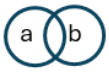
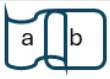
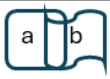
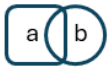
Intersecciones válidas	Premisa – Hipótesis	Camino de Intersección
	P: A dog licks the peanut butter jar. H: The dog is eating pinecones.	a:= butter, is_a, **food** b:= **food**, used_for, eat
	P: A man and a woman are sharing a dishwashing task in a kitchen at the sink. H: A man and woman wash dishes.	a:= sink, synonym, **lavatory** b:= wash, synonym, **lavatory**
	P:Woman climbing an artificial rock wall. H: A woman is training to go mountain climbing.	a:= **climbing**, synonym, climb b:= mountain, used_for, **climbing**
	P: A man is driving a red car. H: A white car is being driven by a man.	a:= red, is_a, **color** b:= white, is_a, **color**
	P: the cat sits inside the house H: A brindle-coated dog is growling at inside a park enclosure.	a:= **garden**, antonym, house b:= **garden**, synonym, park
	P: A boy running on the sidewalk in front of a blue building. H: A boy sits on the sidewalk.	a:= run, synonym, **move** b:= sit, antonym, **move**
	P: Three Asian boys race each other H: A few people are listening to music at a house party.	a:= **place**, manner_of, race b:= house, is_a, **place**
	P: A parking lot filled with bikes and trikes. H: There is a full parking lot of bikes and trikes.	a:= **complete**, manner_of, fill b:= full, synonym, **complete**
	P: People are walking. H: A woman and a man walk along a bridge overlooking a body of water.	a:= people, form_of, **person** b:= man, is_a, **person**

Figura 5.8: Ejemplos de conexiones semánticas válidas a partir de la ampliación de las relaciones de la Figura 5.6. Los términos en negrita indican la intersección entre las categorías abstractas de conceptos, lo que permite establecer conexiones semánticas válidas.

Nota. Adaptado de *Semantic knowledge abstraction: Consistent reasoning in large language models for natural language inference*, por (Torres-Moreno y Hermosillo-Valadez, 2026).

5.2. Mecanismo de representación

Para poder abordar este problema del RIT, para cada par $< P, H >$ se caracterizan las relaciones que ocurren entre P y H. Las características propuestas abordan cada una de las ambigüedades del lenguaje. Por lo que, nuestro mecanismo de representación sigue una lógica de identificación y reducción de elementos léxicos y composición del significado. Las entidades y atributos ayudan a resolver ambigüedades al proporcionar información adicional sobre el rol y significado específico de cada elemento. Crear entidades y atributos en lugar de depender únicamente del nivel léxico para construir significado es crucial para una comprensión más profunda y precisa del lenguaje.

Tradicionalmente, en los distintos modelos enfocados al RIT, se realiza un preprocesamiento para cada P y H. Primero se identifican las palabras y se obtiene su lemma (forma de la palabra

que representa todas sus formas flexionadas) y su Part Of Speech (Etiquetado gramatical del tipo de palabra). Este proceso se realiza con la librería spaCy sobre el análisis de dependencias de las frases. El PoS de cada palabra ayuda a crear entidades con sus atributos o modificadores de los sustantivos y los verbos. Los modificadores de los sustantivos pueden ser otros sustantivos, adjetivos, números, etc. Para el caso de los verbos por lo regular sus modificadores son otros verbos y adverbios. En la siguiente subsección entraremos en detalles.

Se ha categorizado relaciones que provee ConceptNet en grupos donde se identifica compatibilidad semántica e incompatibilidad semántica. De acuerdo a la tarea del RIT, para poder tomar la decisión de la clase final, se requiere contar con información relevante de cada entidad (anteriormente elementos) con sus respectivos atributos del par $< P, H >$. Con esta lógica se puede extender las definiciones para validar relaciones semánticas sobre entidades con sus atributos y agruparlas para incorporar este conocimiento en los LLMs.

Entidades y atributos

Como se ha indicado anteriormente la clasificación de cada relación se realiza según las categorías abstractas sobre las entidades de $< P, H >$. Se aplica el proceso de subgrafos de relaciones encontrado en ConceptNet para cada entidad y cada atributo. La primera categoría es Compatibilidad semántica: G_1 e Incompatibilidad semántica: G_2 y G_3. Sin embargo, dado que se trabaja con un recurso externo, es posible que no se puedan mapear todas las relaciones posibles. Por este motivo, se añadió un grupo que representa la falta de conocimiento G_4. A continuación se muestra una forma forma de identificar las entidades con sus respectivos atributos.

Para comprender el significado profundo del lenguaje humano es fundamental identificar cómo las entidades dentro de una frase se relacionan con sus modificadores, analizando la estructura gramatical de la frase y PoS de las palabras. Estos elementos son cruciales para desglosar la complejidad del texto y construir una representación más precisa del significado. Al analizar cada oración, se puede identificar las entidades presentes (sustantivos) y los modificadores que se relacionan con ellas. Las entidades principales son las que poseen características predominantes de significado. A continuación se describe este proceso:

- Identificación de entidades: Se analiza cada oración para determinar qué entidades están presentes (sustantivos).
- Modificadores: Estos elementos (adjetivos o adverbios) modifican el sustantivo, agregando información sobre su significado.
- Relaciones de significado: Cada frase puede ser analizada a nivel semántico para identificar el significado y la relación entre las entidades.

Un ejemplo se describe con la siguiente frase: Un perro blanco y café corre sobre el césped (Figura 5.9). La entidad principal es perro que contiene atributos relacionados con los conceptos "blanco" y "café" por lo que el proceso identifica atributos de cada uno. El verbo "correr" en su forma lematizada no tiene ningún modificador adverbial por lo que es un concepto único. El caso de "césped" ocurre algo similar. Intrínsecamente, podemos omitir los tipos de palabra artículos *un* y *el* ya que no altera el significado de la misma, pero si la parte gramatical. En cambio la palabra sobre indica información adicional por la cual no puede ser eliminada.

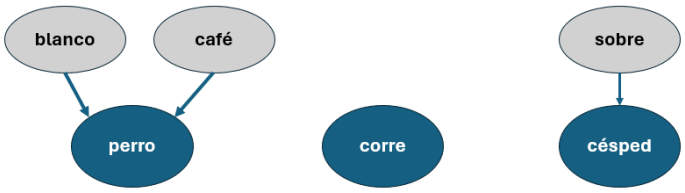


Figura 5.9: Representación de Entidades y atributos

Encontrar una relación entre entidades de $< P, H >$, es una parte fundamental de la arquitectura principal de caracterización de fenómenos lingüísticos para atender la tarea del RIT. Para esto se recurre a la siguiente notación de entidades y atributos.

Primero se establece una notación matemática para representar cada Premisa P y cada Hipótesis H, Para P y H, considerar los conjuntos $\mathcal{E}_p := \{(e_i^p[;\alpha_1^p,\ldots,\alpha_k^p])|i = 1,\ldots,p_1;\ k = 1,\ldots,p_2\}$, donde e_i^p es una entidad de P y $\alpha_1^p,\ldots,\alpha_k^p$ es una lista opcional de atributos correspondientes, y $\mathcal{E}_h := \{(e_i^h[;\alpha_1^h,\ldots,\alpha_k^h])|i = 1,\ldots,h_1;\ k = 0,1,\ldots,h_2\}$, donde e_i^h es una entidad de H y $\alpha_1^h,\ldots,\alpha_k^h$ es una lista opcional de atributos correspondientes. Con esta notación, cada entidad y atributo se puede tratar por separado, por lo que, para recuperar la frase original, escribiremos ε^* para representar una entidad con sus atributos correspondientes en lenguaje natural. Todas las entidades y atributos están lematizados. El siguiente ejemplo[4] muestra el funcionamiento de esta notación para P y H dados.

Example 5.2.1: Notación de entidades y atributos

$P:$	An old man in a long-sleeves white shirt is walking to work in a big city.
$\mathcal{E}_p =$	$\{(man; old), (shirt; white, long\text{-}sleeve), (walk), (work), (city; big)\}$
$\varepsilon_2 =$	$(shirt; white, long\text{-}sleeve); \varepsilon_2^* = long\text{-}sleeve\ white\ shirt$
$H:$	The man is wearing shorts and a t-shirt as he jogs.
$\mathcal{E}_h =$	$\{(man), (wear), (short), (t\text{-}shirt), (jog)\}$

[4]Ejemplo tomado del corpus SNLI

Para poder identificar qué tipo de relaciones vinculan entidades se implementa el marco de compatibilidad e incompatibilidad semántica. Debido a que la relaciones se encuentran sobre palabras, se requiere una forma para extenderlas a la conceptualización de entidades. A continuación se establecen las definiciones de los grupos que creamos y se construyen basándose en las definiciones de las categorías abstractas.

Definición 5.2.1: Grupo 1: Relaciones Generales y equivalentes

Dados dos elementos $\varepsilon_p \in \mathcal{E}_p$ y $\varepsilon_h \in \mathcal{E}_h$. Decimos que ε_h tiene una relación de **Compatibilidad Semántica** con ε_p, sí $e_p \xrightarrow{\text{rel\_SC}} e_h$ y $\alpha_p \xrightarrow{\text{rel\_SC}} \alpha_h$. En este caso, decimos que la relación pertenece al **Grupo 1** (G_1). Entonces nosotros escribimos $(\varepsilon_p \xrightarrow{\text{rel\_SC}} \varepsilon_h) \in G_1$ o alternativamente $(\varepsilon_p, rel\_SC, \varepsilon_h) \in G_1$.

Definición 5.2.2: Grupo 2: Relaciones Opuestas y diferentes

Dados dos elementos $\varepsilon_p \in \mathcal{E}_p$ y $\varepsilon_h \in \mathcal{E}_h$. Decimos que ε_h tiene una relación de **Incompatibilidad Semántica** con ε_p, sí $e_p \xrightarrow{\text{rel\_SID}} e_h$ o $\alpha_p \xrightarrow{\text{rel\_SID}} \alpha_h$ En este caso, decimos que la relación pertenece al **Grupo 2** (G_2). Entonces nosotros escribimos $(\varepsilon_p \xrightarrow{\text{rel\_SID}} \varepsilon_h) \in G_2$ o alternativamente $(\varepsilon_p, rel\_SID, \varepsilon_h) \in G_2$.

Definición 5.2.3: Grupo 3: Relaciones concretas

Dados dos elementos $\varepsilon_p \in \mathcal{E}_p$ y $\varepsilon_h \in \mathcal{E}_h$, Decimos que ε_h tiene una relación de **Incompatibilidad Semántica** con ε_p, sí $e_p \xrightarrow{\text{rel\_SIC}} e_h$ o $\alpha_p \xrightarrow{\text{rel\_SIC}} \alpha_h$ En este caso, decimos que la relación pertenece al **Grupo 3** (G_3). Entonces nosotros escribimos $(\varepsilon_p \xrightarrow{\text{rel\_SIC}} \varepsilon_h) \in G_3$ o alternativamente $(\varepsilon_p, rel\_SIC, \varepsilon_h) \in G_3$.

El término "válido" se refiere a una relación en el grafo de conocimiento que cumple con las definiciones de compatibilidad semántica 5.1.1 y 5.1.2. Las definiciones de los grupos no solo validan la entidad sino también sus atributos deben de considerar las definiciones de compatibilidad semántica. Es fundamental precisar que aunque ConceptNet es una base de conocimiento grande, no contiene todas las relaciones posibles que nosotros podemos identificar entre P y H. El último grupo G_4 reunirá todas las entidades de H que no tengan ninguna relación con ninguna de las entidades de P. En ese caso, se escribe

$$(, \text{UNK}, \varepsilon_h) \in G_4.$$

Una vez que que se obtienen las entidades se realiza el proceso de identificación de relaciones semánticas utilizando ConceptNet. A continuación se describen las etapas de la búsqueda de relaciones semánticas sobre las entidades y atributos sobre el ejemplo anterior 5.2.1.

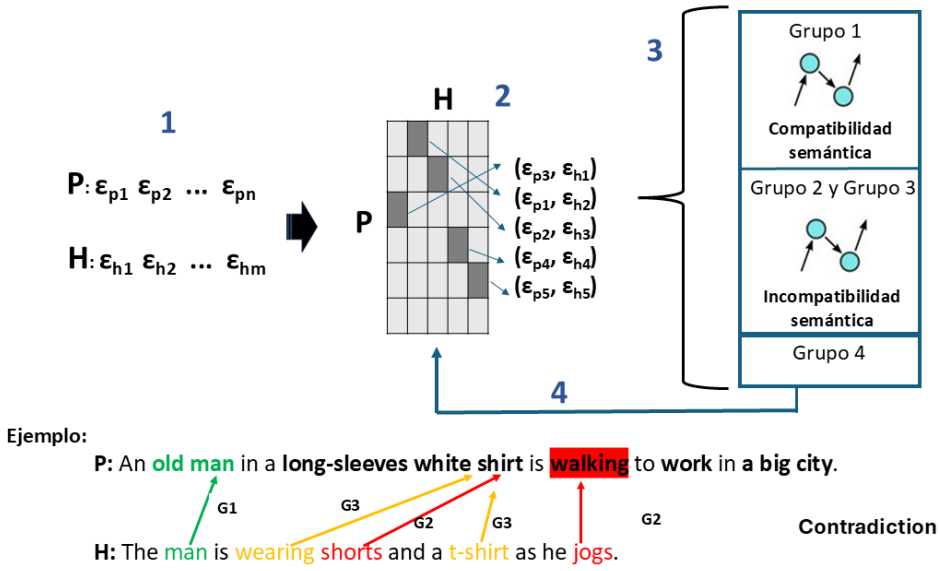


Figura 5.10: Proceso de alineación de entidades e identificación de relaciones con ConceptNet.

Las diferentes etapas para obtener todas las relaciones posibles y válidas se desarrollan a continuación:

- **Etapa uno**: Entidades y atributos. Se realiza el proceso de creación de entidades y atributos de P y H con el análisis de dependencias: Sustantivos con adjetivos y verbos con adverbios. En este proceso se identifican las palabras, así como su lemma y su POS. En caso de identificar algún signo de puntuación o palabra funcional (*stopwords*) se eliminan. Esto es debido a que se trabaja con el significado de las palabras y se puede prescindir de signos de puntuación y palabras funcionales que no agregan información semántica para la comprensión del significado.

- **Etapa dos**: Alineación. Para cada entidad resultante de H se asocia con entidades de P. Para realizar esto se utilizan sus WE normalizados tomados de ConceptNet. Es decir, se obtiene su distancia coseno como parámetro para explorar cuales entidades de H podrían relacionarse con las de P. Priorizamos las entidades de H debido a que queremos identificare la contención de H en P.

- **Etapa tres**: Identificación de relaciones semánticas: A través de ConceptNet y con base a las definiciones se identifica que relación une a los pares alineados sobre las categorías de compatibilidad e incompatibilidad semántica.

- **Etapa cuatro**: Búsqueda para todas las entidades de H. Repetimos la etapa tres hasta ya no

identificar más relaciones en ConceptNet. Este proceso se realiza para los $top - k$[5] entidades con mayor puntaje, que son las entidades dominantes en la matriz de alineamiento sobre las entidades de P. El valor k en este proceso es 3. En este caso, si no se encuentran relaciones, las entidades faltantes de H se agrupan en G_4.

En la Figura 5.10 se muestra un ejemplo de la implementación de la búsqueda de relaciones semánticas sobre las entidades del par $< P, H >$ y como a través de las relaciones de conceptNet se categorizan en los grupos descritos. El resultado de todo este proceso se muestra reutilizando el ejemplo 5.2.1, las siguientes listas constituyen los grupos abstractos de la siguiente manera, donde las relaciones se extraen de ConceptNet:

Example 5.2.2: Grupos de relaciones del ejemplo 5.2.1

$$
\begin{aligned}
G_1 &:= [(\textit{old man}, \textbf{is a}, \textit{man})] \\
G_2 &:= [(\textit{walk}, \textbf{distinct from}, \textit{jog}), \\
&\quad\ (\textit{long-sleeve white shirt}, \textbf{distinct from}, \textit{short})] \\
G_3 &:= [(\textit{long-sleeve white shirt}, \textbf{related to}, \textit{wear}), (\textit{long-sleeve white shirt}, \textbf{is a}, \textit{t-shirt})] \\
G_4 &:= \emptyset
\end{aligned}
$$

En este resultado da cuenta que todas las entidades de H se vincularon con alguna de las entidades de P, por esta razón G_4 se encuentra vacía. En lo que sigue se hará un abuso del lenguaje y al referir a las entidades con atributos simplemente como entidades.

La propuesta presenta algunas similitudes y diferencias claras con la lógica natural (MacCartney y Manning, 2007), que busca modelar formalmente la implicatura (relación de contención) entre P y H ($P \sqsubset H$) (MacCartney y Manning, 2009). En este sentido, la relación $P \sqsubset H$ se distingue de la relación $P \rightarrow H$, que es el tema del presente estudio. Dentro de la propuesta de (MacCartney y Manning, 2009), analizamos las transformaciones de edición (sustitución, eliminación e inserción) que permiten la expansión de una palabra (por ejemplo, vino se expande a bebida) -monotonía ascendente— manteniendo la relación de implicación $P \sqsubset H$, o su contracción (por ejemplo, comida se contrae a cena)—monotonía descendente—, alternando la relación hacia $P \sqsupset H$. La composición de estas transformaciones alterna el tipo de relación entre P y H, lo que permite deducir si finalmente se obtiene $P \sqsubset H$.

Para ilustrar este punto, se analiza el ejemplo (tomado de (MacCartney y Manning, 2009)) P: "Nadie puede entrar sin pantalones" y analicemos su relación de implicatura con H: "Nadie puede entrar sin ropa". Aquí, si se sustituye pantalones por ropa, la relación de contención "pantalones" $\sqsubset$

[5]Este valor permite reducir la búsqueda de relaciones de los elementos de P y H, sin operar sobre todas las combinaciones posibles.

"ropa" mantiene intacta la relación de implicatura. Al analizar la frase "sin pantalones" frente a "sin ropa", esta relación se invierte, de modo que "sin pantalones" $\sqsupset$ "sin ropa". Siguiendo este esquema de composición, se concluye que "Nadie puede entrar sin pantalones" $\sqsubset$ "Nadie puede entrar sin ropa" porque "Nadie" tiene monotonía DESCENDENTE (como la definen los autores), invirtiendo de nuevo la relación de implicatura (contención). Esta teoría ha continuado su trayectoria con el objetivo de construir árboles de razonamiento consistentes utilizando técnicas modernas (Shi y cols., 2025).

