Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 149-161 recibido 05-12-2025 revisado 18-01-2026 aceptado 11-02-2026

Aproximación basada en inferencia lingüística y hechos estructurados para la detección de la alucinación en textos

An Approach Based on Linguistic Inference and Structured Facts for the Detection of Hallucinations in Texts

Jairo Madrigal-Cutillas, Elena Lloret
Department of Language and Computing Systems, University of Alicante, Alicante, Spain
jmcm39@gcloud.ua.es
elloret@dlsi.ua.es

Resumen: El objetivo de este artículo es proponer y evaluar un enfoque para detectar alucinaciones factuales en textos en español, combinando análisis lingüístico con verificación de hechos extraídos de fuentes de conocimiento estructurado, en nuestro caso, Wikidata. Los resultados obtenidos demuestran que el enfoque propuesto permite identificar contradicciones y hechos inventados con una alta precisión (alrededor del 90 %), ofreciendo una herramienta pionera y competitiva con respecto a herramientas similares para la verificación factual en español.
Palabras clave: Alucinación factual, inferencia lingüística, PLN, datos estructurados, generación de lenguaje, verificación factual.

Abstract: The objective of this paper is to propose and evaluate an approach for detecting factual hallucinations in Spanish texts by combining linguistic analysis with fact verification from structured knowledge sources—in our case, Wikidata. The results demonstrate that the proposed approach can identify contradictions and fabricated facts with high precision (around 90 %), providing a pioneering and competitive tool compared to similar frameworks for factual verification in Spanish.
Keywords: Factual hallucination, Linguistic inference, NLP, Structured data, Natural Language Generation, Factual Verification.

1 Introducción

Actualmente, la Inteligencia Artificial (IA) ha adquirido un papel muy protagonista en diversas áreas tanto del conocimiento como de la tecnología. Junto a este impacto, es importante resaltar la aparición de los grandes modelos de lenguaje (LLMs) y su capacidad para generar textos coherentes y fluidos, siendo cada vez más difícil distinguirlos de los creados por los seres humanos (Berber Sardinha, 2024; Sepúlveda-Torres et al., 2025). Este desarrollo ha ido acompañado de un aumento significativo en su complejidad, tamaño y especialización, lo que explica su creciente presencia en ámbitos académicos, profesionales y cotidianos (Bommasani et al., 2022; Weidinger et al., 2021).

Junto con la aparición de los LLMs, su uso ha ido aumentando entre la población en general[1]. Sin embargo, los LLMs pueden producir información convincente pero incorrecta, un fenómeno conocido como alucinación (Maestre et al., 2025). Este fenómeno resulta especialmente problemático en un entorno marcado por el aumento de noticias falsas y contenidos engañosos (Hanley y Durumeric, 2024). Ante esta situación, surge la necesidad de comprobar la veracidad de los textos que generan, especialmente en contextos donde la precisión es fundamental. Esto es crucial puesto que en el caso de que los textos generados por IA presenten información no verídica y se divulguen por internet, estaríamos ante un problema de desinformación muy grave, ya que no podríamos diferenciar la información correcta de la que no lo es.

Por este motivo, el objetivo principal de este artículo es proponer un método novedoso para la detección de alucinación factual en

[1] `https://shre.ink/qour`

ISSN 1135-5948 DOI 10.26342/2026-76-11

Español basado en inferencia lingüística. Para ello, se plantean las siguientes preguntas de investigación:

1. ¿Hasta que punto los modelos basados en inferencia lingüística (NLI) pueden usarse para detectar alucinaciones factuales en textos en español?
2. ¿Qué impacto tiene la integración de información estructurada procedente de fuentes de conocimiento externas, en la mejora de la detección de alucinaciones factuales?

El artículo se organiza en cinco secciones principales. Primero, se presenta la introducción y los trabajos relacionados, que incluye la clasificación de las alucinaciones y una revisión de los métodos y métricas existentes para su detección. Luego, se detalla la aproximación propuesta, describiendo sus componentes. A continuación, se muestran la configuración de la experimentación, así como los resultados y discusión sobre la evaluación del enfoque. Finalmente, se exponen las conclusiones y las posibles líneas de trabajo futuro.