En la presente propuesta, las relaciones de generalidad y concretas funcionan de manera diferente, ya que no se aplican a palabras aisladas, sino al emparejamiento entre entidades P y H. Tomando el ejemplo anterior de P y H, el enfoque tomará la entidad "pantalones" junto con su atributo "sin" y producirá ("sin pantalones", es un tipo, "sin ropa"), una relación de generalidad entre las entidades de P y H. A diferencia de la intención de construir rutas de razonamiento formal, la propuesta busca generar información que contribuya a la inferencia de implicación lógica, neutralidad o contradicción, en ese orden de prioridad en el análisis de entidades con el fin de construir grupos.

5.3. Mecanismo de Inferencia

La presente propuesta radica en evaluar a los LLMs con respecto a la tarea del RIT e identificar que grupo le aporta mayor información y lo guía a través de un razonamiento plausible. Para llevar a cabo esto, se realizaron una serie de experimentos sobre los LLMs elegidos sobre un muestreo de los ejemplos de los distintos corpus del RIT.

Dado que estos modelos suelen carecer de información semántica, relacional y de sentido común, la propuesta se centra en integrar datos externos clave de forma abstracta que les guíen hacia decisiones más precisas y fundamentadas. Para lograrlo, se aplican técnicas de *prompting* utilizando los grupos abstractos de relaciones que son disjuntos entre sí, como en (Dai y cols., 2025).

El marco metodológico permite formular la misma pregunta de la tarea proporcionando información diferente, cada una basada en la información de los grupos categorizados ($G_1 - G_4$), lo que garantiza un análisis multifacético independiente. Este enfoque refuerza la fiabilidad de los resultados al incorporar estratégicamente el contexto externo y tener en cuenta diferentes líneas de razonamiento basadas en la información abstracta proporcionada (Figura 5.11). Como referencia, también se solicitó una respuesta sin información adicional para establecer una línea del modelo base que facilite la evaluación comparativa de la mejora lograda.

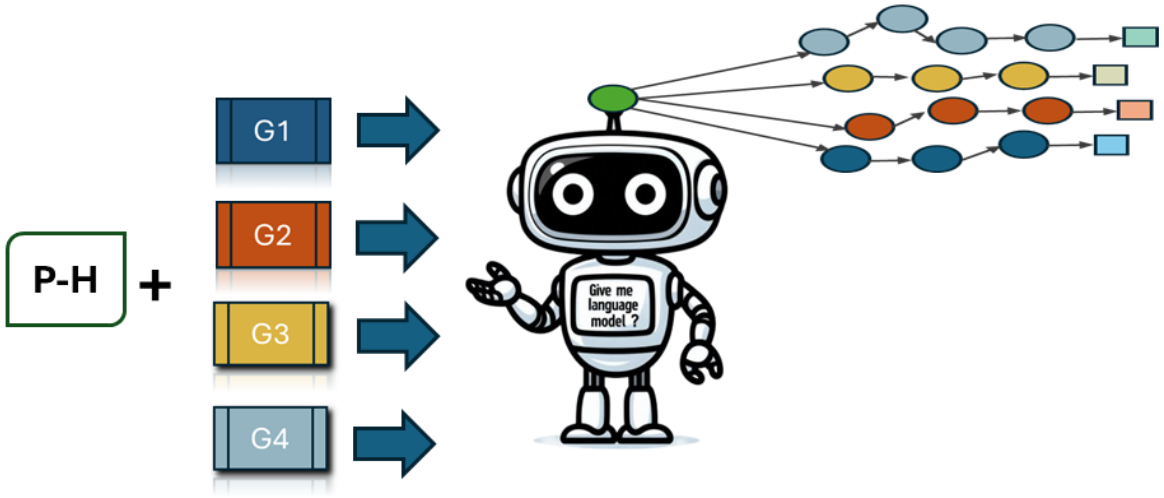


Figura 5.11: Proceso de *prompting* en LLMs: se plantea al LLM una consulta estructurada con base a los pares $< P - H >$ y las relaciones establecidas entre los grupos abstractos, con el objetivo de obtener cuatro respuestas, una correspondiente a cada grupo.

Para el *prompting*[6], se incluye la definición (Figura 5.13) de los grupos abstractos $G_1 - G_4$. Los LLMs podrán tomar la información y utilizarla en su beneficio para tomar una mejor decisión. El prompt utilizado se muestra en la Figura 5.12.

```
You are an expert in Recognizing Textual Entailment over pairs of Premise and Hypothesis. Based on the background
information provide below, classify the relationship between the given Premise and Hypothesis as one of the
following: Entailment, Neutral or Contradiction. Respond only using the template: { ''Answer'': }. Do not modify
the template.
Premise and hypothesis to classify:
Premise: {P_i}
Hypothesis: {H_i}
Background Information: {group_definition_j}
Word relations group: {group_relations_j}
```

Figura 5.12: Prompting con SKA para respuestas de los LLM.

Donde la variable i recorre el número de ejemplos de los corpus. La variable j recorre los grupos $(G_1 - G_4)$ que se proponen. $group\_definition_j$ depende del grupo abstracto de relaciones con el que se está consultando el modelo. La Figura 5.13 muestra las definiciones. $group\_relations_j$ es la lista de relaciones semánticas que pertenecen al grupo abstracto según el marco metodológico de la abstracción semántica. Un ejemplo de uso se da en las siguientes secciones (Figura 6.6).

[6]Los prompts están redactados en inglés, dado que la consulta al LLM se realizó en ese idioma.

- G1: Triplets of Generality or Equivalence. These relations usually correspond to Entailment. triplets: (p_i,rel,h_j) where p_i is in Premise, h_j is in Hypothesis and rel is the relation between p_i and h_j. e.g., (dog,is_a,animal) dog is in the premise, animal is in the hypothesis and the relationship is general from dog to animal.
- G2: Triplets of Contradictory. These often indicate Contradiction. triplets: (p_i,rel,h_j) where p_i is in Premise, h_j is in Hypothesis and rel is the relation between p_i and h_j. e.g., (dog,distinct_from,cat) dog is in the premise, cat is in the hypothesis and relation is distinct_from between dog - cat.
- G3: Triplets of Concretization, Specificity, or Contextuality. These typically correspond to Neutral. triplets: (h_j,rel,p_i) where p_i is in Premise, h_j is in Hypothesis and rel is the relation between p_i and h_j. e.g., (dog, is_a, animal): animal is in the premise, dog is in the hypothesis, and the relationship is specific from animal to dog.
- G4: Relations not identified or categorized.

Figura 5.13: Definiciones de grupo abstracto, depende del grupo abstracto de relaciones a usar.

El LLM genera cuatro respuestas para cada par $< P, H >$ sobre la información del grupo abstracto proporcionada, por lo que es esencial un mecanismo de unificación robusto. El enfoque para consolidar las respuestas finales de los LLMs utiliza diferentes técnicas para la toma de decisiones finales, evaluando el mejor proceso de razonamiento según la información abstracta proporcionada. Para lograrlo, utilizamos dos estrategias clave, como se muestra en la Figura 5.2:

- **SKA_MV**
- **SKA_DT**

Para SKA_DT, se requiere un muestreo adicional de pares $< P, H >$ solo para su entrenamiento, lo que garantiza que el modelo se generalice sin sobreajustarse. La intuición detrás de esta propuesta es que cada grupo $(G_1 - G_4)$ proporciona información abstracta única, sobre el par $< P, H >$ guiándo al LLM sobre diferentes líneas de razonamiento y permitiendo identificar que grupo abstracto aporta más información en la toma de decisiones. Al combinar estas técnicas, no solo capturamos esta variación, sino que también se maximiza la influencia de las respuestas correctas, lo que da como resultado una decisión final más fiable y fundamentada.

Con el fin de medir y contrastar el desempeño de la propuesta, se utiliza como referencia el rendimiento del modelo base, obtenido mediante el siguiente prompt:

```
You are an expert in Recognizing Textual Entailment over pairs of Premise and Hypothesis. Classify the
relationship between the given Premise and Hypothesis as one of the following: Entailment, Neutral or
Contradiction. Respond only using the template: { ‘‘Answer’’: }. Do not modify the template.
Premise and hypothesis to classify:
Premise: {P_i}
Hypothesis: {H_i}
```

Figura 5.14: Prompt del base para los LLM. La variable i recorre el número de ejemplos de los corpus.

En línea con (Hong y cols., 2024), también diseñamos un prompt basado en la metodología AoT para llevar a cabo nuestro análisis comparativo.

```
You are an expert in Recognizing Textual Entailment over pairs of Premise and Hypothesis. Classify the
relationship between the given Premise and Hypothesis as one of the following: Entailment, Neutral or
Contradiction.
Premise and hypothesis to classify:
Premise: {P_i}
Hypothesis: {H_i}

Let’s think step by step

Step 1: Identify all the relationships between terms from the premise to the hypothesis. This process is performed
for each of the terms in the hypothesis.
Use the next format for relationships: (p,rel,h) where p is a term from the premise, h is a term from the
hypothesis, and rel is the relation linking these terms. If any terms of the hypothesis with an unknown
relationship with terms in the premise, identify them as (,unknown,h) where h is in the hypothesis. List the
relationships found.

Step 2: Align all the relationships found with the RIT classes: Entailment, Neutral, Contradiction; classifying
them into groups G1, G2, G3 and G4 according to the following:
G1: will contain the list of relationships that align with the entailment label.
G2: will contain the list of relationships that align with the contradiction label.
G3: will contain the list of relationships that align with the neutrality label.
G4: will contain the list of terms of the hypothesis with an unknown relationship with the premise.

Step 3: Analyze each group of relationships and decide on the correct label for the premise and hypothesis
presented.

Respond only using the template: { ‘‘G1’’:[], ‘‘G2’’:[], ‘‘G3’’:[], ‘‘G4’’:[], ‘‘Answer’’: }. Do not modify
the template.
```

Figura 5.15: AoT prompt. La variable i recorre el número de ejemplos en los corpus. También se solicitan los grupos que obtiene según el prompt.

Capítulo 6

Experimentación y Resultados

En esta sección se detallan los experimentos para identificar el rendimiento de la metodología establecida anteriormente. Con el fin de investigar el impacto del mecanismo de abstracción propuesto y el grado en que el LLM lo incorpora a su conocimiento para apoyar su razonamiento, analizamos el impacto en el rendimiento por grupo y con los mecanismos propuestos. A continuación se describe la configuración de los experimentos realizados.

Los experimentos se realizaron sobre 6 LLMs: `llama 3.1` y `llama 3.2` de *Meta*, `phi3:medium` y `phi3` de *Microsoft* y `gemma2` y `gemm2:2b` de *Google*. Estos modelos se usaron directamente sin ninguna configuración adicional. Los modelos elegidos han sido probados anteriormente en el *benchmarck* *MMLU*[1], donde se aplican diferentes tareas de comprensión y cuenta con un total de 16 mil preguntas de diversas áreas. La siguiente Figura 6.1 muestra un comparativo del rendimiento de los LLMs seleccionados y el tamaño de sus parámetros. Para fines ilustrativos se agregó el rendimiento obtenido de `GPT4-o`, del cuál no se tiene el número de parámetros.

La elección de estos modelos toma en cuenta el rendimiento del *MMLU* y seleccionando una versión pequeña y una con más parámetros de la misma familia. En la gráfica se puede observar que `phi3:medium` es un modelo con el mayor número de parámetros, seguido de `gemma2` y `llama3.1`. El modelo con menos parámetros es `gemma2:2b`, seguido de `llama3.2` y por último `phi3`. Aunque el rendimiento de los modelos mostrado en la Figura 6.1 es consecuente con el número de parámetros de los modelos, notamos que `phi3` (que es un modelo pequeño) tiene un rendimiento equiparable con los modelos más grandes. El modelo que tiene el mayor rendimiento es `GPT4-o`.

[1]Massive Multitask Language Understanding (Comprensión Masiva de Lenguaje en Múltiples Tareas), es uno de corpus que se utilizan para medir el conocimiento general de los LLMs.

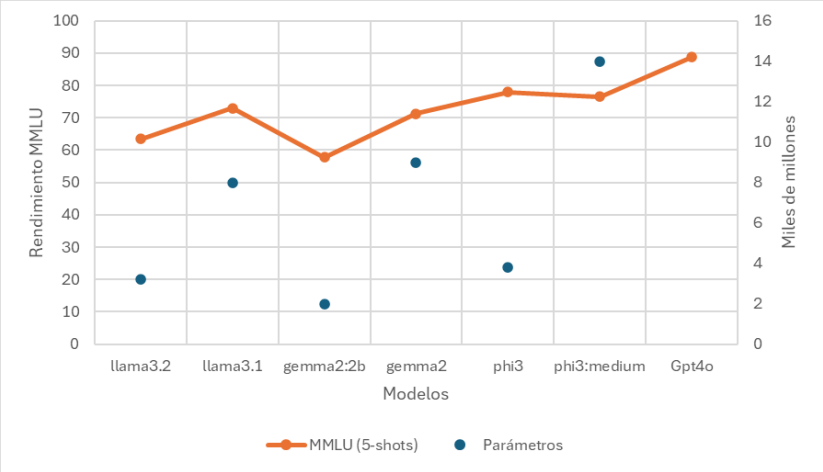


Figura 6.1: Accuracy (y parámetros) de LLMs en MMLU

Para poder usar estos modelos, se configuró un entorno con dos GPUs; una *NVIDIA RTX4090* y una *NVIDIA RTX3070* en un servidor con un procesador *Intel core-i9* y 128 GB de RAM. Se utilizó el aplicativo *Ollama*[2] que permite descargar y ejecutar de manera local los modelos ya entrenados.

Los corpus que se consideraron para evaluar a los modelos con la propuesta SKA son: SNLI y SICK con tres clases; RTEGLUE, Scitail y el diagnóstico del *benchmark GLUE* con dos clases. Debido al alto costo de recursos computacionales y el tiempo de procesamiento, se generan muestreos[3] aleatorios de los corpus, equilibrando el número de ejemplos de las clases y tener suficiente información para generar pruebas estadísticas. Únicamente para el corpus de diagnóstico se tomaron todos los pares de ejemplos. A continuación se muestra la Tabla 6.1 con los corpus usados, con su número de clases y la selección de muestreos[4]:

Corpus	Clases	Muestreos	Pares de ejemplos por muestreo
SNLI	3	13	600; 200 por class
SICK	3	13	300; 100 por class
RTE	2	13	200; 100 por class
Scitail	2	13	400; 200 por class
Diagnóstico	2	1	1,104; 460 (E) y 644 (NE)

Cuadro 6.1: Corpus usados y muestreos de ejemplos de los corpus con clases equilibradas para evaluar el SKA. Las 3 clases de los corpus son: Entailment, Neutral y Contradiction. Para los corpus con dos clases: Entailment y Not entailment.

[2]https://ollama.com/

[3]Parte representativa de un conjunto de datos para realizar análisis, estimaciones o pruebas, con el fin de obtener conclusiones válidas sobre el conjunto total.

[4]Para los experimentos se establecieron semillas aleatorias para garantizar la reproducibilidad.

Para los experimentos, usamos el *prompt* (Figura 5.12) que cumple con lo siguiente: 1) Permite tener una respuesta adecuada por parte de los LLMs. 2) Permite proporcionar información de forma clara y útil para dar una respuesta. 3) Responde en un mismo formato para validar la respuesta (estructura en formato JSON). 4) Permite validar el impacto de la información proporcionada. A pesar de este alineamiento aún se tuvo dificultades en respuestas inadecuadas o sin formato, las cuales se intentaron procesar y si ocurría el mismo error se excluyeron.

6.1. Resultados de LLMs

Los grupos de abstracción propuestos se diseñaron para capturar las relaciones de compatibilidad e incompatibilidad semánticas con el fin de inducir una línea de razonamiento en los LLMs. La hipótesis es que cada grupo de abstracción inducirá respuestas con cierta tendencia; por ejemplo, una tendencia hacia la implicación o hacia la contradicción. Por lo tanto, la primera pregunta es ¿hasta qué punto estos grupos de abstracción semántica influyen en el proceso de razonamiento de cada modelo?. Para este análisis, se solicitaron respuestas a los LLM utilizando el *prompt* de la Figura 5.12. Una vez obtenidas las respuestas, se calculó el rendimiento promedio de los LLM para cada grupo y por corpus.

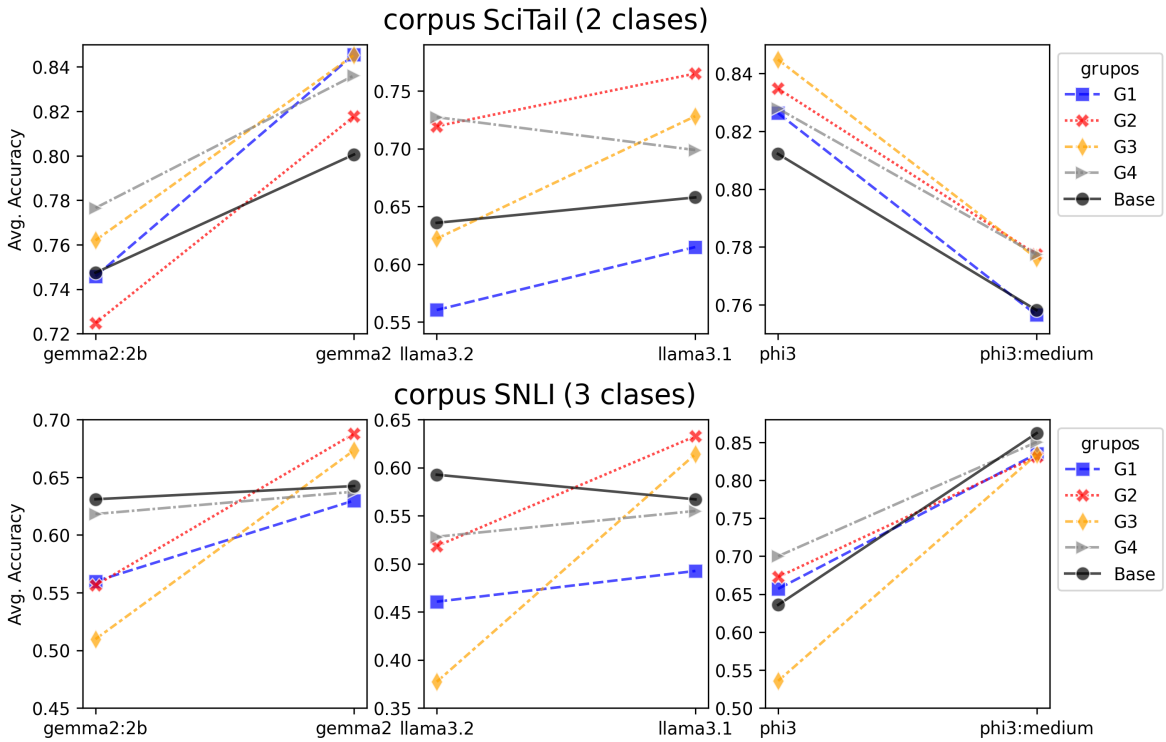


Figura 6.2: Accuracy promedio de la influencia de los grupos en los LLMs. *Base* hace referencia a las respuestas del modelo sin SKA.