2 Trabajos relacionados

El término "alucinación" se usa en psicología para describir una percepción falsa que se vive como real (Macpherson y Platchias, 2024). Siguiendo esta analogía, en el campo de la generación de lenguaje natural (GLN), se dice que un texto alucina cuando este contiene información aparentemente verdadera, pero que, al verificarla, resulta ser falsa o inventada. Este fenómeno ha cobrado especialmente importancia a raíz de la generación automática de textos, y más concretamente, de textos generados por los primeros modelos de lenguaje y posteriores LLMs. Duran- te estos años, se han propuesto diferentes marcos conceptuales que buscan identificar cuándo un texto generado introduce información incorrecta o no sustentada. Es decir, se trata de un caso en el que el modelo genera una salida que parece adecuada, pero que en realidad es errónea porque inventa hechos o genera contenido sin sentido lógico o factual (Farnschläder, 2025).

Según Cossio (2025), los tipos de alucinación se clasifican en varias categorías, entre las que podemos encontrar: i) factuales, ii) contextuales, iii) lógicas y iv) éticas. En esta investigación vamos a centrarnos en las alucinaciones factuales.

Las alucinaciones factuales, también denominadas errores de hecho, se definen como aquellos que tienen lugar cuando un modelo de IA produce información incorrecta, contrastándola con una fuente fiable de información. Estos pueden ser cálculos incorrectos, inexactitudes históricas o falsedades científicas. Ouyang (2025) estudió este tipo de alucinación, a partir del desarrollo de un conjunto de datos (TreeCut) para evaluar la tendencia de los LLM a generar respuestas erróneas en problemas matemáticos no resolubles. Entre los resultados obtenidos, destacaron la dificultad de estos modelos para identificar y manejar adecuadamente tales problemas.

2.1 Detección de la alucinación

Entre las primeras aproximaciones de detección de alucinaciones, destacan las propuestas de Dušek y Kasner (2020), quienes diferenciaron entre alucinación y omisión dependiendo de la correspondencia entre la entrada y la salida. Por otro lado, Thomson y Reiter (2020) clasificaron los errores en categorías más específicas (números incorrectos, entidades erróneas, etc.), proporcionando una visión más léxica del problema. Más adelante, Ji et al. (2023) introdujeron la distinción entre alucinaciones intrínsecas y extrínsecas, atendiendo a si la salida contradice la entrada o simplemente no puede verificarse a partir de ella.

A partir de estas propuestas, van Deemter (2024) plantea un marco más formal basado en la relación semántica *"follows from"*. Este enfoque ofrece una visión más estructurada, aunque presenta limitaciones en aspectos pragmáticos, análisis contextual y dominios especializados.

Cuando la alucinación implica hechos verificables, una de las estrategias que se pueden utilizar es el uso de modelos de inferencia de lenguaje natural (NLI). Estos modelos clasifican la relación entre una premisa y una hipótesis en *entailment*, *contradiction* o *neutral* (Bowman et al., 2015), permitiendo verificar si una afirmación generada está respaldada por un conocimiento dado. La Tabla 1 muestra ejemplos típicos de estas relaciones.

Aunque estos modelos son herramientas clave para la verificación factual, su rendimiento depende de la disponibilidad de hechos relevantes y de su capacidad para manejar relaciones semánticas complejas.

Por ejemplo, trabajos recientes utilizan

Premisa	Hipótesis	Relación
La conferencia comenzó a las 10 de la mañana	El evento empezó por la mañana	Implicación textual
Mozart era futbolista	Mozart era músico	Contradicción
El restaurante sirve comida japonesa	El restaurante es caro	Neutral

Tabla 1: Ejemplos de relaciones NLI.

NLI para detectar alucinaciones en textos generados automáticamente. Badathala, Saxena, y Bhattacharyya (2023) aplican NLI para mapear relaciones de *implicación textual*, *contradicción* y *neutral* a categorías de alucinación en resúmenes. Kang, Blevins, y Zettlemoyer (2024) muestran que las métricas basadas en NLI correlacionan bien con juicios humanos, aunque presentan limitaciones en "single-fact hallucinations". Además, Heo, Son, y Park (2025) implementan un proceso de descomposición de respuestas y verificación factual mediante NLI a nivel de hechos atómicos, mejorando la granularidad de la detección.

2.1.1 Métricas y benchmarks

El auge de los LLM ha impulsado el desarrollo de métricas y benchmarks destinados a evaluar la consistencia factual de los textos generados. Entre los más relevantes encontramos FactCHD (Chen et al., 2024), que evalúa la detección de alucinaciones que contradicen hechos verificables e incluye múltiples dominios y cadenas de evidencia (Xu et al., 2025).

La mayoría de los trabajos previos sobre detección de alucinaciones se han centrado casi exclusivamente en el inglés, idioma en el que se encuentran los principales benchmarks, recursos de verificación factual y modelos NLI más avanzados. Aunque recientemente han surgido iniciativas en otras lenguas (Zhang et al., 2025), la investigación en español sigue siendo muy limitada: no existen enfoques consolidados, ni datasets, ni sistemas orientados específicamente a la detección automática de alucinación factual en textos generados por IA. En este contexto, nuestro enfoque constituye una de las primeras propuestas centradas explícitamente en el castellano y aporta una estrategia novedosa al introducir un uso sistemático de conocimiento no estructurado para la verificación factual, complementado con análisis lingüístico y modelos NLI. Esta combinación metodológica contribuye tanto a llenar el vacío existente en la literatura como a la creación de recursos que pueden impulsar nuevas líneas de investigación en este ámbito.

Para evaluar la alucinación en los textos generados, se han propuesto diversas métricas automáticas. Entre las más utilizadas se encuentran FEQA (Durmus, He, y Diab, 2020) y QAGS (Wang, Cho, y Lewis, 2020), que permiten medir la consistencia factual comparando el contenido generado con fuentes de referencia. Además, benchmarks como FactCHD (Chen et al., 2024) reportan métricas simples de proporción de afirmaciones falsas o no verificables (hallucination rate), ofreciendo una visión cuantitativa de la fidelidad de los textos. De nuevo, todas las métricas existentes se han centrado principalmente en el idioma inglés.

3 Wikidata como fuente de conocimiento.

Para verificar la validez factual de las afirmaciones en el texto, nuestro enfoque se apoya en Wikidata [2], una base de conocimiento multilingüe, colaborativa y estructurada que organiza la información en forma de entidades y propiedades. Cada hecho se representa como una tripleta *(entidad, propiedad, valor)*, donde las entidades y propiedades están identificadas mediante códigos únicos llamados QID (para entidades) y PID o números de propiedad (para propiedades). Esta estructura permite acceder a datos normalizados y verificables sobre personas, lugares, obras, organizaciones y otros conceptos, facilitando la recuperación automática de información.

Por ejemplo, la información sobre el escritor "Gabriel García Márquez" se representa en Wikidata mediante tripletas, tales como:

- (Q5582, P569, 1927-03-06) — Q5582 corresponde a Gabriel García Márquez y

[2] `https://www.wikidata.org/`

P569 es la propiedad "fecha de nacimiento".

- (Q5582, P27, Q739) — P27 indica "nacionalidad" y Q739 corresponde a Colombia.

4 Propuesta de enfoque para la detección de alucinaciones factuales

En esta sección describimos el enfoque propuesto para la detección de alucinaciones factuales en textos en castellano. Nuestra metodología se articula como un *pipeline* modular dividido en cinco etapas que permiten analizar el texto de forma progresiva, desde el tratamiento lingüístico inicial hasta la verificación factual.