La Figura 6.2 resume el promedio de rendimiento de cada LLM en diferentes corpus bajo la influencia de cada grupo de abstracción. La fila superior muestra el promedio de rendimiento en el corpus Scitail (dos clases), mientras que la fila inferior muestra su rendimiento en el corpus SNLI (tres clases). En el caso del corpus de dos clases, se observan generalmente mayores precisiones bajo la influencia de los grupos de abstracción, dependiendo del tamaño de cada modelo, excepto en los grupos G_2 y G_1 para los modelos `gemma2:2b` y `Llama`, respectivamente. La situación cambia para el corpus de tres clases. En general, el rendimiento del modelo disminuye en comparación con el modelo base, excepto en algunos grupos y algunos modelos. Pero se muestra que al menos un grupo obtiene mejores resultados. Por lo tanto, la segunda pregunta es ¿cómo llegar a un consenso sobre las respuestas que suscita cada grupo de abstracción?.

Para unificar las respuestas, se experimentó con las estrategias propuestas en la sección 5.3. Esto permite lidiar con la variabilidad de las respuestas de los LLMs e identificar la mejor estrategia. En el caso de SKA_DT, es necesario elegir 3 muestreos de entrenamiento distintos por corpus. Para la votación por mayoría SKA_MV, solo se toma la predicción más frecuente. En caso de empate, la decisión se tomó al azar entre la clase más votada. Se realizó una validación cruzada[5] (kfold=5); 10 muestreos para la prueba y 3 para el entrenamiento de la propuesta SKA_DT.

		Corpus				
Modelos	Característica	Scitail	RTEGLUE	SICK	SNLI	DIAG
gemma2	root	G_1,G_3	G_1	G_1,G_4	G_4	G_4,G_1
	Avg. depth	5	5	7	6	5.2
gemma2:2b	root	G_4	G_1,G_4	G_1	G_2	G_4,G_1
	Avg. depth	5	5	7	7	5.2
llama3.1	root	G_2,G_3	G_3,G_2	G_1	G_1	G_4
	Avg. depth	4	4	7	7	4
llama3.2	root	G_2,G_4	G_4	G_3	G_4,G_2	G_1
	Avg. depth	4	4	7	7	4.2
phi3:medium	root	G_3,G_2	G_4,G_2	G_1	G_1	G_4
	Avg. depth	5	5.2	7.2	8	4
phi3	root	G_3,G_4	G_4	G_1,G_4	G_2	G_4,G_3
	Avg. depth	4	4	6.2	6	4

Cuadro 6.2: Nodo raíz y profundidad media del árbol de decisión para el esquema SKA_DT según el criterio de Gini, tras la validación cruzada.

[5]Se utiliza para evaluar cómo los resultados de un análisis predictivo se generalizarán a un conjunto de datos independiente y no visto.

Para replicar los experimentos, se genera una semilla aleatoria. La Tabla 6.2 muestra las características del árbol de decisión después del entrenamiento de validación cruzada. Al analizar la Tabla, se nota que los corpus de 3 clases (SNLI y SICK) tienen una profundidad promedio del árbol notablemente mayor que los de 2 clases (RTEGLUE y SciTail). Además, en los corpus SNLI y SICK la raíz suele ser G_1 para la mayoría de modelos. Esto indica que para distinguir entre las clases Entailment, Neutral y Contradiction se necesita un proceso de decisión más elaborado y jerárquico, que parte típicamente de la cabeza de atención G_1, sugiriendo que esta esa información captura patrones fundamentales para la ambigüedad semántica.

El criterio de *Gini* y la validación cruzada generan árboles eficientes y generalizables, no estructuras triviales ni complejas. La profundidad media de (4-5) para tareas binarias sugiere que con solo 4-5 preguntas binarias, SKA_DT puede clasificar con alta precisión, lo que apunta a un mecanismo notablemente eficiente e interpretable. En la siguiente gráfica se muestra un comparativo sobre los mecanismos de inferencia que se evaluaron (Figura 6.3).

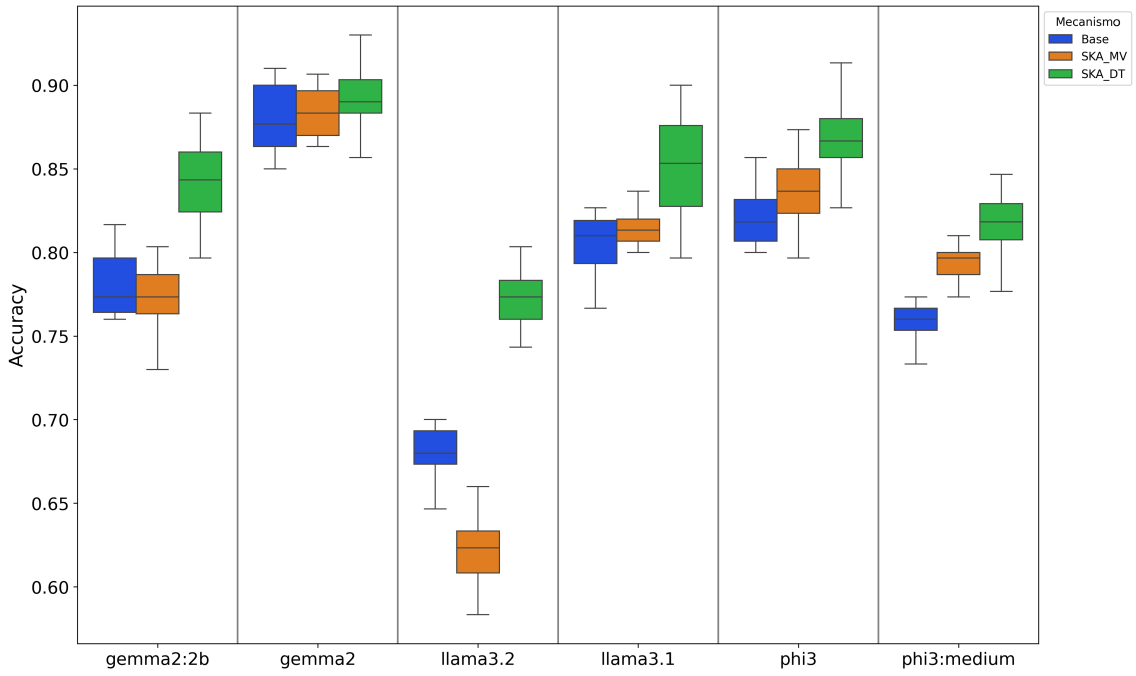


Figura 6.3: Accuracy de SKA_MV y SKA_DT en el corpus SICK

Los algoritmos para unificar en una respuesta se basan en las predicciones de clases recuperadas de las solicitudes a los LLMs. El algoritmo de SKA_DT logra obtener el rendimiento equilibrado sobre todos los LLMs en comparación con el SKA_MV. En este sentido, por su simpleza de entender e interpretar la toma de las decisiones finales, optamos por usar este algoritmo SKA_DT para la unificación de las respuestas en una sola etiqueta. Este método no solo unifica las respuestas de manera eficiente, sino que también pondera diferentes razonamientos para refinar continuamente el proceso, lo

que destaca la adaptabilidad y la solidez de la propuesta.

El rendimiento del mecanismo SKA_DT se muestra para los corpus de tres clases y dos clases en las Tablas 6.3 y 6.4, respectivamente. Además, se han incluido los resultados de nuestra implementación de AoT y calculado pruebas estadísticas con respecto a la línea de base.

SNLI				
Modelos	**Base**	**AoT**	**SKA_MV**	**SKA_DT**
gemma2	64.2 ± 1.5	64.9♦ ± 1.6	65.6♠ ± 1.8	71.1♠ ± 1.1
gemma2:2b	63.1 ± 1.8	57.5 ± 1.4	56.5 ± 2.12	**66.5♠ ± 1.7**
llama3.1	56.7 ± 2.1	53.1 ± 1.0	57.4 ± 1.57	**72.4♠ ± 1.3**
llama3.2	59.3 ± 1.6	44.4 ± 1.1	46.5 ± 1.7	**63.4♠ ± 1.5**
phi3:medium	86.2 ± 1.1	81.5 ± 1.4	84.5 ± 1.2	85.9 ± 1.1
phi3	63.6 ± 5.4	66.7 ± 2.1	64.5 ± 3.9	**76.2♠ ± 2.0**
SICK				
Modelos	**Base**	**AoT**	**SKA_MV**	**SKA_DT**
gemma2	88.0 ± 2.0	88.1 ± 1.8	88.4 ± 1.5	**89.1♦ ± 1.7**
gemma2:2b	78.0 ± 2.1	59.8 ± 1.9	77.2 ± 1.9	**84.1♠ ± 2.2**
llama3.1	80.4 ± 1.9	72.6 ± 1.5	81.1 ± 1.7	**85.1♠ ± 2.7**
llama3.2	68.0 ± 1.6	62.5 ± 1.9	62.4 ± 2.2	**77.1♠ ± 1.5**
phi3:medium	76.0 ± 1.4	76.5 ± 1.7	79.4♠ ± 1.6	81.6♠ ± 1.8
phi3	82.1 ± 1.6	84.1♠ ± 1.8	83.3♠ ± 2.5	**86.7♠ ± 1.8**

Cuadro 6.3: Comparación del rendimiento promedio de exactitud entre el modelo base, AoT, SKA_MV y SKA_DT para los corpus de 3 clases con p-value de las pruebas estadísticas de Mann-Whitney: ♠ indica un p-value<0,001, ♦ un p-value<0,01 y $*$ un p-value<0,05.

El método SKA_DT demuestra una superioridad en las tareas de clasificación de tres categorías. En el corpus SNLI, logra mejoras estadísticamente significativas (p-value$<$ 0.001) sobre el baseline, especialmente en modelos como `llama3.1` (de 56.7 % a 72.4 %) y `phi3` (63.6 % $\rightarrow$ 76.2 %). Este salto de rendimiento, que supera los 15 puntos porcentuales en `llama3.1`, evidencia que SKA_DT posee la capacidad para capturar relaciones semánticas complejas y matices lógicos necesarios para distinguir entre las tres clases. En el corpus SICK, aunque las mejoras absolutas son más modestas debido a un rendimiento base más alto, SKA_DT sigue siendo consistentemente el mejor método, logrando las puntuaciones más altas en 5 de los 6 modelos evaluados y confirmando su robustez en diferentes corpus.

RTEGLUE				
Modelos	**Base**	**AoT**	**SKA_MV**	**SKA_DT**
gemma2	87.6 ± 2.6	86.6 ± 2.8	87.7 ± 3.0	88.0 ± 2.5
gemma2:2b	74.5 ± 3.4	70.0 ± 4.3	73.0 ± 3.1	**75.1 ± 3.5**
llama3.1	75.3 ± 2.6	67.9 ± 1.8	77.2 ♠ ± 2.7	**79.4♠ ± 2.3**
llama3.2	71.4 ± 3.1	57.6 ± 2.3	71.8 ± 3.0	**74.5♠ ± 2.4**
phi3:medium	**86.8 ± 1.9**	82.4 ± 2.7	85.9 ± 2.4	86.5 ± 2.8
phi3	85.0 ± 3.8	**85.5 ± 1.94**	83.0 ± 2.7	85.1 ± 2.7
Scitail				
Modelos	**Base**	**AoT**	**SKA_MV**	**SKA_DT**
gemma2	80.1 ± 2.2	83.8♠ ± 2.2	84.1♠ ± 2.0	84.7♠ ± 1.7
gemma2:2b	74.7 ± 1.9	66.4 ± 2.2	77.3♠ ± 2.0	77.9♠ ± 2.2
llama3.1	65.8 ± 2.4	57.1 ± 1.9	71.3♠ ± 2.0	**77.2♠ ± 2.4**
llama3.2	63.6 ± 3.0	54.9 ± 1.0	73.0♠ ± 2.4	**73.0♠ ± 1.7**
phi3:medium	75.8 ± 2.6	71.9 ± 2.2	77.4* ± 3.1	**80.3♠ ± 2.9**
phi3	81.2 ± 1.8	83.2♠ ± 1.6	83.4♠ ± 1.7	**84.0♠ ± 1.6**

Cuadro 6.4: Comparación del rendimiento promedio de exactitud entre el modelo base, AoT, SKA_MV y SKA_DT para los corpus de dos clases con p-value de las pruebas estadísticas de Mann-Whitney: ♠ indica un p-value<0,001, ♦ un p-value<0,01 y $*$ un p-value<0,05.

Es crucial destacar que SKA_DT supera de manera sistemática a su variante SKA_MV. Esta comparación directa revela que la estrategia de integración mediante un árbol de decisiones (DT) es fundamental. Mientras que SKA_MV ya aporta mejoras (como se abordó en la literatura), el árbol de decisiones de SKA_DT aprende a combinar de forma óptima y la información de las diferentes lineas de razonamiento, extrayendo patrones más sofisticados. Esto se corrobora que, en los pocos casos donde SKA_MV no supera el base (`llama3.2` en SNLI con 46.5 %), SKA_DT sí lo hace de manera contundente (63.4 %), demostrando su capacidad para rescatar y sintetizar conocimiento donde otros métodos fallan.

En los corpus con dos clases, SKA_DT consolida su rendimiento, aunque las diferencias son menores. En el corpus RTEGLUE, logra la mayor exactitud en 4 de los 6 modelos, con mejoras estadísticamente significativas (♠) para la familia `llama`. Su desempeño es particularmente notable en modelos más pequeños o con mayor margen de mejora, como `gemma2:2b`, donde supera a todos los demás métodos. En el corpus SciTail, su dominio es aún más claro: SKA_DT obtiene la puntuación más alta en 5 de los 6 modelos, frecuentemente con diferencias estadísticas significativas.

Un hallazgo clave en esta tabla es la consistencia de SKA_DT frente a la estrategia de AoT y SKA_MV. Mientras que el método AoT muestra un rendimiento muy irregular, llegando a perjudicar el rendimiento base en algunos casos como `llama3.2` en SNLI o `gemma2:2b` en SICK, SKA_DT no lo hace y casi siempre lo mejora. Esta fiabilidad, sumada a que en el corpus SciTail supera incluso a SKA_MV en todos los modelos, refuerza la conclusión de que este mecanismos de SKA_DT no es solo un integrador eficaz, sino un mecanismos que extrae de manera confiable la información más relevante para la decisión, independientemente de que sea binaria o multiclase. Los resultados confirman que el verdadero salto cualitativo surge cuando el modelo puede apoyar el razonamiento, valorando la pertinencia y coherencia sobre la frecuencia.

Como se ha mencionado no es suficiente, validar con una métrica el impacto en el rendimiento del modelo, si no también explorar en dónde se ve el beneficio. Es así que para profundizar en el análisis de las predicciones realizadas, se requiere no solo considerar las predicciones correctas, si no de que etiqueta. En las siguientes tablas se muestran los resultados sobre la métrica F1-Score agrupando los resultados de los corpus de 3 clases: SNLI y SICK.

La Tabla 6.5 expone la principal dificultad de los modelos: clasificar correctamente la clase "Neutral". Para el corpus SNLI, mientras que el rendimiento en "Entailment" suele ser alto, las métricas para "Neutral" son consistentemente las más bajas y volátiles, por ejemplo en `llama3.2` (Baseline) de 15.4 %. SKA_DT aborda este problema y no solo mejora el F1-score para "Neutral"lo incrementa de 32.6 % a 51.6 %, sino que lo hace sin sacrificar el rendimiento en las otras clases. De hecho, su mayor logro es en la clase "Contradiction", donde logra mejoras con diferencias significativas (♠) en casi todos los modelos: `gemma2` de 36.5 % a 59.2 %, `phi3` de 48.8 % a 81.7 %, `llama3.1` de 58.7 % a 83.5 %. Esto demuestra que SKA_DT fortalece las debilidades críticas del modelo base.

En el corpus SICK, el patrón se mantiene pero con un rendimiento base más alto. SKA_DT logra el F1-score más alto en la gran mayoría los resultados, frecuentemente con diferencia significativa (♠). La mejora es particularmente notable en modelos más pequeños y con dificultades, para `llama3.2`, SKA_DT duplica el F1-score en "Neutral"de 27.4 % a 63.9 % y mejora sustancialmente "Contradiction"de 76.7 % a 84.3 %. Incluso en modelos con buenos resultados en el base, como `phi3`, SKA_DT proporciona ganancias marginales pero consistentes en las tres clases.

SNLI					
Modelos	**Clases**	**Base**	**AoT**	**SKA_MV**	**SKA_DT**
gemma2	Entail	86.1 ± 1.5	85.2 ± 1.7	86.1 ± 1.6	**86.4 ± 1.5**
	Neutral	59.1 ± 1.5	60.6♠ ± 1.4	61.3♠ ± 1.9	65.2♠ ± 1.4
	Contradiction	36.5 ± 4.5	40.7♠ ± 3.6	40.7♠ ± 4.3	59.2♠ ± 2.7
gemma2:2b	Entail	**74.8 ± 1.5**	64.2 ± 1.2	60.7 ± 2.9	**72.5 ± 2.4**
	Neutral	41.5 ± 2.8	19.3 ± 3.0	52.9♠ ± 2.1	49.9♠ ± 3.1
	Contradiction	69.3 ± 2.8	72.9♠ ± 2.5	57.8 ± 3.2	75.6♠ ± 2.1
llama3.1	Entail	72.5 ± 1.4	66.7 ± 1.1	76.4♠ ± 1.1	76.6♠ ± 2.2
	Neutral	32.6 ± 3.5	16.8 ± 3.2	42.7♠ ± 2.9	**51.6♠ ± 3.2**
	Contradiction	58.7 ± 3.7	62.7♠ ± 1.5	46.5 ± 2.9	**83.5♠ ± 1.4**
llama3.2	Entail	65.8 ± 1.4	59.2 ± 0.9	36.0 ± 3.1	66.5 ± 2.1
	Neutral	15.4 ± 3.5	11.4 ± 2.1	**45.0♠ ± 2.3**	31.8♠ ± 5.6
	Contradiction	79.0 ± 2.0	45.5 ± 3.7	58.6 ± 2.3	**80.6♠ ± 1.3**
phi3:medium	Entail	**87.5 ± 0.8**	84.5 ± 1.1	86.3 ± 1.1	87.1 ± 1.2
	Neutral	78.6 ± 1.8	71.9 ± 2.2	76.5 ± 2.0	**78.6 ± 1.8**
	Contradiction	**91.8 ± 1.8**	87.6 ± 2.1	90.2 ± 1.7	91.5 ± 1.8
phi3	Entail	75.5 ± 3.3	**79.5♠ ± 2.3**	62.4 ± 5.8	77.9♦ ± 3.2
	Neutral	62.9 ± 3.7	63.9 ± 1.8	64.1 ± 2.7	**70.1♠ ± 2.0**
	Contradiction	48.8 ± 13.2	54.9 ± 4.2	67.1♠ ± 5.0	**81.7♠ ± 2.6**
SICK					
Modelos	**Clases**	**Base**	**AoT**	**SKA_MV**	**SKA_DT**
gemma2	Entail	91.8 ± 1.7	91.3 ± 1.3	92.0 ± 1.0	92.1 ± 1.3
	Neutral	82.1 ± 3.2	81.8 ± 3.2	82.1 ± 2.5	**83.8♦ ± 2.7**
	Contradiction	89.9 ± 1.6	90.8♦ ± 1.3	90.6* ± 1.6	**91.2♠ ± 1.9**
gemma2:2b	Entail	86.8 ± 2.1	69.8 ± 2.13	80.4 ± 3.7	**87.4 ± 1.8**
	Neutral	58.8 ± 4.8	13.5 ± 4.2	65.5♠ ± 2.5	**75.3♠ ± 3.6**
	Contradiction	82.5 ± 1.8	70.9 ± 2.9	84.3♠ ± 1.7	**88.8♠ ± 2.3**
llama3.1	Entail	84.8 ± 1.3	77.2 ± 1.8	86.2♠ ± 1.4	**88.4♠ ± 2.1**
	Neutral	62.4 ± 5.4	40.3 ± 4.5	65.5♠ ± 4.8	**77.0♠ ± 4.8**
	Contradiction	88.0 ± 1.9	86.1 ± 1.8	87.0 ± 2.3	**89.2♦ ± 2.5**
llama3.2	Entail	79.8 ± 2.9	66.3 ± 1.6	59.4 ± 4.6	**81.6♦ ± 1.3**
	Neutral	27.4 ± 5.9	32.6♠ ± 5.2	46.4♠ ± 3.8	**63.9♠ ± 2.4**
	Contradiction	76.7 ± 1.3	76.6 ± 2.6	76.3 ± 1.4	**84.3♠ ± 1.5**
phi3:medium	Entail	89.5 ± 1.6	**91.9♠ ± 1.1**	90.7♠ ± 1.7	91.4♠ ± 1.3
	Neutral	47.5 ± 4.9	49.7* ± 6.1	58.8♠ ± 4.8	65.9♠ ± 4.3
	Contradiction	79.6 ± 1.6	78.5 ± 1.5	81.8♠ ± 1.4	83.5♠ ± 1.8
phi3	Entail	85.7 ± 2.0	88.2♠ ± 2.2	88.9♠ ± 3.0	**90.8♠ ± 1.6**
	Neutral	77.3 ± 2.9	76.6 ± 3.0	74.7 ± 3.7	**80.0♠ ± 3.2**
	Contradiction	83.3 ± 2.2	87.3♠ ± 1.7	86.0♠ ± 1.9	88.9♠ ± 2.0

Cuadro 6.5: Rendimiento promedio de F1-score del modelo base, AoT, SKA_MV y SKA_DT para los corpus de 3 clases con pruebas estadísticas de Mann-Whitney: ♠ indica un p-value<0,001, ♦ un p-value<0,01 y $*$ un p-value<0,05.

Un hallazgo es que SKA_DT nunca perjudica el rendimiento en ninguna clase, a diferencia de AoT y, en ocasiones, SKA_MV. AoT muestra una caída en "Neutral"para gemma2:2b: de 58.8 % a 13.5 %. SKA_DT parece identificar de manera confiable qué líneas de razonamiento son relevantes para las clases objetivo.

RTEGLUE					
Modelos	**Clases**	**Base**	**AoT**	**SKA_MV**	**SKA_DT**
gemma2	Entail	88.5 ± 2.3	87.4 ± 2.4	88.3 ± 2.6	88.4 ± 2.3
	Not-entailment	86.6 ± 3.0	85.7 ± 3.2	87.0 ± 3.4	87.5* ± 2.9
gemma2:2b	Entail	**75.3 ± 3.3**	74.8 ± 3.4	70.2 ± 3.8	75.2 ± 3.6
	Not-entailment	73.6 ± 3.7	62.9 ± 6.0	**75.3♦ ± 2.6**	75.0 ± 3.6
llama3.1	Entail	79.8 ± 1.7	75.4 ± 1.1	81.0♦ ± 2.0	80.9* ± 2.1
	Not-entailment	68.0 ± 4.3	53.8 ± 3.6	71.7♠ ± 3.9	**77.5♠ ± 3.3**
llama3.2	Entail	76.1 ± 2.4	69.3 ± 1.5	71.4 ± 3.2	**76.4 ± 2.0**
	Not-entailment	64.5 ± 4.4	31.1 ± 4.9	72.1♠ ± 3.0	**72.2♠ ± 3.1**
phi3:medium	Entail	**87.9 ± 1.6**	84.8 ± 2.1	87.3 ± 2.0	87.4 ± 2.8
	Not-entailment	**85.5 ± 2.2**	78.9 ± 3.8	84.1 ± 3.0	85.4 ± 3.1
phi3	Entail	84.7 ± 4.0	**85.0 ± 2.05**	81.5 ± 2.9	84.7 ± 2.8
	Not-entailment	85.2 ± 3.7	**86.0 ± 1.9**	84.3 ± 2.5	85.5 ± 2.7
Scitail					
Modelos	**Clases**	**Base**	**AoT**	**SKA_MV**	**SKA_DT**
gemma2	Entail	82.5 ± 1.8	85.1♠ ± 2.1	85.3♠ ± 1.9	85.0♠ ± 1.6
	Not-entailment	76.8 ± 2.6	82.2♠ ± 2.4	82.7♠ ± 2.2	84.4♠ ± 1.9
gemma2:2b	Entail	78.8 ± 1.4	74.1 ± 1.4	78.2 ± 1.9	79.4 ± 2.0
	Not-entailment	68.7 ± 2.9	52.1 ± 4.2	76.4♠ ± 2.1	76.2♠ ± 2.5
llama3.1	Entail	74.3 ± 1.4	69.8 ± 1.0	77.5♠ ± 1.3	**79.1♠ ± 2.0**
	Not-entailment	48.9 ± 5.0	25.6 ± 5.4	60.6♠ ± 3.6	**74.8♠ ± 3.1**
llama3.2	Entail	72.2 ± 2.0	68.6 ± 0.6	**74.5♠ ± 2.4**	72.2 ± 2.1
	Not-entailment	47.1 ± 5.8	20.2 ± 3.0	71.3♠ ± 2.5	**73.7♠ ± 1.6**
phi3:medium	Entail	80.1 ± 1.9	77.8 ± 1.5	81.2* ± 2.3	**83.1♠ ± 2.3**
	Not-entailment	69.3 ± 3.9	61.6 ± 3.8	71.4 ± 4.7	**76.4♠ ± 4.0**
phi3	Entail	83.1 ± 1.5	84.2♠ ± 1.4	**84.3♠ ± 1.5**	84.2♠ ± 1.6
	Not-entailment	78.9 ± 2.4	81.9♠ ± 2.0	82.4♠ ± 1.9	**83.7♠ ± 1.7**

Cuadro 6.6: Rendimiento promedio de F1-score del modelo base, AoT, SKA_MV y SKA_DT para los corpus de dos clases con pruebas estadísticas de Mann-Whitney: ♠ indica un p-value<0,001, ♦ un p-value<0,01 y * un p-value<0,05.

En el corpus RTEGLUE, SKA_DT se consolida especialmente donde más se necesita. Mientras que los modelos base ya tienen un buen rendimiento en Entailment para familia de modelos como `Gemma2` y `phi3`, la verdadera prueba está en la clase Not-entailment. Not-Entailment, es una categoría que agrupa Neutral y Contradiction. Aquí, SKA_DT logra mejoras significativas (♠) en los modelos que más les cuesta esta distinción. Por ejemplo, para `llama3.1`, eleva el F1-score de Not-entailment de un 68.0 % base a un 77.5 %, y para Llama3.2, de un 64.5 % a un 72.2 %. Este patrón demuestra que SKA_DT no solo mantiene el rendimiento en las tareas fáciles, sino que proporciona una capacidad de discriminación crítica en las clases difíciles, donde la simple votación por mayoría (SKA_MV) o otros métodos (AoT) muestran una confiabilidad inconsistente y a veces perjudicial.

Un hallazgo crucial en RTEGLUE es la estabilidad de SKA_DT frente a la volatilidad de AoT. El método AoT presenta caídas dramáticas y perjudiciales en el rendimiento, particularmente en la clase Not-entailment (ejemplo en `Llama3.2`, colapsa del 64.5 % base a un 31.1 %). En contraste, SKA_DT nunca perjudica al modelo base de manera significativa y consistentemente lo iguala o supera. Esto subraya que la ventaja de SKA_DT no es solo de rendimiento, sino de confiabilidad.

Por último, en el corpus SciTail, SKA_DT reafirma su superioridad como el mecanismo definitivo para el razonamiento en dominios distintos, logrando mejores rendimiento en toda la evaluación. El corpus revela una debilidad inherente de los modelos base, particularmente en la clase Not-entailment. Para `llama3.1`, duplica el F1-score en Not-entailment"pasando de de 48.9 % a 74.8 % y para `llama3.2` lo lleva de un 47.1 % a un 73.7 %. Estas mejoras, con una diferencia significativa (♠) demuestran la capacidad única de SKA_DT para la clase Not-entailment.

En los resultados presentados se muestra que el conocimiento aprendido de los LLM (base) no es suficiente para la tarea, comparado cuando se le proporciona la información relevante sobre las relaciones que ocurren entre el par $< P, H >$, ya unificada la respuesta final con SKA_DT. En las tablas, también se muestran los resultados de las pruebas estadísticas para medir diferencias significativas sobre los resultados. La prueba estadística Mannwhitneyu demostró diferencias con p-value < 0.001 para 5 modelos. Para el caso del LLM `gemma2` obtuvo un p-value < 0.01.

El método considera la direccionalidad requerida en la evaluación de la implicación para definir clases abstractas de compatibilidad e incompatibilidad semántica que guían el proceso de razonamiento de los LLM, revelando inconsistencias en las respuestas y debilidades.

6.2. Diagnóstico sobre fenómenos lingüísticos

El *benchmark General Language Understanding Evaluation* (GLUE) contiene una colección de recursos para entrenar, evaluar y analizar sistemas de comprensión del lenguaje natural. Además, brinda un corpus de diagnóstico[6], con respecto a fenómenos lingüísticos que se encuentran en el lenguaje natural. El corpus se proporciona únicamente como una herramienta de análisis para describir a grandes rasgos los tipos de fenómenos que un modelo puede o no capturar, para análisis de errores y comprender mejor las capacidades de los modelos. Los creadores del corpus de diagnóstico hacen una anotación de que el proceso de etiquetado manual de clases de neutralidad y contradicción puede volverse subjetivo por parte de los etiquetadores. Por esta razón, para el caso del diagnóstico, en este benchmark, solo se establecen dos clases: entailment y not-entailment.

El objetivo es profundizar en los fenómenos lingüísticos que se producen en las tareas del RIT y que ya se han identificado en el corpus de diagnóstico. El corpus de diagnóstico tiene 1104 pares $< P, H >$. La clase Entailment tiene 460 ejemplos, mientras que la clase Not-entailment tiene 644. Para entrenar SKA_DT, seleccionamos aleatoriamente una configuración equilibrada de 200 pares $< P, H >$ por clase. Los 704 pares restantes (260 de Entailment y 444 de not-entailment) se utilizaron para las pruebas. Esta configuración se utilizó para la validación cruzada (kfold=5). Los dominios de donde provienen estos pares $< P, H >$ del diagnóstico son: 1) Artificial, 2) News, 3) Wikipedia, 4) ACl y 5) Reddit. Los aspectos lingüísticos se dividen en 4 categorías que se enlistan a continuación con algunos ejemplos tomados directamente del corpus de diagnóstico:

- **Lexical Semantics**: El entailment se puede aplicar a nivel de frase y a nivel de palabra. También existen relaciones simétricas y entidades nombradas.
  - "I am refusing to do X" contradice "I am doing X"
  - "John married Gary" implica "Gary married John"
  - SNL significa Saturday Night Live
- **Predicate-Argument Structure**: En esta categoría es necesario tratar con la ambigüedad sintáctica, roles semánticos y coreferencia. Aunado a esto, la identificación de la acción del sujeto sobre su objeto.
  - "Jake broke the vase" implica "the vase broke".
  - "Jake broke the vase" no implica "Jake broke".
- **Logic**: En la implicatura también se toma en cuenta la lógica matemática, a través del procesa-

[6] https://gluebenchmark.com/diagnostics

miento de cuantificadores y razonamiento sobre temporalidad.

- "I have had more than 2 drinks tonight" implica "I have had more than 1 drink tonight".
- "The cat sat on the mat" contradice "The cat did not sit on the mat".

- **Knowledge and Common Sense**: En esta categoría se toma en cuenta la aplicación de conocimiento adicional o sentido común sobre significados de palabras.
  - "This is the most oniony article I've seen on the entire internet" implica "This article reads like satire".

Se analiza cuánta mejora o pérdida hay comparando los ejemplos que el modelo LLM responde correctamente e incorrectamente con la propuesta de forma general. La Tabla 6.7 muestra los resultados. "Mejora" se refiere a la proporción de predicciones correctas según la propuesta SKA_DT del número total de fallos del modelo de referencia y "Pérdida" se refiere a la proporción de fallos según la propuesta SKA_DT del número total de predicciones correctas del modelo de referencia.

	`gemma2:2b`	`gemma2`	`llama3.2`	`llama3.1`	`phi3`	`phi3:medium`
Mejora	27.6 %	16.5 %	37.3 %	27.0 %	19.8 %	9.9 %
Pérdida	11.7 %	4.7 %	18.1 %	8.5 %	6.6 %	3.6 %

Cuadro 6.7: Mejora general y pérdida para el corpus de diagnóstico.

La capacidad de SKA_DT tiene debilidades, en la Figura 6.4 se muestra lo siguiente: Cuando todas las respuestas $(G_1 - G_4)$ del LLM coinciden y son correctas entonces se dice que hay consistencia en las respuestas de los LLMs. Cuando al menos una de las respuestas $(G_1 - G_4)$ del LLM es correcta entonces se dice que no hay consistencia. SKA_DT muestra el rendimiento actual.

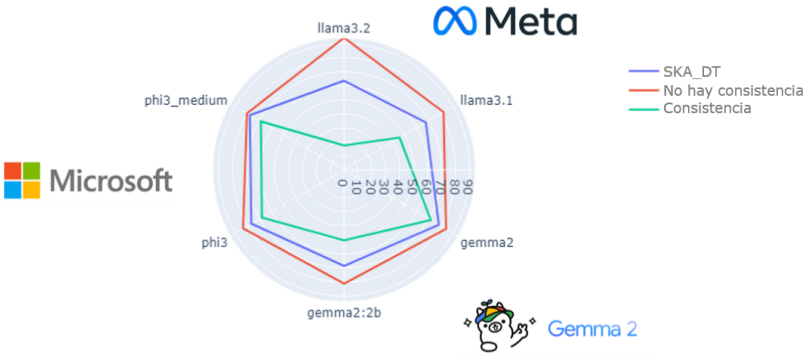


Figura 6.4: Respuestas con consistencia, no consistencia y SKA_DT.

Aquí, el LLM demuestra que su conocimiento es accesible pero inestable. El árbol de decisión actúa como un mecanismo de estabilización y consolidación: al analizar y ponderar las múltiples (y a veces contradictorias) respuestas del LLM, puede "rescatar" la señal correcta del ruido y producir una salida final más precisa y confiable que cualquiera de las respuestas individuales inconsistentes. En otras palabras, el SKA_DT actúa como un optimizador de confiabilidad para respuestas de LLMs cuyo rendimiento es variable. Su éxito se mediría por su capacidad para mover resultados de la categoría "No hay consistencia" hacia una decisión final correcta y robusta, acercándose así al ideal de "Consistencia" que los modelos base, por sí solos, no alcanzan de forma fiable. Esto indica que la consistencia es una métrica crucial para la fiabilidad en aplicaciones reales, y que técnicas de post-procesamiento como SKA_DT pueden ser esenciales para lograr un rendimiento robusto a partir de modelos subyacentes imperfectos.

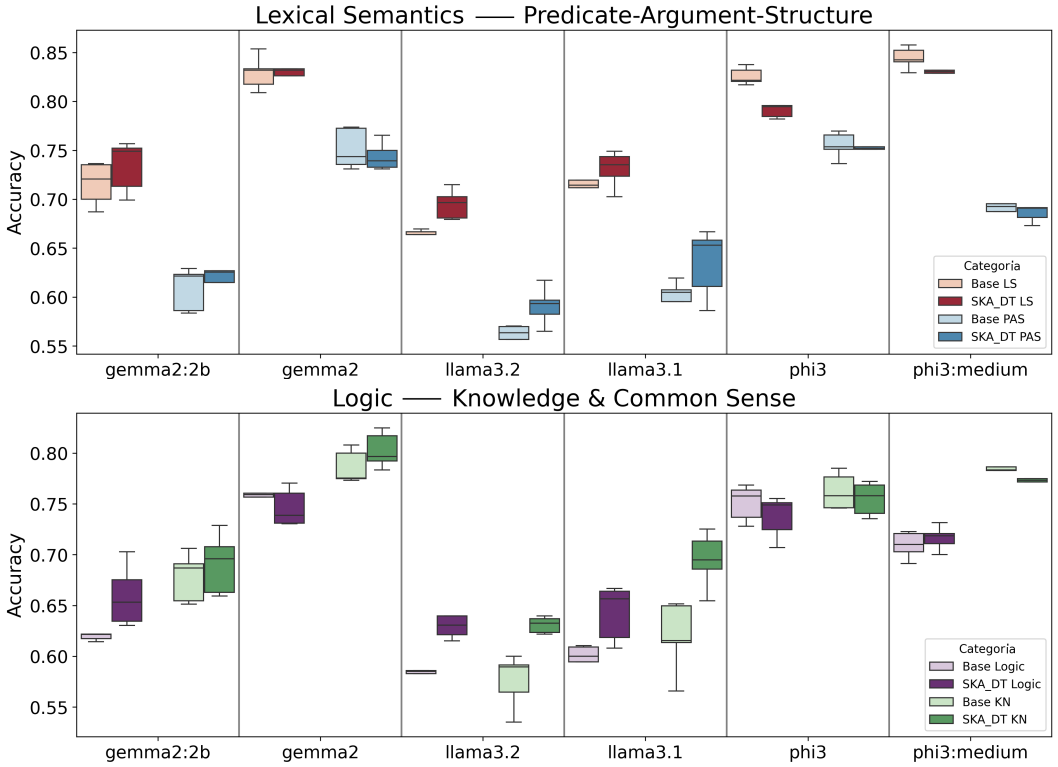


Figura 6.5: Accuracy de la validación cruzada de todos los modelos en el corpus de diagnóstico por categoría lingüística. La fila superior muestra el rendimiento en las categorías Lexical Semantics (LS) y Predicate-Argument Structure (PAS). La fila inferior muestra el rendimiento en las categorías Knowledge and Common Sense (KN) y Logic. Los colores intensos y tenues se refieren a SKA_DT y al modelo base, respectivamente. En promedio, hay 233 ejemplos en la categoría LS, 263 en la categoría PAS, 234 en la categoría Lógica y 182 en la categoría KN. Algunos ejemplos pertenecen a más de una categoría.