Este diseño combina técnicas de análisis lingüístico, recuperación de información y modelos de inferencia textual (NLI) con conocimiento estructurado procedente de Wikidata. El núcleo del enfoque integra componentes específicos para la segmentación de oraciones, la identificación de sujetos y la extracción de entidades, lo que permite contrastar cada afirmación con fuentes fiables. Gracias a esta arquitectura modular, es posible combinar herramientas externas, como modelos de lenguaje y bases de conocimiento, para clasificar cada afirmación según su grado de veracidad.

En la Figura 1 se muestra un esquema del flujo principal de procesamiento, que ilustra cómo el enfoque propuesto integra las distintas etapas mencionadas. En las siguientes secciones se detallan cada una de estas fases.

4.1 Segmentación del texto en oraciones

En esta fase, dividimos el texto de entrada en oraciones para analizarlas individualmente. En nuestro enfoque, esta función se realiza con la librería *SpaCy*[3], debido a su reconocida eficiencia y versatilidad dentro del ámbito del Procesamiento de Lenguaje Natural (PLN) actual. Para ello, se utiliza el modelo *es_core_news_md*[4]. Este modelo es adecuado para tareas como tokenización, segmentación de oraciones, lematización, reconocimiento de entidades y análisis sintáctico.

4.2 Detección y extracción del sujeto

Con las oraciones identificadas, el siguiente paso es identificar el sujeto de las mismas. Esta tarea es fundamental, ya que, en nuestro caso, el sujeto es el componente al que se asocia la información que posteriormente se contrastará con bases de conocimiento externas. En este enfoque se han desarrollado dos aproximaciones para la detección y extracción de sujetos: la primera utilizando técnicas de análisis lingüístico, usando *SpaCy* y la segunda, usando *GPT-4o*[5].

El método que utiliza SpaCy se basa en el análisis sintáctico y reconocimiento de entidades nombradas de la oración, utilizando el modelo mencionado en el apartado anterior. Podemos dividir este proceso en tres etapas:

1. Primero, se intentan extraer entidades nombradas de tipo **PER** (persona) u **ORG** (organización) en español.
2. Si no se encuentra ninguna entidad relevante se repite el proceso con un modelo en inglés, *en_core_web_trf*[6], dado que algunas oraciones pueden contener nombres propios o extranjerismos.
3. Si aún no se detectan entidades, se analiza la estructura sintáctica para identificar el sujeto gramatical mediante las dependencias **nsubj** y **nsubj:pass**.

Por otro lado, el método basado en el LLM GPT-4o consiste en la obtención del sujeto mediante el diseño de un *prompt* explicado más abajo. La principal ventaja de este último método es su capacidad para interpretar oraciones complejas, desambiguar sujetos implícitos y aplicar reglas semánticas específicas mediante una instrucción en lenguaje natural.

El *prompt* utilizado se ha diseñado en base a estas reglas:

- Si el sujeto es una persona u organización, devuelve su nombre completo.
- Si se menciona un cargo o título sin identificar al individuo (por ejemplo, "el presidente" sin especificar quién), la oración se ignora.
- Si el sujeto es una obra (por ejemplo, película, libro, serie), devuelve únicamente su título.

[3] https://spacy.io/
[4] https://spacy.io/models/es#es_core_news
[5] https://openai.com/index/hello-gpt-4o/
[6] https://spacy.io/models/en#en_core_news

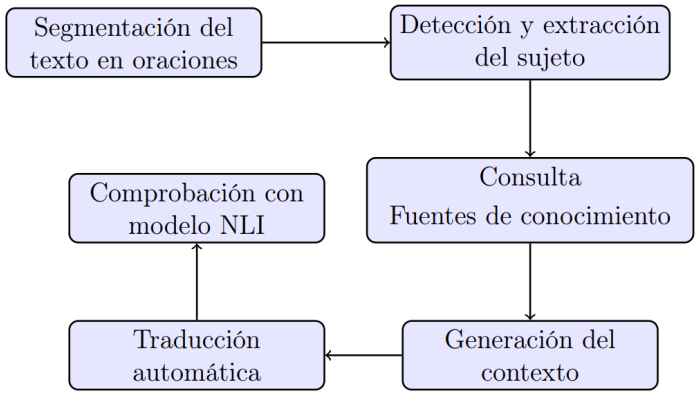

Figura 1: Pipeline principal del enfoque propuesto para la detección de alucinaciones factuales.