La Figura 6.5 muestra la precisión de nuestra propuesta en comparación con las categorías lingüísticas del corpus de diagnóstico.

A continuación se muestran algunos ejemplos por categoría, donde el SKA_DT acertó para la clase entailment y not_entailment.

Lexical Semantic	P: A calm wind rolled across the glade. H: A serene wind rolled across the glade. $\mathcal{C}$: Entailment	M: gemma2:2b G1: [(calm, synonym, serene), (calm wind, same, serene wind), (roll, same, roll), (glade, same, glade)] DT: $\mathbf{G_1 : E}$, $G_2 : NE$, $G_3 : NE$, $G_4 : NE$ Pred.: Not Entailment
Predicate Argument Structure	P: They should be attached to the lifting mechanism in the faucet. H: They should be attached in the faucet. $\mathcal{C}$: Entailment	M: phi3 G1: [(should attach, same, should attach), (faucet, same, faucet)] DT: $G_1 : NE$, $G_2 : NE$, $G_3 : NE$, $\mathbf{G_4 : E}$ Pred.: Not Entailment
Logic	P: Susan knows how turtles reproduce. H: Cedric doesn't know how turtles reproduce. $\mathcal{C}$: Not Entailment	M: llama3.1 G1: [(turtle, same, turtle), (how reproduce, same, how reproduce)] G2: [(susan know, distinct_from, cedric do not know)] G4: [(, UNK, cedric)] DT: $\mathbf{G_1}$: **NE**, $\mathbf{G_2}$: **NE**, $\mathbf{G_3}$: **NE**, $G_4 : E$ Pred.: Entailment
Knowledge & Common Sense	P: I ate until I was full. H: I ate until it was uncomfortable to eat more. $\mathcal{C}$: Entailment	M: phi3:medium G1: [(until, same, until)] G3: [(more eat, is_a, eat)] G4: [(, UNK, uncomfortable)] DT: $G_1 : NE$, $\mathbf{G_2 : E}$, $G_3 : NE$, $G_4 : NE$ Pred.: Not Entailment

Cuadro 6.8: Ejemplos de pares $< P - H >$ en los que el esquema SKA_DT falla.

La propuesta presentada tiene áreas de mejora, en la Tabla 6.8 se muestran algunos ejemplos donde no logra realizar la predicción correcta. La última columna representa el modelo, los grupos abstractos de SKA utilizados en el prompt, las respuestas de cada prompt basada en grupos (la entrada al SKA_DT) y la clase predicha (salida del SKA_DT). Los grupos solo se muestran si contienen relaciones semánticas. Un ejemplo de *prompting* a los LLMs se encuentra en la Figura 6.6.

```
You are an expert in Recognizing Textual Entailment over pairs of Premise and Hypothesis. Based on
the background information provide below, classify the relationship between the given Premise and
Hypothesis as one of the following: Entailment, Neutral or Contradiction. Respond only using the
template:{‘‘Answer’’:}. Do not modify the template.

Premise and hypothesis to classify:
Premise: A calm wind rolled across the glade.
Hypothesis: A serene wind rolled across the glade.

Background Information:
G1: Triplets of Generality or Equivalence. These relations usually correspond to Entailment. Triplets:
(p_i,rel,h_j) where p_i is in Premise, h_j is in Hypothesis and rel is the relation between p_i and h_j. e.g.,
(dog,is_a,animal) dog is in the premise, animal is in the hypothesis and the relationship is general from dog to
animal.

Word relations group:
G1: [(' calm', 'synonym', ' serene'), ('calm wind', 'same', 'serene wind'), (' roll', 'same', ' roll'), (' glade', 'same',
' glade')]
```

Figura 6.6: Ejemplo de prompt SKA con G_1. G_1 contiene la lista de relaciones semánticas que pertenecen al grupo abstracto según el marco metodológico de la abstracción semántica. Además, se añaden su definición de lo que contiene el grupo y la alineación con la clase de la tarea.

Estudio de ablación

Los resultados del estudio de ablación revelan un panorama respecto a la eficacia de las distintas estrategias de integración de conocimiento semántico externo. La Tabla 6.9 muestra los resultados obtenidos.

La estrategia más consistentemente beneficiosa es SKE (Conocimiento Semántico de Entidades de ConceptNet), la cual tiene diferencias significativas, particularmente en los modelos de la familia `Gemma2`. Mientras que `Gemma2` responde positivamente a SKE e incluso a SKC (Conocimiento Semántico directo de ConceptNet) los modelos de la familia `llama3` presentan un comportamiento más heterogéneo. Para `llama3.1` y `llama3.2`, la estrategia SKA_DT obtiene el mejor rendimiento, lo que apunta a que este modelo se beneficia de una representación más procesada y abstracta del conocimiento, en línea con la hipótesis central de esta investigación.

Modelos	Accuracy	Entailment F1-score	Not_entailment F1-score	Modelos	Accuracy	Entailment F1-score	Not_entailment F1-score
`Vega v2`	43.3	56.8	17.6	`Vega v2`	43.3	56.8	17.6
`Majority class`	58.3	0	73.7	`Majority class`	58.3	0	73.7
`gemma2` base	77.2 ± 1.2	74.5 ± 1.2	79.3 ± 1.2	`gemma2:2b` base	65.5 ± 1.2	65.3 ± 1.0	65.7 ± 1.4
+AoT	76.6 ± 1.1	67.4 ± 1.6	81.8* ± 0.9	+AoT	53.5 ± 1.0	59.2 ± 0.8	46.0 ± 1.3
+SKC	77.6 ± 1.03	74.8 ± 1.0	79.9 ± 1.1	+SKC	67.2 ± 0.85	63.8 ± 0.8	70.1♦ ± 0.9
+SKE	**80.0* ± 1.3**	**76.2 ± 1.4**	**82.8♦ ± 1.2**	+SKE	**70.2♦ ± 1.7**	**67.3 ± 1.7**	**72.6♦ ± 1.8**
+SKA	77.1 ± 1.0	73.9 ± 1.0	79.6 ± 1.1	+SKA	66.1 ± 1.5	61.4 ± 1.9	69.8 ± 1.19
`llama3.1` base	62.3 ± 0.5	64.8 ± 0.2	59.4 ± 0.9	`llama3.2` base	58.6 ± 0.7	60.3 ± 0.7	56.7 ± 0.8
+AoT	53.7 ± 0.5	61.0 ± 0.2	43.2 ± 1.2	+AoT	49.3 ± 1.1	57.9 ± 0.6	36.4 ± 2.2
+SKC	61.8 ± 0.9	64.7 ± 0.6	58.3 ± 1.4	+SKC	59.0 ± 0.82	**62.3♦ ± 0.9**	55.2 ± 1.0
+SKE	62.7 ± 0.8	65.1 ± 0.5	59.9 ± 1.3	+SKE	59.4 ± 1.0	62.1♦ ± 1.0	56.2 ± 1.2
+SKA	**67.2♦ ± 1.2**	**65.4 ± 1.2**	**68.7♦ ± 3.0**	+SKA	**63.4* ± 0.8**	54.5 ± 3.2	**69.3♦ ± 1.2**
`phi3:medium` base	74.9 ± 0.6	74.0 ± 0.4	75.6 ± 0.8	`phi3` base	**76.3 ± 0.8**	72.4 ± 0.9	79.3 ± 0.7
+AoT	**76.0* ± 1.0**	73.6 ± 0.9	**78.1♦ ± 1.0**	+AoT	75.6 ± 0.5	71.2 ± 1.1	78.8 ± 0.4
+SKC	73.7 ± 0.93	73.1 ± 0.6	74.2 ± 1.2	+SKC	75.0 ± 0.86	71.1 ± 1.1	78.0 ± 0.7
+SKE	75.5 ± 1.0	**74.5 ± 0.8**	76.5 ± 1.2	+SKE	76.1 ± 0.7	**72.8 ± 0.8**	78.7 ± 0.7
+SKA_DT	74.6 ± 1.1	73.1 ± 1.0	75.9 ± 1.35	+SKA_DT	75.9 ± 1.1	70.4 ± 0.5	**79.7 ± 1.4**

Cuadro 6.9: Resultados del estudio de ablación en el corpus de diagnóstico. Añadimos el enfoque de SKC: conocimiento semántico directo de ConceptNet, y SKE: conocimiento semántico de entidades de ConceptNet. Pruebas estadísticas de Mann-Whitney: ♠ indica un p-value<0,001, ♦ un p-value<0,01 y * un p-valuep<0,05.

En contraste, el modelo base `phi3`, es muy competente, muestra ganancias marginales o incluso ligeras perdidas con las estrategias mostradas. Esto sugiere un posible que su proceso interno de razonamiento ya es lo suficientemente robusto, haciendo redundante o incluso contraproducente dotar de ciertos tipos de conocimiento estructurado de forma externa. Sin embargo, una interpretación posible es la contaminación en los datos del pre-entrenamiento.

Al tratarse de un corpus público, existe la posibilidad de que los LLMs evaluados hayan sido expuestos al mismo durante su fase de entrenamiento. Estrategias diseñadas para guiar y mejorar un proceso de razonamiento genuino, como SKA_DT, se vuelven irrelevantes o incluso contraproducentes ante esta situación. Este mecanismo SKA_DT, que en otros corpus demostró una utilidad considerable al aprender a seleccionar las rutas de razonamiento más confiables, no logra superar el modelo base `phi3` en este corpus en particular.

Los resultados demuestran empíricamente que la integración de conocimiento semántico externo es una forma viable para mejorar el razonamiento de los LLMs, pero su éxito está condicionado por una mezcla del tipo de integración (entidades y/o abstracción), la arquitectura del modelo y la naturaleza de la tarea.

El desempeño de SKA_DT sirve como valiosa evidencia de que los avances más sustanciales en comprensión del lenguaje vendrán de técnicas que permitan a los modelos construir y manipular representaciones semánticas y conocimiento semántico abstracto, tal como se propone con el enfoque de abstracción semántica que motiva esta investigación.

6.3. Comparación de SKA con otros métodos

Para comprender el valor diferencial del marco metodológico propuesto SKA en el contexto actual de la investigación en IA, resulta esencial contrastarlo explícitamente con las principales familias de enfoques del estado del arte. Esta comparación permite visualizar no solo las diferencias técnicas, sino el nicho específico que SKA ocupa dentro del ecosistema de soluciones para mejorar el razonamiento en modelos de lenguaje.

SKA vs Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) propuesto por (Lewis y cols., 2020) se ha consolidado como una de las aproximaciones más extendidas para mitigar las alucinaciones y mejorar la veracidad factual de los LLMs. RAG opera mediante la recuperación de documentos o fragmentos relevantes desde una base de conocimiento externa, que luego se incorporan al contexto del modelo para fundamentar su respuesta.

La distinción fundamental reside en que RAG actúa como una memoria extensible: el modelo accede a información no presente en su entrenamiento, pero procesa esa información de manera fundamentalmente lineal y factual. El conocimiento recuperado se integra como texto adicional en el *prompt*, y el modelo aplica sobre él sus capacidades estándar de procesamiento secuencial.

SKA, por el contrario, no se limita a recuperar hechos, atiende la tarea del RIT al reconfigurar activamente las relaciones semánticas entre las entidades presentes en la premisa y la hipótesis. Mientras RAG responde a la pregunta ¿qué información adicional existe sobre estas entidades?, SKA aborda la cuestión más profunda de ¿cómo se relacionan en un plano abstracto y qué tipo de inferencia permite esa relación?.

Los resultados reportados en la literatura sugieren que, aunque RAG mejora significativamente tareas intensivas en conocimiento factual, su contribución en tareas de inferencia que requieren comprensión de relaciones jerárquicas, como distinguir entre “perro” y “animal” frente a “perro” y “gato”, es limitada, precisamente porque carece del mecanismo de abstracción que SKA introduce.

SKA vs Knowledge-Augmented Language Models (KALMs)

Los modelos como IERL, ERNIE, K-BERT o KEPLER ((Zi y cols., 2023a), (Baidu-ERNIE-Team, 2025), (W. Liu y cols., 2020) o (X. Wang y cols., 2021) respectivamente) representan un esfuerzo por incorporar conocimiento estructurado directamente en la arquitectura o el entrenamiento de los modelos de lenguaje. Estos enfoques integran tripletas de grafos de conocimiento mediante mecanismos como la alineación de entidades, la inyección en árboles de frases o el entrenamiento conjunto con objetivos de *knowledge embedding*.

La diferencia de la estrategia se caracteriza como una oposición entre fusión y abstracción. Los *KALMs* modifican el espacio de representación del modelo para que las representaciones de entidades y relaciones queden alineadas con el conocimiento factual. *ERNIE*, por ejemplo, aprende representaciones conjuntas de texto y entidades; *KEPLER* optimiza simultáneamente objetivos de lenguaje y de incrustación de conocimiento. IERL opera en el plano vectorial y un mero correlato superficial. Para el RIT, no es suficiente detectar asociaciones frecuentes entre palabras, sino evaluar si una hipótesis se sigue necesariamente de una premisa.

SKA no compite en este terreno de la representación interna, sino que opera en un nivel metacognitivo superior. En lugar de preguntarse ¿cómo incorporamos el conocimiento en los pesos del modelo?, SKA se plantea como un modelo con representaciones fijas, y se enfoca en: ¿cómo organizamos el conocimiento externo en categorías lógicas que guíen su proceso de inferencia?. Esta distinción es crucial porque permite que SKA sea aplicable a modelos ya entrenados sin necesidad de reentrenamiento, y porque aborda un aspecto que los *KALMs* no resuelven: la rigidez cognitiva para navegar entre dominios semánticos distantes.

Los KALMs actúan como una memoria extendida de hechos y entidades; SKA actúa como un meta-razonador que organiza ese conocimiento en categorías abstractas para asegurar que la lógica de la inferencia sea consistente y sistemática.

SKA vs abstracción en LLMS

Investigaciones recientes han explorado la abstracción desde ángulos complementarios. AbsPyramid, propuesto por (Z. Wang y cols., 2024) constituye un corpus y grafo de relaciones para evaluar la capacidad de los LLMs de detectar o generar abstracciones. Sus resultados experimentales demuestran que los LLM actuales enfrentan serios desafíos para comprender el conocimiento de abstracción. A pesar de sus carencias iniciales, el estudio concluye que los LLM pueden adquirir habilidades básicas de abstracción si son entrenados específicamente con un corpus rico en este tipo de conocimiento. (Hong y cols., 2024) proponen Abstraction of Thought (AoT), un formato de razonamiento estructurado que obliga al modelo a generar un abstracciones del proceso antes de incorporar detalles concretos.

La diferenciación clave se sitúa en el plano de la direccionalidad y el propósito. AoT es fundamentalmente una estrategia de *prompting* o *fine-tuning* que organiza el razonamiento interno del modelo y parte del supuesto de que los LLMs cuentan con el conocimiento suficiente para atender las tareas del PLN. Pero como hemos analizado en las secciones anteriores, los LLMs carecen de flexibilidad en su red conceptual y no logran reconocer relaciones semánticas de manera explícita. Más aún (Shojaee y cols., 2025) sugieren que los LRM (Large Reasoning Models) diseñados explícitamente para abordar tareas de razonamiento complejas, enfrentan barreras fundamentales que el simple aumento de cómputo o datos no podrá resolver fácilmente. En este sentido, *AbsPyramid* es un recurso de evaluación que permite medir capacidades de abstracción, pero no prescribe cómo aplicarlas en tareas concretas.

SKA integra ambas dimensiones: utiliza conocimiento externo (ConceptNet) y lo categoriza según compatibilidad semántica para inducir vías de razonamiento alternativas que luego se consolidan mediante el mecanismo de consenso SKA_DT. Los resultados muestran que esta aproximación produce mejoras sustanciales en la tarea del RIT, especialmente en la clase de no-implicatura.

Capítulo 7

Discusión

Actualmente, ha aumentado la preocupación por desarrollar sistemas que sean capaces de tomar decisiones similares a los humanos. En este sentido, (A. Wang, Singh, y cols., 2019) (A. Wang, Pruksachatkun, y cols., 2019a) proponen el *benchmark GLUE* y *SUPERGLUE* para distintas tareas de PLN. Además proporcionan corpus etiquetados que prometen elevar la dificultad de las tareas y un corpus de diagnóstico sobre categorías lingüísticas para analizar las habilidades de los modelos. Los pares de $< P, H >$ de este corpus fueron creados para evitar que los modelos dependan únicamente de señales léxicas simples y estadísticas para tomar una decisión. El corpus de diagnóstico permite identificar que fenómenos lingüísticos capturan los modelos, para mejorarlos y para compararse con los demás. Este corpus ha sido etiquetado por humanos con la intención de identificar áreas de oportunidad y fortalezas de los modelos. También muestra una tabla de clasificación para el seguimiento del rendimiento de los modelos que atienden tareas en el PLN[1]. Lo cuál es necesario para desarrollar nuevas estrategias que realmente atiendan las tareas del PLN.

La implementación de la propuesta, por ejemplo, combina técnicas de prompting, información de bases de conocimientos y mecanismos de consenso. En línea con la metodología (Y. Li y cols., 2023; Kasner y cols., 2023; Huang y cols., 2024; Mu y cols., 2024), exploramos diferentes líneas de razonamiento para la misma pregunta. En nuestro trabajo, construimos grupos abstractos de relaciones semánticas que influyen en el razonamiento de los LLMs.

Se evaluó un esquema tradicional de votación mayoritaria. Donde se parte del supuesto de que, cuando se formula la misma pregunta, independientemente de cómo se plantee al modelo, este debería llegar a la misma respuesta y ser coherente en su razonamiento. En otras palabras, el modelo tendría que seguir la misma línea de razonamiento independientemente de cómo se le plantee. Sin embargo, se

[1] https://super.gluebenchmark.com/leaderboard

ha demostrado que los LLMs pueden simular patrones de razonamiento no monótonos en escenarios controlados, gestionando excepciones sin ignorar las reglas generales. No obstante, esta capacidad sigue siendo superficial y frágil (Leidinger y cols., 2024; Z. Li y cols., 2025). Cuando se exponen a información irrelevante o contradictoria, los LLM tienden a abandonar las creencias previamente establecidas. No logran demostrar de manera consistente el dominio de esta capacidad de una manera robusta, generalizable y lógicamente coherente, tanto en contextos de lenguaje natural como en lógica formal (Xiu y cols., 2022; Leidinger y cols., 2024; Z. Li y cols., 2025).

La Figura 6.2 confirma esta debilidad, ya que el rendimiento de los modelos mejora o empeora en función del tipo de información proporcionada por cada grupo. No obstante, las tablas 6.3 y 6.4 muestran que nuestro mecanismo SKA_DT contribuye a mejorar la precisión de forma sistemática, teniendo en cuenta las respuestas dadas por los procesos de razonamiento provocados por cada grupo, lo que aumenta la precisión en todos los modelos.

Lo que muestran los resultados es que, cuando se les formula la misma pregunta, al menos uno de los grupos proporciona información disyuntiva relevante sobre los elementos de la pregunta (Figura 6.2). Por lo tanto, los modelos siguen diferentes líneas de razonamiento en cada caso, como también muestra la Tabla 6.8 con las entradas al árbol de decisión.

Lo que hace SKA_DT es identificar el patrón de acuerdo entre los grupos que conduce a la decisión correcta. Si los grupos de abstracción no proporcionaran información útil de manera coherente, no habría ningún patrón que identificar. Por lo tanto, los grupos de abstracción cubren las lagunas de conocimiento en los modelos, lo que les lleva a adoptar líneas de razonamiento coherentes y a dar respuestas correctas de manera sistemática.

Coherente refiere al hecho de que el SKA y la generación de relaciones semánticas (mecanismo de representación), junto con SKA_DT (mecanismo de inferencia), provoca respuestas que generalmente siguen un patrón correcto, dentro de los límites de la información disponible. Este patrón puede mostrar los grupos que se encuentran en el nodo raíz del árbol, como se muestra en la Tabla 6.2. Este nodo es el que mejor divide el corpus durante el entrenamiento. Por ejemplo, a partir de esta tabla, notamos que en el corpus SciTail destaca el grupo G_3, mientras que en el corpus SICK destaca el G_1. También podemos notar que para el modelo `phi3` predomina el grupo G_4, y para el modelo `gemma2`, el G_1. Esto da una idea del patrón que surge en los diferentes modelos y los corpus.

Una consideración sobre la relevancia del uso del árbol de decisión para la toma de decisiones finales en un LLM, es comparar con estrategias recientes basadas en el debate entre múltiples agentes (Q. Wang y cols., 2024). En este sentido, el mecanismo de razonamiento utiliza un árbol de decisión jerárquico explícito para evitar las inconsistencias del debate entre los LLMs, como la propagación de

errores o el sesgo de los jueces, y para garantizar un mayor control y fiabilidad en las respuestas. Por lo tanto, la propuesta SKA se centra en compensar las deficiencias inherentes de los LLM, independientemente de su tamaño: el agente se vuelve más robusto, lo que podría dar lugar a mejores argumentos en escenarios que impliquen un debate entre agentes. Esto permite obtener resultados sólidos sin incurrir en los altos costes asociados al uso de LLM estándar de la industria (más grandes).

Sin embargo, en la Tabla 6.8 se ilustra un área de mejora potencial del SKA_DT, que muestra ejemplos en los que el enfoque propuesto no da la respuesta correcta, mientras que el base sí lo hace. Estos ejemplos se muestran para cada categoría del corpus de diagnóstico. Se puede observar que hay dos situaciones en las que el esquema DT falla: una en la que los grupos mayoritarios dan una opinión incorrecta que se toma en cuenta, y otra en la que los grupos mayoritarios tienen la respuesta correcta, pero no se toma en cuenta.

El primer ejemplo de *LS* ilustra muy bien el primer caso. El único grupo que tiene información relevante coherente con la clases es G_1. La respuesta provocada por los otros grupos, que no contienen información pero están alineados con las otras clases en la tarea, provoca respuestas erróneas, lo que hace que el patrón aprendido dé la respuesta incorrecta. El tercer ejemplo de *Logic* muestra el comportamiento opuesto: hay tres grupos con información que conduce a la respuesta correcta, pero solo uno de ellos (G_4) se tiene en cuenta con una respuesta incorrecta. Ambas situaciones revelan que el SKA_DT sigue mostrando rigidez, siendo vulnerable a los patrones generales sin poder abordar casos particulares.

Los métodos de evaluación actuales para los LLMs evalúan principalmente las capacidades inferenciales a través de métricas basadas en resultados. Sin embargo, esto no permite captar la complejidad de sus procesos de razonamiento subyacentes. En particular, cuando se enfrentan a preguntas sencillas que requieren la aplicación de conocimientos generales, estos modelos suelen tener dificultades para abstraer y utilizar la información relevante, lo que pone de manifiesto una deficiencia significativa en sus capacidades de razonamiento abstracto (Xiong y cols., 2024; Lee y cols., 2025). En este estudio, se define un proceso AoT (Hong y cols., 2024) que comprende los pasos de identificación, alineación e inferencia.

Cuando los LLMs trabajan con relaciones sin procesar, tienden a quedarse atascados en asociaciones literales y no logran captar la profundidad semántica (Kim y cols., 2024). Por el contrario, al estructurar el conocimiento en categorías abstractas que imitan el razonamiento humano, los modelos se benefician de nuevos conceptos o conexiones entre conceptos y evitan utilizar patrones léxicos aprendidos; en otras palabras, proporcionar relaciones novedosas podría evitar la tendencia al sobreajuste o a los atajos basados en ciertas pistas de la premisa o hipótesis.

Del mismo modo, las abstracciones resuelven las ambigüedades al identificar que dos conceptos pueden estar relacionados de múltiples maneras, e incluso detectan contradicciones y neutralidad con mayor precisión (figuras 6.5 y 6.6), áreas en las que los LLM suelen fallar.

Los resultados (Tabla 6.9) demuestran que SKA_DT es un enfoque robusto, ya que equilibra la generalización y la especificidad, minimizando los falsos positivos. Podemos entonces destacar las ventajas de SKA_DT que se enlistan a continuación:

1. Permite controlar el conocimiento externo proporcionado al modelo al caracterizarlo en un nivel más alto de abstracción, ampliando así las relaciones inmediatas capturadas en el grafo de conocimiento.

2. Permite establecer relaciones semánticas entre entidades en P y H, que pueden clasificarse sistemáticamente y vincularse a clases específicas de tareas. Esto permite explorar con mayor precisión las deficiencias lingüísticas y de razonamiento de los modelos.

3. Al definir grupos y vincularlos a la tarea, el enfoque permite centrar la atención del modelo, lo que reduce su campo de inferencia. Esto, a su vez, puede facilitar un análisis más detallado de cómo se configura la atención y cómo se relaciona con el rendimiento del modelo.

4. Incluso cuando las definiciones de los grupos no contienen información específica sobre las entidades P o H, siguen siendo informativas. Guían al modelo para que se alinee con estas definiciones, revelando su conocimiento previo del tema o la facilidad con la que aprovecha la información proporcionada.

Capítulo 8

Conclusiones

El RIT requiere analizar relaciones implícitas/explícitas complejas, lo que exige tanto el conocimiento interno del LLM como el sentido común externo (C. Liu y cols., 2021). Si bien los conocimientos preentrenados de los LLMs son insuficientes para el RIT, los datos externos por sí solos tampoco son suficientes. El enfoque híbrido propuesto dirige la atención hacia las relaciones entre palabras que se pasan por alto, lo que permite a los modelos sintetizar la información que falta.

El marco metodológico SKA tiene como objetivo enriquecer los LLMs con conocimientos relevantes, abstractos y estructurados para mejorar su rendimiento en tareas del RIT, aunque puede aplicarse a otras tareas, por ejemplo, similaridad semántica-textual, paráfrasis, generación de resúmenes, entre otras. La abstracción no solo mejora el rendimiento inmediato, sino que redefine la forma en que los LLMs integran el conocimiento y cómo lo utilizan. Cuando la instrucción introduce información realmente nueva, compensa las lagunas en la red semántica del modelo al especificar a qué relación hay que prestar atención. En consecuencia, el modelo aplica la inferencia solo a las relaciones consideradas esenciales.

La estrategia propuesta para guiar las decisiones de los LLMs mejora la detección de Not-entailment. La eficacia de este enfoque depende de la calidad y la cobertura de la base de conocimiento, ya que las relaciones incompletas, con sesgos culturales o mal definidas pueden limitar la construcción de los grupos abstractos. También existe el problema de la ambigüedad en los límites conceptuales, es decir, algunas relaciones pueden encajar en varias categorías (Ilievski y cols., 2021). A pesar de esto, los resultados demuestran que el marco metodológico de Abstracción de Conocimiento Semántico compensa eficazmente las lagunas de conocimiento en los LLMs.

En la actualidad se observa un cambio de paradigma en el desarrollo de los LLMs, lo que pone de relieve la necesidad de estrategias más inteligentes y sostenibles, como la creación de modelos más pequeños capaces de razonar de forma más eficaz (X. L. Li y cols., 2022; Hoffmann y cols., 2022; Touvron y cols., 2023; Ranaldi y Freitas, 2024; T. Feng y cols., 2024). Las pruebas sugieren que modelos como `phi3` pueden superar a modelos más grandes en tareas de razonamiento. El futuro no reside en ampliar el tamaño de los modelos, sino en crear un andamiaje semántico que imite y permita la flexibilidad del pensamiento humano.

Los *Small Language Models* (SLMs) constituyen el futuro para los agentes de IA, de acuerdo a su superioridad operacional y económica: ya que una porción significativa de las tareas en agentes existentes podría ser gestionada eficazmente por SLMs especializados (Belcak y cols., 2025). Nuestra propuesta permite el desarrollo de agentes más robustos y un razonamiento interpretable, guiando a la IA hacia una comprensión fiable del lenguaje (Gendron y cols., 2024; Z. Li y cols., 2025; Yun y cols., 2025).

Las contribuciones de la presente investigación son las siguientes:

- Presentación del marco metodológico *Semantic Knowledge Abstraction* (SKA) para descubrir y compensar las lagunas de conocimiento semántico de los LLM en el RIT.
- El método permite la representación de conocimiento externo a un nivel más alto de abstracción mediante la definición de categorías de *compatibilidad semántica* e *incompatibilidad* que guían el proceso de razonamiento del LLM y revelan sus inconsistencias y debilidades.
- El marco añade flexibilidad al proceso de razonamiento y toma de decisiones, ya que las relaciones de compatibilidad e incompatibilidad semánticas permiten nuevas conexiones válidas entre conceptos, distintas vías de inferencia y la implementación de estrategias de decisión robustas en el RIT.
- Los resultados muestran que el marco metodológico es particularmente útil para mejorar la inferencia en Not-entailment y sugieren que también podría contribuir al desarrollo de agentes más robustos y fiables en el campo de la CLN.

Propuesta de Marco Metodológico para Modelar la Abstracción Semántica en Inteligencia Artificial: Aplicación a la Tarea del RIT.

David Torres Moreno

5 de marzo de 2026

Referencias

Agrawal, G., Kumarage, T., Alghamdi, Z., y Liu, H. (2024, junio). Can knowledge graphs reduce hallucinations in LLMs? : A survey. En K. Duh, H. Gomez, y S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 3947–3960). Mexico City, Mexico: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.naacl-long.219/ doi: 10.18653/v1/2024.naacl-long.219

Alammar, J. (2018). *The illustrated transformer.* Descargado 2022, de https://jalammar.github.io/illustrated-gpt2

Al-Saeedi, A., y Harma, A. (2025, enero). Emergence of symbolic abstraction heads for in-context learning in large language models. En K. Liu, Y. Song, Z. Han, R. Sifa, S. He, y Y. Long (Eds.), *Proceedings of bridging neurons and symbols for natural language processing and knowledge graphs reasoning @ coling 2025* (pp. 86–96). Abu Dhabi, UAE: ELRA and ICCL. Descargado de https://aclanthology.org/2025.neusymbridge-1.9/

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020, 6). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82-115. doi: 10.1016/j.inffus.2019.12.012

Bahdanau, D., Cho, K., y Bengio, Y. (2014, 9). Neural machine translation by jointly learning to align and translate. *Published as a conference paper at ICLR 2015*. Descargado de http://arxiv.org/abs/1409.0473

Baidu-ERNIE-Team. (2025). *Ernie 4.5 technical report.* https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf.

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., ... Molchanov, P. (2025). *Small language models are the future of agentic ai.* Descargado de https://arxiv.org/abs/2506.02153

Beloucif, M., y Biemann, C. (2021, noviembre). Probing pre-trained language models for semantic attributes and their values. En M.-F. Moens, X. Huang, L. Specia, y S. W.-t. Yih (Eds.), *Findings of the association for computational linguistics: Emnlp 2021* (pp. 2554–2559). Punta Cana, Dominican Republic: Association for Computational Linguistics. Descargado de https://aclanthology.org/2021.findings-emnlp.218/ doi: 10.18653/v1/2021.findings-emnlp.218

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., Ca, J. U., Kandola, J., ... Shawe-Taylor, J. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137-1155.

Bovi, C. D., Telesca, L., y Navigli, R. (2015). Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*. Descargado de http://lcl.uniroma1.it/defie

Bowman, S. R., Angeli, G., Potts, C., y Manning, C. D. (2015, 8). A large annotated corpus for learning natural language inference. *arxiv*. Descargado de http://arxiv.org/abs/1508.05326

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020, 5). Language models are few-shot learners. *arxib*. Descargado de http://arxiv.org/abs/2005.14165

Cao, Z., Yamada, H., Teufel, S., y Tokunaga, T. (2025). A comprehensive evaluation of semantic relation knowledge of pretrained language models and humans. *Lang Resources & Evaluation*. doi: https://doi.org/10.1007/s10579-025-09858-9

Chen, L., Davis, J. Q., Hanin, B., Bailis, P., Stoica, I., Zaharia, M., y Zou, J. (2024, 3). Are more llm calls all you need? towards scaling laws of compound inference systems. *Arxiv*. Descargado de http://arxiv.org/abs/2403.02419

Chen, Q., Zhu, X., Ling, Z.-H., Inkpen, D., y Wei, S. (2018, julio). Neural natural language inference models enhanced with external knowledge. En I. Gurevych y Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2406–2417). Melbourne, Australia: Association for Computational Linguistics. Descargado de https://aclanthology.org/P18-1224/ doi: 10.18653/v1/P18-1224

Chen, W., Wang, W., Chu, Z., Ren, K., Zheng, Z., y Lu, Z. (2024, agosto). Self-para-consistency: Improving reasoning tasks at low cost for large language models. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 14162–14167). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.findings-acl.842/ doi: 10.18653/v1/2024.findings-acl.842

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... Fiedel, N. (2022).

Palm: Scaling language modeling with pathways. *ArXiv*, *abs/2204.02311*. Descargado de https://api.semanticscholar.org/CorpusID:247951931

Cruse, A. (2004). *Meaning in language: An introduction to semantics and pragmatics*. Oxford: Oxford University Press UK.

Dagan, I., Glickman, O., y Magnini, B. (2006). The pascal recognising textual entailment challenge. *In Quionero-Candela et al., editor, LNAI 3944: MLCW2005*.

Dai, X., Hua, Y., Wu, T., Sheng, Y., Ji, Q., y Qi, G. (2025). Large language models can better understand knowledge graphs than we thought. *Knowledge-Based Systems*, *312*, 113060. Descargado de https://www.sciencedirect.com/science/article/pii/S0950705125001078 doi: https://doi.org/10.1016/j.knosys.2025.113060

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2018, 10). Bert: Pre-training of deep bidirectional transformers for language understanding. *arxib*. Descargado de http://arxiv.org/abs/1810.04805

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., y Weston, J. (2024, agosto). Chain-of-verification reduces hallucination in large language models. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 3563–3578). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.findings-acl.212/ doi: 10.18653/v1/2024.findings-acl.212

Dogan, A., y Birant, D. (2019a). A weighted majority voting ensemble approach for classification. *2019 4th International Conference on Computer Science and Engineering (UBMK)*, 1-6. Descargado de https://api.semanticscholar.org/CorpusID:208207259

Dogan, A., y Birant, D. (2019b). *A weighted majority voting ensemble approach for classification.*

Dutt, R., Ray Choudhury, S., Rao, V. V., Rose, C., y Vydiswaran, V. (2024, noviembre). Investigating the generalizability of pretrained language models across multiple dimensions: A case study of NLI and MRC. En D. Hupkes y cols. (Eds.), *Proceedings of the 2nd genbench workshop on generalisation (benchmarking) in nlp* (pp. 165–182). Miami, Florida, USA: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.genbench-1.11/ doi: 10.18653/v1/2024.genbench-1.11

Echterhoff, J. M., Liu, Y., Alessa, A., McAuley, J., y He, Z. (2024, noviembre). Cognitive bias in decision-making with LLMs. En Y. Al-Onaizan, M. Bansal, y Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 12640–12653). Miami, Florida,

USA: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.findings-emnlp.739/ doi: 10.18653/v1/2024.findings-emnlp.739

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., y Goldberg, Y. (2021). Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, *9*, 1012–1031. Descargado de https://aclanthology.org/2021.tacl-1.60/ doi: 10.1162/tacl_a_00410

Feng, S., Shi, W., Wang, Y., Ding, W., Balachandran, V., y Tsvetkov, Y. (2024, agosto). Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 14664–14690). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.acl-long.786/ doi: 10.18653/v1/2024.acl-long.786

Feng, T., Li, Y., Chenglin, L., Chen, H., Yu, F., y Zhang, Y. (2024, noviembre). Teaching small language models reasoning through counterfactual distillation. En Y. Al-Onaizan, M. Bansal, y Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 5831–5842). Miami, Florida, USA: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.emnlp-main.333/ doi: 10.18653/v1/2024.emnlp-main.333

Fu, Y., Feng, Y., y Cunningham, J. P. (2020, 1). Paraphrase generation with latent bag of words. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.*. Descargado de http://arxiv.org/abs/2001.01941