- Si el sujeto es un territorio o país, lo devuelve solo si no hay una persona asociada.
- En caso de ambigüedad entre persona y lugar, se prioriza a la persona.

En otras palabras, el *prompt* solicita al modelo que identifique y extraiga el sujeto principal de cada oración siguiendo estas reglas específicas, asegurando que la información relevante se capture de manera consistente. En el Anexo A se puede ver el prompt completo utilizado. Cabe mencionar que en esta primera versión del enfoque no se realiza un proceso exhaustivo de detección y resolución de anáforas y correferencias. Sólo, en caso de que una oración no contenga sujeto, porque esté omitido, se establece una regla para asociarlo con el sujeto de la oración más cercana.

4.3 Consulta a fuentes de conocimiento

Una vez extraído el sujeto principal de la oración, el siguiente paso en el *pipeline* consiste en consultar la base de conocimiento seleccionada para identificar la entidad correspondiente y recuperar la información factual relevante. En nuestro enfoque utilizamos Wikidata como fuente de conocimiento, extrayendo nuestra información estructurada a través de su API oficial.

Para obtener el identificador único de la entidad, se realiza una búsqueda directa del sujeto en la API de Wikidata y se selecciona el primer resultado como la coincidencia más probable. A partir de este identificador se recuperan las *claims* asociadas —esto es, propiedades y valores relevantes— así como las descripciones breves disponibles en español. En la Tabla 2 se muestran algunos ejemplos de las propiedades utilizadas en función del tipo de entidad.

El objetivo de este proceso es construir un conjunto de afirmaciones concisas y verificables que describan la entidad, a partir de propiedades comunes en distintos dominios del conocimiento. Para asegurar la coherencia y la utilidad de los hechos recuperados, la extracción se limita a un conjunto predefinido de propiedades de interés, agrupadas según el tipo de entidad (persona, lugar, obra, organización, etc.). Esta restricción es necesaria para garantizar que posteriormente sea posible generar oraciones bien formadas a partir de los valores obtenidos, de manera que los modelos NLI puedan evaluar con precisión la relación factual entre las afirmaciones originales y la información recuperada.

De este modo, para cada propiedad definida en el conjunto de interés, si la entidad la contiene, se extrae su valor y se incorpora a una lista final de hechos. Por ejemplo, la propiedad P19 (lugar de nacimiento) se recupera únicamente si forma parte del grupo de propiedades predefinidas para entidades tipo “persona”, permitiendo después construir afirmaciones verificables a partir de dicha información. Este conjunto estructurado de hechos constituye la base sobre la que se realizará la verificación factual de las oraciones de entrada.

4.4 Generación del contexto

Una vez extraído el sujeto y los hechos factuales con los que está relacionado, el siguiente paso es la creación de oraciones con sentido para formar un pequeño contexto asociado

Categoría	Propiedades utilizadas
Personas	P106 (Ocupación), P166 (Premio recibido)
Obras y medios	P50 (Autor), P57 (Director), P58 (Guionista)
Lugares	P17 (País), P131 (Ubicación administrativa)
Astronomía	P2583 (Clase espectral), P2120 (Gravedad)
Deportes	P54 (Miembro de equipo), P1350 (Victorias)

Tabla 2: Ejemplo de propiedades extraídas de Wikidata según la categoría de la entidad.

a dicho sujeto, que pueda ser utilizado posteriormente para contrastar la información. Para este proceso, se ha diseñado manualmente una plantilla donde se van añadiendo oraciones conteniendo la información extraída de la fuente de información, para esta investigación, Wikidata.