Gendron, G., Bao, Q., Witbrock, M., y Dobbie, G. (2024, 8). Large language models are not strong abstract reasoners. En K. Larson (Ed.), *Proceedings of the thirty-third international joint conference on artificial intelligence, IJCAI-24* (pp. 6270–6278). International Joint Conferences on Artificial Intelligence Organization. Descargado de https://doi.org/10.24963/ijcai.2024/693 (Main Track) doi: 10.24963/ijcai.2024/693

Guo, M., Chen, Y., Xu, J., y Zhang, Y. (2022). Dynamic knowledge integration for natural language inference. En *2022 4th international conference on natural language processing (icnlp)* (p. 360-364). doi: 10.1109/ICNLP55136.2022.00066

Haji, F., Bethany, M., Tabar, M., Chiang, J., Rios, A., y Najafirad, P. (2024, 9). Improving llm reasoning with multi-agent tree-of-thought validator agent. *Arxiv*. Descargado de http://arxiv.org/abs/2409.11527

He, P., Liu, X., Gao, J., y Chen, W. (2020, 6). Deberta: Decoding-enhanced bert with disentangled

attention. *Axib*. Descargado de http://arxiv.org/abs/2006.03654

He, Q., Wang, Y., Yu, J., y Wang, W. (2025, enero). Language models over large-scale knowledge base: on capacity, flexibility and reasoning for new facts. En O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, y S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 1736–1753). Abu Dhabi, UAE: Association for Computational Linguistics. Descargado de https://aclanthology.org/2025.coling-main.118/

Hendel, R., Geva, M., y Globerson, A. (2023, diciembre). In-context learning creates task vectors. En H. Bouamor, J. Pino, y K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 9318–9333). Singapore: Association for Computational Linguistics. Descargado de https://aclanthology.org/2023.findings-emnlp.624/ doi: 10.18653/v1/2023.findings-emnlp.624

Hochreiter, S., y Schmidhuber, J. (1997, 11). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. Descargado de https://doi.org/10.1162/neco.1997.9.8.1735 doi: 10.1162/neco.1997.9.8.1735

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... Sifre, L. (2022). *Training compute-optimal large language models.* Descargado de https://arxiv.org/abs/2203.15556

Hong, R., Zhang, H., Pan, X., Yu, D., y Zhang, C. (2024, noviembre). Abstraction-of-thought makes language models better reasoners. En Y. Al-Onaizan, M. Bansal, y Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 1993–2027). Miami, Florida, USA: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.findings-emnlp.110/ doi: 10.18653/v1/2024.findings-emnlp.110

Huang, B., Lu, S., Wan, X., y Duan, N. (2024, agosto). Enhancing large language models in coding through multi-perspective self-consistency. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1429–1450). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.acl-long.78/ doi: 10.18653/v1/2024.acl-long.78

Huber, T., y Niklaus, C. (2025, enero). LLMs meet bloom's taxonomy: A cognitive view on large language model evaluations. En O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, y S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 5211–5246). Abu Dhabi, UAE: Association for Computational Linguistics. Descargado de https://aclanthology.org/2025.coling-main.350/

Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., ... Jin, Z. (2023, 10). A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, *5*, 1161-1174. doi: 10.1038/s42256-023-00729-y

Hurford, J. R., Heasley, B., y Smith, M. B. (2007). *Semantics: A coursebook* (2.ª ed.). Cambridge University Press.

Ilievski, F., Oltramari, A., Ma, K., Zhang, B., McGuinness, D. L., y Szekely, P. (2021). Dimensions of commonsense knowledge. *Knowledge-Based Systems*, *229*, 107347. Descargado de https://www.sciencedirect.com/science/article/pii/S0950705121006092 doi: https://doi.org/10.1016/j.knosys.2021.107347

Jeffries, L. (1998). *Meaning in english: An introduction to language study.* St. Martin's Press. Descargado de https://books.google.com.mx/books?id=WrLDQgAACAAJ

Jiang, S., Li, B., Liu, C., y Yu, D. (2019). Knowledge augmented inference network for natural language inference. En J. Zhao, F. v. Harmelen, J. Tang, X. Han, Q. Wang, y X. Li (Eds.), *Knowledge graph and semantic computing. knowledge computing and language understanding* (pp. 129–135). Singapore: Springer Singapore.

Jin, D., Jin, Z., Zhou, J., y Szolovits, P. (2020, 04). Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 8018-8025. doi: 10.1609/aaai.v34i05.6311

Kapanipathi, P., Thost, V., Sankalp Patel, S., Whitehead, S., Abdelaziz, I., Balakrishnan, A., ... Fokoue, A. (2020, Apr.). Infusing knowledge into the textual entailment task using graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8074-8081. Descargado de https://ojs.aaai.org/index.php/AAAI/article/view/6318 doi: 10.1609/aaai.v34i05.6318

Kasner, Z., Konstas, I., y Dusek, O. (2023, mayo). Mind the labels: Describing relations in knowledge graphs with pretrained models. En A. Vlachos y I. Augenstein (Eds.), *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (pp. 2398–2415). Dubrovnik, Croatia: Association for Computational Linguistics. Descargado de https://aclanthology.org/2023.eacl-main.176/ doi: 10.18653/v1/2023.eacl-main.176

Khot, T., Sabharwal, A., y Clark, P. (2018). Scitail: A textual entailment dataset from science question answering. En *Aaai conference on artificial intelligence.* Descargado de https://api.semanticscholar.org/CorpusID:24462950

Kim, S., Jeong, S., y Kim, H. (2024). Bridge to better understanding: Syntax extension with virtual linking-phrase for natural language inference. *Knowledge-Based Systems*, *305*, 112608. Descargado de https://www.sciencedirect.com/science/article/pii/S0950705124012425 doi: https://doi.org/10.1016/j.knosys.2024.112608

Kouylekov, M., y Magnini, B. (2006). Recognizing textual entailment with tree edit distance algorithms. *arxiv*. Descargado de http://www.cs.ualberta.ca/

Lake, B. M., y Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological review*. doi: https://doi.org/10.1037/rev0000297

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., y Müller, K. R. (2019, 12). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, *10*. doi: 10.1038/s41467-019-08987-4

Lauscher, A., Majewska, O., Ribeiro, L. F. R., Gurevych, I., Rozanov, N., y Glavaš, G. (2020, noviembre). Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. En E. Agirre, M. Apidianaki, y I. Vulić (Eds.), *Proceedings of deep learning inside out (deelio): The first workshop on knowledge extraction and integration for deep learning architectures* (pp. 43–49). Online: Association for Computational Linguistics. Descargado de https://aclanthology.org/2020.deelio-1.5/ doi: 10.18653/v1/2020.deelio-1.5

Lederman, H., y Mahowald, K. (2024, 09). Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of llms. *Transactions of the Association for Computational Linguistics*, *12*, 1087-1103. Descargado de https://doi.org/10.1162/tacl_a_00690 doi: 10.1162/tacl_a_00690

Lee, S., Sim, W., Shin, D., Seo, W., Park, J., Lee, S., ... Kim, S. (2025, enero). Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Trans. Intell. Syst. Technol.*. Descargado de https://doi.org/10.1145/3712701 (Just Accepted) doi: 10.1145/3712701

Leidinger, A., Van Rooij, R., y Shutova, E. (2024, agosto). Are LLMs classical or nonmonotonic reasoners? lessons from generics. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 558–573). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.acl-short.51/ doi: 10.18653/v1/2024.acl-short.51

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... others (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. En *Neurips 2020.*

Li, C., y Flanigan, J. (2024, Mar.). Task contamination: Language models may not be few-shot anymore. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 18471-18480. Descargado de https://ojs.aaai.org/index.php/AAAI/article/view/29808 doi: 10.1609/aaai.v38i16.29808

Li, J., Cheng, X., Zhao, X., Nie, J.-Y., y Wen, J.-R. (2023, diciembre). HaluEval: A large-scale hallucination evaluation benchmark for large language models. En H. Bouamor, J. Pino, y K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 6449–6464). Singapore: Association for Computational Linguistics. Descargado de https://aclanthology.org/2023.emnlp-main.397/ doi: 10.18653/v1/2023.emnlp-main.397

Li, X. L., Kuncoro, A., Hoffmann, J., de Masson d'Autume, C., Blunsom, P., y Nematzadeh, A. (2022, diciembre). A systematic investigation of commonsense knowledge in large language models. En Y. Goldberg, Z. Kozareva, y Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 11838–11855). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Descargado de https://aclanthology.org/2022.emnlp-main.812/ doi: 10.18653/v1/2022.emnlp-main.812

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.-G., y Chen, W. (2023, julio). Making language models better reasoners with step-aware verifier. En A. Rogers, J. Boyd-Graber, y N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5315–5333). Toronto, Canada: Association for Computational Linguistics. Descargado de https://aclanthology.org/2023.acl-long.291/ doi: 10.18653/v1/2023.acl-long.291

Li, Z., Chen, C., Li, M., y Liao, B. (2025). Exploring formal defeasible reasoning of large language models: A chain-of-thought approach. *Knowledge-Based Systems*, *319*, 113564. Descargado de https://www.sciencedirect.com/science/article/pii/S0950705125006100 doi: https://doi.org/10.1016/j.knosys.2025.113564

Lin, L., Fu, J., Liu, P., Li, Q., Gong, Y., Wan, J., ... Gai, K. (2024, agosto). Just ask one more time! self-agreement improves reasoning of language models in (almost) all scenarios. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 3829–3852). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.findings-acl.230/ doi: 10.18653/v1/2024.findings-acl.230

Liu, C., Cohn, T., y Frermann, L. (2021, noviembre). Commonsense knowledge in word associations and ConceptNet. En A. Bisazza y O. Abend (Eds.), *Proceedings of the 25th conference on computational natural language learning* (pp. 481–495). Online: Association for Computational Linguistics. Descargado de https://aclanthology.org/2021.conll-1.38/ doi: 10.18653/v1/2021.conll-1.38

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., y Wang, P. (2020, Apr.). K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(03), 2901-2908. Descargado de https://ojs.aaai.org/index.php/AAAI/article/view/5681 doi: 10.1609/aaai.v34i03.5681

Liu, Z., Li, D., Hu, X., Zhao, X., Chen, Y., Hu, B., y Zhang, M. (2024, noviembre). Take off the training wheels! progressive in-context learning for effective alignment. En Y. Al-Onaizan, M. Bansal, y Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 2743–2757). Miami, Florida, USA: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.emnlp-main.160/ doi: 10.18653/v1/2024.emnlp-main.160

Lundberg, S. M., Allen, P. G., y Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Arxiv*. Descargado de https://github.com/slundberg/shap

Lyons, J. (1977). *Semantics*. Cambridge University Press.

MacCartney, B., y Manning, C. D. (2007, junio). Natural logic for textual inference. En S. Sekine, K. Inui, I. Dagan, B. Dolan, D. Giampiccolo, y B. Magnini (Eds.), *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing* (pp. 193–200). Prague: Association for Computational Linguistics. Descargado de https://aclanthology.org/W07-1431/

MacCartney, B., y Manning, C. D. (2009, enero). An extended model of natural logic. En H. Bunt (Ed.), *Proceedings of the eight international conference on computational semantics* (pp. 140–156). Tilburg, The Netherlands: Association for Computational Linguistics. Descargado de https://aclanthology.org/W09-3714/

Madaan, L., Esiobu, D., Stenetorp, P., Plank, B., y Hupkes, D. (2024). *Lost in inference: Rediscovering the role of natural language inference for large language models.* Descargado de https://arxiv.org/abs/2411.14103

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., y Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517-540. Descargado de https://www.sciencedirect.com/science/article/pii/S1364661324000275 doi: https://doi.org/10.1016/j.tics.2024.01.011

Manakul, P., Liusie, A., y Gales, M. (2023, diciembre). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. En H. Bouamor, J. Pino, y K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 9004–9017). Singapore: Association for Computational Linguistics. Descargado de https://aclanthology.org/2023.emnlp-main.557/ doi: 10.18653/v1/2023.emnlp-main.557

Marcus, G., thank Christina, I., Chollet, F., Davis, E., Lipton, Z., Pacifico, S., ... Vouloumanos, A. (2017). Deep learning: A critical appraisal. *arxiv*. Descargado de http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., y Zamparelli, R. (2014, mayo). A SICK cure for the evaluation of compositional distributional semantic models. En N. Calzolari y cols. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 216–223). Reykjavik, Iceland: European Language Resources Association (ELRA). Descargado de https://aclanthology.org/L14-1314/

McCoy, T., Pavlick, E., y Linzen, T. (2019, julio). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. En A. Korhonen, D. Traum, y L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3428–3448). Florence, Italy: Association for Computational Linguistics. Descargado de https://aclanthology.org/P19-1334/ doi: 10.18653/v1/P19-1334

Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013, 1). Efficient estimation of word representations in vector space. *ResearchGate*. Descargado de http://arxiv.org/abs/1301.3781

Mu, L., Zhang, W., Zhang, Y., y Jin, P. (2024, agosto). DDPrompt: Differential diversity prompting in large language models. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 168–174). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.acl-short.17/ doi: 10.18653/v1/2024.acl-short.17

Nangia, N., y Bowman, S. R. (2019, julio). Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. En A. Korhonen, D. Traum, y L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4566–4575). Florence, Italy: Association for Computational Linguistics. Descargado de https://aclanthology.org/P19-1449 doi: 10.18653/v1/P19-1449

Nie, Y., Zhou, X., y Bansal, M. (2020a, 10). What can we learn from collective human opinions on natural language inference data? *Arxiv*. Descargado de http://arxiv.org/abs/2010.03532

Nie, Y., Zhou, X., y Bansal, M. (2020b, noviembre). What can we learn from collective human opinions on natural language inference data? En B. Webber, T. Cohn, Y. He, y Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 9131–9143). Online: Association for Computational Linguistics. Descargado de https://aclanthology.org/2020.emnlp-main.734/ doi: 10.18653/v1/2020.emnlp-main.734

Niu, T., Yavuz, S., Zhou, Y., Keskar, N. S., Wang, H., y Xiong, C. (2020, 10). Unsupervised paraphrasing with pretrained language models. *SemanticSchoolar*. Descargado de http://arxiv.org/abs/2010.12885

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., ... Zoph, B. (2024). *Gpt-4 technical report.* Descargado de https://arxiv.org/abs/2303.08774

Ormazabal, A., Artetxe, M., Soroa, A., Labaka, G., y Agirre, E. (2022). Principled paraphrase generation with parallel corpora. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*, 1621-1638. Descargado de https://github.com/aitorormazabal/

Parikh, A. P., Täckström, O., Das, D., y Uszkoreit, J. (2016, 6). A decomposable attention model for natural language inference. *Computer Science*. Descargado de http://arxiv.org/abs/1606.01933

Peng, H., Wang, X., Hu, S., Jin, H., Hou, L., Li, J., ... Liu, Q. (2022, diciembre). COPEN: Probing conceptual knowledge in pre-trained language models. En Y. Goldberg, Z. Kozareva, y Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 5015–5035). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Descargado de https://aclanthology.org/2022.emnlp-main.335/ doi: 10.18653/v1/2022.emnlp-main.335

Poliak, A. (2020). A survey on recognizing textual entailment as an nlp evaluation. *arxiv*.

Poliak, A., Haldar, A., Rudinger, R., Hu, J. E., Pavlick, E., White, A. S., y Van Durme, B. (2018, octubre-noviembre). Collecting diverse natural language inference problems for sentence representation evaluation. En E. Riloff, D. Chiang, J. Hockenmaier, y J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 67–81). Brussels, Belgium: Association for Computational Linguistics. Descargado de https://aclanthology.org/D18-1007/ doi: 10.18653/v1/D18-1007

Proebsting, G., y Poliak, A. (2025, enero). Biases in large language model-elicited text: A case study in natural language inference. En O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, y S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 5836–5851). Abu Dhabi, UAE: Association for Computational Linguistics. Descargado de https://aclanthology.org/2025.coling-main.389/

Qi, J., Fernández, R., y Bisazza, A. (2023, diciembre). Cross-lingual consistency of factual knowledge in multilingual language models. En H. Bouamor, J. Pino, y K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 10650–10666). Sin-

gapore: Association for Computational Linguistics. Descargado de https://aclanthology.org/2023.emnlp-main.658/ doi: 10.18653/v1/2023.emnlp-main.658

Qian, L., Qiu, L., Zhang, W., Jiang, X., y Yu, Y. (2019). Exploring diverse expressions for paraphrase generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processingand the 9th International Joint Conference on Natural Language Processing,*, 3173-3182.

Radford, A., y Narasimhan, K. (2018). Improving language understanding by generative pre-training.. Descargado de https://api.semanticscholar.org/CorpusID:49313245

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., y Sutskever, I. (2019). Language models are unsupervised multitask learners. *Arxiv*. Descargado de https://github.com/codelucas/newspaper

Ranaldi, L., y Freitas, A. (2024, marzo). Aligning large and small language models via chain-of-thought reasoning. En Y. Graham y M. Purver (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 1: Long papers)* (pp. 1812–1827). St. Julian's, Malta: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.eacl-long.109/ doi: 10.18653/v1/2024.eacl-long.109

Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. T. I., Chadha, A., ... Das, A. (2023, diciembre). The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. En H. Bouamor, J. Pino, y K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 2541–2573). Singapore: Association for Computational Linguistics. Descargado de https://aclanthology.org/2023.emnlp-main.155/ doi: 10.18653/v1/2023.emnlp-main.155

Regneri, M., Abdelhalim, A., y Laue, S. (2024, mayo). Detecting conceptual abstraction in LLMs. En N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, y N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 4697–4704). Torino, Italia: ELRA and ICCL. Descargado de https://aclanthology.org/2024.lrec-main.420/

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., y Blunsom, P. (2015, 9). Reasoning about entailment with neural attention. *conference paper at ICLR 2016*. Descargado de http://arxiv.org/abs/1509.06664

Ríos, M., Gelbukh, A., y Bandyopadhyay, S. (2010a). Recognizing textual entailment using a machine learning approach. *ResearchGate*.