La plantilla para crear la oración consiste en una estructura fija que comienza siempre con la descripción del sujeto (en caso de estar disponible). Esta descripción proporciona un contexto breve pero útil para entender de quién o qué se está hablando. En caso de no contar con una descripción, la oración inicial se construye simplemente con el nombre del sujeto seguido de un verbo introductorio, en nuestro caso "es".

A continuación, se van concatenando diferentes oraciones independientes, cada una correspondiente a un hecho específico extraído del sujeto. El uso de plantillas definidas manualmente nos asegura tanto la corrección sintáctica como la claridad semántica de las oraciones generadas. Cada tipo de hecho está asociado a una plantilla concreta, lo que facilita la generación controlada del texto y permite mantener la coherencia del contexto. La Tabla 3 muestra algunos ejemplos de oraciones creadas dependiendo del tipo de la propiedad.

Estas plantillas son fácilmente ampliables a medida que se incorporan nuevas propiedades o tipos de hechos. El objetivo principal es que cada oración transmita un hecho verificable y específico, manteniendo una estructura clara y sencilla para facilitar su posterior análisis por parte del modelo de inferencia.

Por ejemplo, si introducimos la oración *"El poeta Miguel Hernández no es español"*, el sujeto extraído será Miguel Hernández y realizará toda la búsqueda de hechos asociados a *"Miguel Hernández"* en la base de conocimiento. Con los hechos extraídos y la plantilla explicada anteriormente, crea la siguiente oración:

Según Wikidata, Miguel Hernández, poeta español, es ser humano, nació en Orihuela el Fecha de nacimiento: 1910-10-30, falleció en Alicante el 1942-03-28, su cónyuge es Josefina Manresa, fue dramaturgo, poeta, escritor, con apellido: Hernández y nombre de pila: Miguel, recibió premios como Alta Distinción de la Generalitat Valenciana (fecha: 2022-00-00), Concurso Nacional de Literatura (fecha: 1938-00-00), es conocido por obras como Cancionero y romancero de ausencias (fecha: 1958-00-00), El hombre acecha, Viento del pueblo (fecha: 1937-00-00), Perito en lunas (fecha: 1933-00-00), El rayo que no cesa (fecha: 1936-00-00), Elegía a Ramón Sijé, Nanas de la cebolla (fecha: 1939-00-00), su página web oficial es http://www.miguelhernandezvirtual.es/.

Como se puede observar, a partir de los hechos extraídos se han ido creando diferentes oraciones con los verbos adecuados para crear un pequeño fragmento donde está contenida toda la información de una manera resumida, además de ser sintáctica y semánticamente correcta.

4.5 Traducción automática

Con la oración con los hechos factuales creada, ya se dispone tanto de la oración hipótesis, como de la oración generada para contrastarla, que serán las entradas para el modelo de inferencia lingüística (modelo NLI).

La razón por la que se decidió integrar una etapa de traducción automática se debe a que los modelos NLI entrenados para otros idiomas distintos del inglés, y en especial para el castellano (idioma en el que se centra nuestra investigación), no alcanzan los mismos niveles de rendimiento que los modelos en inglés, especialmente en tareas complejas de detección de contradicciones o implicaciones. Sin embargo los sistemas de traducción automática, y en concreto, los que traducen del español al inglés, sí que obtienen actual-

Tipo de hecho	Plantilla de oración generada
Descripción	*Según Wikidata, [Entidad]* es [descripción], [tipos sujeto].
Nacimiento	nació en [lugar de nacimiento] el [fecha de nacimiento]
Fallecimiento	falleció en [lugar de fallecimiento] el [fecha de fallecimiento]
Guionista	escrito por [nombre del guionista]
Fecha de fundación	fundada el [fecha de fundación]
Industria	pertenece a la industria de [industria]
Accionista	con accionistas como [accionistas]
Propietario	es propiedad de [propietario]

Tabla 3: Ejemplo de plantillas de generación de oraciones por tipo de hecho.

mente un rendimiento muy bueno (Singhal et al., 2024) y por tanto, traduciendo las frases al inglés, nos permitiría beneficiarnos: i) de los modelos NLI más avanzados entrenados para inglés; ii) poder adaptar nuestro enfoque de detección de alucinación para otros idiomas, convirtiéndolo en un enfoque multilingüe. Además, esta solución también ofrece flexibilidad para probar diferentes modelos NLI sin necesidad de ajustar el resto del *pipeline*.

Para la traducción del español al inglés, se ha utilizado el modelo *opus-mt-es-en*[7], desarrollado por el proyecto *Helsinki-NLP* (Tiedemann et al., 2023; Tiedemann y Thottingal, 2020). Siguiendo con el ejemplo anterior, la traducción obtenida es la siguiente:

The poet Miguel Hernandez was not born in Spain.

According to Wikidata, Miguel Hernandez, a Spanish poet, is a human being, was born in Orihuela on the date of birth: 1910-10-30, died in Alicante on the date of death: 1942-03-28, his spouse is Josefina Manresa, was playwright, poet, writer, with surname: Hernández and first name: Miguel, received awards as High Distinction of the Valencian Government (date: 2022-00-00), National Literature Contest (date: 1938-00-00), is known for works such as Cancionero and romancero of absences (date: 1958-00-00), The man stalks, Wind of the people (date: 1937-00-00), Perito on moons (date: 1933-00-00), The ray that does not cease (date: 1936-00-00), Elegía a Ramón Sijé, Nanas de la onion (date: 1939-00-00), its official website is http://www.miguelhernandezvirtual.es/.

4.6 Comprobación con modelo NLI

El último paso es la aplicación del modelo NLI. Como ya se explicó en la sección 2, el objetivo de este modelo es determinar la relación semántica entre ambas oraciones para comprobar si la hipótesis está respaldada por los hechos (*entailment*), si los contradice (*contradiction*) o si la relación es neutral, es decir, que no ha sido capaz de determinar la relación semántica entre las dos oraciones.

En particular, se ha optado por emplear modelos basados en la arquitectura RoBERTa, debido a su buen rendimiento en benchmarks estándar de NLI. En concreto, el modelo utilizado ha sido *roberta-large-mnli*[8].

Este modelo ha sido elegido por varias razones. En primer lugar, se trata de una versión de gran tamaño de RoBERTa (con aproximadamente 355 millones de parámetros), preentrenada sobre un corpus extenso de texto no etiquetado y posteriormente afinada específicamente para la tarea de inferencia lógica natural utilizando el conjunto de datos MNLI. Esto le permite capturar con precisión relaciones semánticas complejas entre pares de oraciones. En segundo lugar, roberta-large-mnli ha demostrado un rendimiento sobresaliente, alcanzando una precisión de aproximadamente un 90 % en la tarea MNLI-matched, lo que lo sitúa entre los modelos más precisos para esta tarea (Liu et al., 2019). Esta fiabilidad es fundamental para un sistema de verificación automática de hechos, ya que permite inferir con alta confianza si la información generada concuerda, contradice o es irrelevante con respecto a la hipótesis. Cabe destacar que esta comprobación se rea-

[7]https://huggingface.co/Helsinki-NLP/opus-mt-es-en

[8]https://huggingface.co/FacebookAI/roberta-large-mnli

Método	Precisión	Cobertura	Exactitud
SpaCy	0.90	0.99	0.89
GPT-4o	0.96	0.95	0.92

Tabla 4: Precisión, cobertura y exactitud en la extracción de sujetos vinculados a Wikidata.

liza utilizando los sujetos y hechos extraídos automáticamente en las fases anteriores del *pipeline*, de modo que la evaluación final refleja el rendimiento del sistema completo, incluyendo el impacto de posibles errores en la extracción o vinculación de entidades.

Así pues, siguiendo con el ejemplo de Miguel Hernández, al introducir tanto la oración hipótesis, como la oración que hemos generado, se obtienen los siguientes resultados.