Ríos, M., Gelbukh, A., y Bandyopadhyay, S. (2010b). Recognizing textual entailment with statistical

methods. En (Vol. 6256 LNCS, p. 372-381). doi: 10.1007/978-3-642-15992-3_39

Sahu, P., Cogswell, M., Gong, Y., y Divakaran, A. (2022). *Unpacking large language models with conceptual consistency.* Descargado de https://arxiv.org/abs/2209.15093

Sanyal, S., Singh, H., y Ren, X. (2022, mayo). FaiRR: Faithful and robust deductive reasoning over natural language. En S. Muresan, P. Nakov, y A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1075–1093). Dublin, Ireland: Association for Computational Linguistics. Descargado de https://aclanthology.org/2022.acl-long.77/ doi: 10.18653/v1/2022.acl-long.77

Sedova, A., Litschko, R., Frassinelli, D., Roth, B., y Plank, B. (2024, noviembre). To know or not to know? analyzing self-consistency of large language models under ambiguity. En Y. Al-Onaizan, M. Bansal, y Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 17203–17217). Miami, Florida, USA: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.findings-emnlp.1003/ doi: 10.18653/v1/2024.findings-emnlp.1003

Shajalal, M., Atabuzzaman, M., Baby, M. B., Karim, M. R., y Boden, A. (2023). Textual entailment recognition with semantic features from empirical text representation. En A. K. M y cols. (Eds.), *Speech and language technologies for low-resource languages* (pp. 183–195). Cham: Springer International Publishing.

Shi, J., Ding, X., Xiong, K., Zhao, H., Qin, B., y Liu, T. (2025, julio). Natural logic at the core: Dynamic rewards for entailment tree generation. En W. Che, J. Nabende, E. Shutova, y M. T. Pilehvar (Eds.), *Findings of the association for computational linguistics: Acl 2025* (pp. 17372–17382). Vienna, Austria: Association for Computational Linguistics. Descargado de https://aclanthology.org/2025.findings-acl.893/ doi: 10.18653/v1/2025.findings-acl.893

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., y Farajtabar, M. (2025). *The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity.* Descargado de https://arxiv.org/abs/2506.06941

Silva, V. S., Freitas, A., y Handschuh, S. (2020, 9). Xte: Explainable text entailment. *Artificial Intelligence*. Descargado de http://arxiv.org/abs/2009.12431

Speer, R., Chin, J., y Havasi, C. (2016). Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR, abs/1612.03975*. Descargado de http://arxiv.org/abs/1612.03975

Speer, R., y Lowry-Duda, J. (2017). Conceptnet at semeval-2017 task 2: Extending word embeddings

with multilingual relational knowledge. *CoRR, abs/1704.03560*. Descargado de http://arxiv.org/abs/1704.03560

Torres-Moreno, D., y Hermosillo-Valadez, J. (2026). Semantic knowledge abstraction: Consistent reasoning in large language models for natural language inference. *Knowledge-Based Systems*, *332*, 114825. Descargado de https://www.sciencedirect.com/science/article/pii/S0950705125018635 doi: https://doi.org/10.1016/j.knosys.2025.114825

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Lample, G. (2023). *Llama: Open and efficient foundation language models.* Descargado de https://arxiv.org/abs/2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017, 6). Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS) 2017*. Descargado de http://arxiv.org/abs/1706.03762

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... Bowman, S. R. (2019a). Superglue: A stickier benchmark for general-purpose language understanding systems. *ArXiv*, *abs/1905.00537*. Descargado de https://api.semanticscholar.org/CorpusID:143424870

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... Bowman, S. (2019b, 05). *Superglue: A stickier benchmark for general-purpose language understanding systems.* doi: 10.48550/arXiv.1905.00537

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., y Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. En *International conference on learning representations*. Descargado de https://openreview.net/forum?id=rJ4km2R5t7

Wang, H., Prasad, A., Stengel-Eskin, E., y Bansal, M. (2024, agosto). Soft self-consistency improves language models agents. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 287–301). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.acl-short.28/ doi: 10.18653/v1/2024.acl-short.28

Wang, Q., Wang, Z., Su, Y., Tong, H., y Song, Y. (2024, agosto). Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 6106–6131). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.acl-long.331/ doi: 10.18653/v1/2024.acl-long.331

Wang, S., Wei, Z., Choi, Y., y Ren, X. (2024, agosto). Can LLMs reason with rules? logic scaffolding for stress-testing and improving LLMs. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 7523–7543). Bangkok, Thailand: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.acl-long.406/ doi: 10.18653/v1/2024.acl-long.406

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., y Tang, J. (2021, 03). Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, *9*, 176-194. Descargado de https://doi.org/10.1162/tacl_a_00360 doi: 10.1162/tacl_a_00360

Wang, X., Kapanipathi, P., Musa, R., Yu, M., Talamadupula, K., Abdelaziz, I., ... Witbrock, M. (2018). Improving natural language inference using external knowledge in the science questions domain. En *Aaai conference on artificial intelligence.* Descargado de https://api.semanticscholar.org/CorpusID:52291548

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... Zhou, D. (2023). *Self-consistency improves chain of thought reasoning in language models.* Descargado de https://arxiv.org/abs/2203.11171

Wang, Y., Gangi Reddy, R., Mujahid, Z. M., Arora, A., Rubashevskii, A., Geng, J., ... Nakov, P. (2024, noviembre). Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. En Y. Al-Onaizan, M. Bansal, y Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 14199–14230). Miami, Florida, USA: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.findings-emnlp.830/ doi: 10.18653/v1/2024.findings-emnlp.830

Wang, Z., Shi, H., Wang, W., Fang, T., Zhang, H., Choi, S., ... Song, Y. (2024, junio). AbsPyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. En K. Duh, H. Gomez, y S. Bethard (Eds.), *Findings of the association for computational linguistics: Naacl 2024* (pp. 3991–4010). Mexico City, Mexico: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.findings-naacl.252/ doi: 10.18653/v1/2024.findings-naacl.252

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. En *Proceedings of the 36th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Westera, M., y Boleda, G. (2019). Don't blame distributional semantics if it can't do entailment. *ArXiv*, *abs/1905.07356*. Descargado de https://api.semanticscholar.org/CorpusID:92994351

Williams, A., Nangia, N., y Bowman, S. (2018, junio). A broad-coverage challenge corpus for sentence understanding through inference. En M. Walker, H. Ji, y 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). New Orleans, Louisiana: Association for Computational Linguistics. Descargado de https://aclanthology.org/N18-1101/ doi: 10.18653/v1/N18-1101

Xiong, K., Ding, X., Liu, T., Qin, B., Xu, D., Yang, Q., ... Cao, Y. (2024). Meaningful learning: Enhancing abstract reasoning in large language models via generic fact guidance. En A. Globerson y cols. (Eds.), *Advances in neural information processing systems* (Vol. 37, pp. 120501–120525). Curran Associates, Inc. Descargado de https://proceedings.neurips.cc/paper_files/paper/2024/file/da5498f88193ff61f0daea1940b819da-Paper-Conference.pdf

Xiu, Y., Xiao, Z., y Liu, Y. (2022, diciembre). LogicNMR: Probing the non-monotonic reasoning ability of pre-trained language models. En Y. Goldberg, Z. Kozareva, y Y. Zhang (Eds.), *Findings of the association for computational linguistics: Emnlp 2022* (pp. 3616–3626). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Descargado de https://aclanthology.org/2022.findings-emnlp.265/ doi: 10.18653/v1/2022.findings-emnlp.265

Xue, M., Liu, D., Lei, W., Ren, X., Yang, B., Xie, J., ... Lv, J. (2023, diciembre). Dynamic voting for efficient reasoning in large language models. En H. Bouamor, J. Pino, y K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 3085–3104). Singapore: Association for Computational Linguistics. Descargado de https://aclanthology.org/2023.findings-emnlp.203/ doi: 10.18653/v1/2023.findings-emnlp.203

Xue, M., Liu, D., Lei, W., Xingzhang, Baosong, R., Xie, Y. J., ... Lv, J. (2023). *Dynamic voting for efficient reasoning in large language models.*

Yang, D., Li, N., Zou, L., y Ma, H. (2022). Lexical semantics enhanced neural word embeddings. *Knowledge-Based Systems*, *252*, 109298. Descargado de https://www.sciencedirect.com/science/article/pii/S0950705122006517 doi: https://doi.org/10.1016/j.knosys.2022.109298

Yang, K., Deng, J., y Chen, D. (2022, diciembre). Generating natural language proofs with verifier-guided search. En Y. Goldberg, Z. Kozareva, y Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 89–105). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Descargado de https://aclanthology.org/

2022.emnlp-main.7/ doi: 10.18653/v1/2022.emnlp-main.7

Yang, X., Zhu, X., Zhao, H., Zhang, Q., y Feng, Y. (2019, 04). Enhancing unsupervised pretraining with external knowledge for natural language inference. En (p. 413-419). doi: 10.1007/978-3-030-18305-9_38

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., y Le, Q. V. (2019, 6). Xlnet: Generalized autoregressive pretraining for language understanding. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*. Descargado de http://arxiv.org/abs/1906.08237

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., y Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, *abs/2305.10601*. Descargado de https://api.semanticscholar.org/CorpusID:258762525

Yun, T., Sun, C., y Pavlick, E. (2025, julio). What is an "abstract reasoner"? revisiting experiments and arguments about large language models. En G. Boleda y M. Roth (Eds.), *Proceedings of the 29th conference on computational natural language learning* (pp. 156–168). Vienna, Austria: Association for Computational Linguistics. Descargado de https://aclanthology.org/2025.conll-1.11/ doi: 10.18653/v1/2025.conll-1.11

Zanzotto, F. M., y Pennacchiotti, M. (2010). Expanding textual entailment corpora from wikipedia using co-training. *Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources*, 28-36. Descargado de http://mturk.com

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J.-R. (2025). *A survey of large language models.* Descargado de https://arxiv.org/abs/2303.18223

Zhong, Q., Ding, L., Zhan, Y., Qiao, Y., Wen, Y., Shen, L., ... Tao, D. (2022). *Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue.* Descargado de https://arxiv.org/abs/2212.01853

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., ... Duan, N. (2024, junio). AGIEval: A human-centric benchmark for evaluating foundation models. En K. Duh, H. Gomez, y S. Bethard (Eds.), *Findings of the association for computational linguistics: Naacl 2024* (pp. 2299–2314). Mexico City, Mexico: Association for Computational Linguistics. Descargado de https://aclanthology.org/2024.findings-naacl.149/ doi: 10.18653/v1/2024.findings-naacl.149

Zhou, J., y Bhat, S. (2021). Paraphrase generation: A survey of the state of the art. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,*, 5075-5086.

Zi, Y., Roy, K., Narayanan, V., Gaur, M., y Sheth, A. (2023a, 06). Ierl: Interpretable ensemble representation learning – combining crowdsourced knowledge and distributed semantic representations. *ArXiv*. doi: 10.48550/arXiv.2306.13865

Zi, Y., Roy, K., Narayanan, V., Gaur, M., y Sheth, A. (2023b, 6). Ierl: Interpretable ensemble representation learning – combining crowdsourced knowledge and distributed semantic representations. *Arxiv*. Descargado de http://arxiv.org/abs/2306.13865

Cuernavaca, Mor., a 26 de Febrero de 2026

DRA. LINA ANDREA RIVILLAS ACEVEDO
COORDINADORA DEL POSGRADO EN CIENCIAS
P R E S E N T E

Atendiendo a la solicitud para emitir DICTAMEN sobre la revisión de la tesis titulada: “Propuesta de Marco Metodológico para Modelar la Abstracción Semántica en Inteligencia Artificial: Aplicación a la Tarea del RIT” que presenta el M.C.C. David Torres Moreno (10025502) para obtener el título de Doctor en Ciencias.

Director de tesis: Dr. Jorge Hermosillo Valadez
Unidad Académica: Instituto de Investigación en Ciencias Básicas y Aplicadas (IICBA)

Nos permitimos informarle que nuestro voto es:

NOMBRE	DICTAMEN	FIRMA
Dr. Markus Franciskus Müller Bender CInC – UAEM	APROBADO	
Dr. Bruno Lara Guzmán CInC – UAEM	APROBADO	
Dra. María Asela Reig Alamillo CIIHu – UAEM	APROBADO	
Dr. José Daniel Arzate Mena CInC – UAEM	APROBADO	
Dr. Manuel Montes y Gómez INAOE	APROBADO	
Dr. Noé Alejandro Castro Sánchez CENIDET	APROBADO	
Dra. Marta Lilia Eraña Díaz CINCCO - UAEM	APROBADO	

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Se expide el presente documento con firma electrónica UAEM, soportada por el certificado vigente a la fecha de su elaboración y con efectos plenos de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS PUBLICADOS en el ÓRGANÓ INFORMATIVO UNIVERSITARIO "ADOLFO MENÉNDEZ SAMARÁ" número 117 de fecha 20 de abril de 2021.

Sello electrónico

MARKUS FRANCISKUS MULLER BENDER | Fecha:2026-02-26 15:48:32 | FIRMANTE
u5surlMhDdSVXakClYdQOL91JVqQMDBzznEu5mBKni09tRA6iVK5A003rrEl6aH8TkXyYmvH84jRRuL3qWc39RlqubIOjSS/jmneD09tYIF2CngeKsrc552IWlhlxdiVT3iPbe9RDUhqHBCOzEIMZ6DU+9YBQP4dMtU19TJlfpMDxe5WjQXUtEUjgjiNp4EjN2Vt7nzIXOMUDDFLQvVkm/UF+WTk0PlF0zE4em+E0qzW28zvBfjoZbQSJlI1nQbPPAxKIqLurdkObxX2oU2uuImeMV7wNYy8P0I7WjDvs77icTcjx60UF4I7f2YGIw9Kqh+VO8jzb6gfN75gixeNmA==

MARTA LILIA ERAÑA DIAZ | Fecha:2026-02-26 17:01:06 | FIRMANTE
w/PfEpVkAR2TKmgJEQhSeKWL6itBQ7VUxkLtYd1s6Ez/L9vBAV/Tf1NbqWM0G1iC0sLADq6UwCdRKJGmDvBrK+h/mkvxyOlN6vXHBj7Q30Tx61p6D0XJCgF8pJTNszaVXEHEekYpIbFK+D6nB+lvEdm50RCPoit6QNsskpdUlPjL1hOChK8KtQGoVBy61XySmkQxxQPiYjD7eO77NL3s4dQeicgYMxCI6ANWHFlr3rG/NCW3rMIGRJRD9zxFw7oks4cNjk3BYpNZCznzeUrXdrRjFi2LJJQkXU1F4yQ/9d58/TKwoxfQtijPc8thtBvF3NXvwELwD4fWL8sFNqTESg==

MARIA ASELA REIG ALAMILLO | Fecha:2026-02-28 10:41:49 | FIRMANTE
eZ+fgo6HmscIiBOx7a9aLgEJIviA2pJhtFChwCs07ONfMQH9joPZxIOc3pxZrHaGQFMwMJ6+87nxdlcEHAxpjxpDBQomjnatS0vvwx0OYesHSwLnyrmYqB+82AqgLARiMnbhKaGz9PluuMmtMbWqQc5dFlBfe+XsxqabnxMMXXVNCPreH8EVh/hxNo3t559W6NCesdqjoVc3vet7QvLvM16npwzNb07E7SBKQCkr9bAL3r7WU9m4unKuRsa+pBPn7fn1VHMq2TEXnD7OnRfd5CmYs1vdGf9Ba5LTSwtD/QUqGYLwjrnSrVIRFlj3yJNfIPjpg5xpRE6Ivc0Xm1dF7Q==

MANUEL MONTES Y GOMEZ | Fecha:2026-02-28 13:24:57 | FIRMANTE
rRaheybq9FBogktbyY/OVIMf3Aa+ArtlcQ7Ph3W8IFSR3fGdTMH0B3Mwa6eA/Oi4HgrfWU9aqwINibxa+pyu6XKIgBXe2evPRPs1PPJ0L5z05qNUDhM1RvNpXXyNqIgJAVy8J63FUPDPXrW9n4ub6anf1crle2bRo2Q79nB5b3n/g3aqcf/ONBCQiZT2SEV4QACXy+i483He4aBJ5mXU3e1HA3ksmGU3Q1+EtSBEfrVvTm1hJX8Ap1vDMLDsX+SOesTBUWXsyhLs0u/w0dpFoaefbS9kPhxN6icGCpTasL25+obivPjRJJeFSLuJvRGtu/57nf5EJNR3WppPkvg/Eg==

JOSE DANIEL ARZATE MENA | Fecha:2026-03-03 07:45:05 | FIRMANTE
WdNKifzLyQmx+9HsgPLR2HemV33U0mf3eOuwY7fVLlU4kFyCGHvfTBaoxSMCXMrm99jPVyl70y2MhkULpRKaW3kxxpEkdGJuXSbE6QsJ9wAs8E8+LbGqYQiDP//Z1qLeB2PJDATFQ07lwRq4+8hAWUtkOwQJ3pqB/2uT8nAjWTSfDBNsO+TXEJc5Vr/PGp5WD/cQgAFeVYyEKFTAcPnf85UQiL6wOV6d2gATg/7gPIE2KnVyj5ViTkknpt/FaoM/XlSsL6unQekX09gU43OaN6ZQadRXZFxd3i0h1s5M/2j8SLUfS8hi/XZIvKnjWCpUletRFHOkGE6p/pNbxiqCdw==

NOÉ ALEJANDRO CASTRO SÁNCHEZ | Fecha:2026-03-03 21:08:04 | FIRMANTE
DbXLYcbIh2KyAru1f7NzcFG2lanibrwPz6SykEvGZzkAdF8SxWxHODttLwGK7WRbLuaFSS70Z+7f8XXAd638xZUgW/0QZMZsmP6whwgeOBj0KC66/r2rFbEEnelC4lCwwktErV4/dZDv7csx0scdpih2z2OPkzSpGqKK0e7aE9zyCrIA1cHXsGDo/Bp+ACg3YOU9+pmNNuxCbBcUeFfqtXn/lZr0EooanRlqQODIYyJ4GlGhUYMCeU6AEqqF4Kuu0kOlPPmomlCAz2tv2g9xPq/MsLngFJZdgY7Yeet6YCub3A6VELolunkOfrR8LeKfk5uLPkVERxm996QgwfA7Ig==

BRUNO LARA GUZMAN | Fecha:2026-03-04 13:07:25 | FIRMANTE
P2GMYSDXrejJPvI9sKFYUaWDmKN1TwQNwj4/NZGNlLteDOn3fl8k9tJvNhNB4jsAGkmXIHHlMKDfb6CDUJDr4dUrmNQ7E3jBNBREhBCl7lBPO4W6OqhUBsv+4qcrznzZvVE707gf/CQwezyCZq9Iy07lRyhABa7c5UQJ8E1jLltWytPVFD50oJYlz86EogYT+hLJM6vqlfCKO2UMCLVmGDBqZvqzkAV7Jwn3x0Vm6T2CGyll+ZGTC2Uyyo/oM7zgiGa3W0j9BMZb6H8An5XIsZLOaJLamS8hxrN9P9eXARaa1NbgmGR0dyWezGQ9x/awZ3FgBpODM/MR7g1WeuDD3A==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

BX0uqWncy

https://efirma.uaem.mx/noRepudio/riZhBStexEDf0rWaLSM3JbYJ6m0Z19Ah