Predicción : contradiction — Confianza: [0.99463, 0.00332, 0.00203]

Estos valores representan las probabilidades asignadas por el modelo a las etiquetas de *contradiction*, *neutral* y *entailment*, respectivamente. Atendiendo a la confianza proporcionada (0.99463), podemos observar que el modelo está casi completamente seguro de que la hipótesis es falsa, ya que afirma que este poeta no nació en España. Sin embargo, los hechos extraídos de Wikidata confirman que este poeta sí nació en España.

5 Configuración de la experimentación

En esta sección se describen los datos, los experimentos realizados y las métricas utilizadas para evaluar el enfoque propuesto.

5.1 Conjunto de datos

Para llevar a cabo los experimentos se ha utilizado como base el conjunto de datos *FactCHD* (Chen et al., 2024). Este *benchmark* está estructurado en oraciones siguiendo el patrón *Sujeto + Verbo + Hecho*, acompañadas de una etiqueta que indica si el hecho expresado contradice o no la información conocida. Esta estructura facilita la evaluación del enfoque, permitiendo introducir la oración y comparar el resultado obtenido con el esperado. Se ha seleccionado una muestra de 2000 ejemplos de este conjunto de datos para la evaluación, y se tradujeron al castellano, ya que originalmente estaban en inglés. Además, estos ejemplos muestran una diversidad de contenidos, incluyendo autores, películas, hechos históricos y otros ámbitos temáticos.

5.2 Experimentos realizados

Se han planteado dos experimentos principales:

- **Extracción de sujetos:** se comparan los dos enfoques para la identificación de sujetos (*SpaCy* vs. *GPT-4o*) en las oraciones extraidas del conjunto de datos, evaluando cuántos son correctamente extraídos.
- **Detección de alucinaciones factuales:** se analiza la capacidad del enfoque propuesto para clasificar las oraciones como correctas, incorrectas o neutras en base a la salida obtenida con el modelo NLI, considerando cada uno de los dos métodos de extracción de sujetos.

5.3 Métricas de evaluación

Para evaluar el rendimiento del enfoque, se han utilizado las métricas de precisión, cobertura (*recall*), exactitud (*accuracy*) y medida F1 (*F1-Score*), la cual es esencial para evaluar el balance entre la precisión y la cobertura, aportando también las matrices de confusión. Además, se estudió la confianza de las predicciones, aplicando un umbral del 70 % para filtrar aquellas clasificaciones poco seguras, y se analizó el tipo de error en la extracción de sujetos para identificar limitaciones de cada método.

6 Resultados y discusión

En esta sección se presentan los resultados obtenidos en los experimentos de extracción de sujetos y detección de alucinaciones factuales, así como un análisis de las matrices de confusión y la discusión sobre las fortalezas y limitaciones del enfoque propuesto.

Para la evaluación se utilizaron los 2000 ejemplos del corpus traducido mencionados en la sección 5.1.

6.1 Estudio comparativo de extracción de sujeto

Se compararon los dos enfoques desarrollados para la extracción de sujetos: SpaCy y GPT-4o. Este análisis permite evaluar cuántos su-

Modelo	Estrategia	Precisión	Cobertura	F1-Score	Exactitud
GPT-4o	Sin Neutros, Sin Umbral	0.9158	0.5562	0.6922	0.7247
	Sin Neutros, Con Umbral	0.9394	0.6165	0.7441	0.7756
SpaCy	Sin Neutros, Sin Umbral	0.9129	0.5540	0.6897	0.7291
	Sin Neutros, Con Umbral	0.9461	0.6266	0.7536	0.7896

Tabla 5: Resultados globales para el enfoque de detección de alucinaciones factuales.

jetos son capaces de identificar correctamente.

Los valores de precisión, cobertura y exactitud para cada método se resumen en la Tabla 4. Respecto a los resultados obtenidos, se observa que ambos enfoques extraen aproximadamente la misma cantidad de sujetos, siendo ligeramente superior la extracción con GPT-4o (+100 sujetos). La precisión, cobertura y exactitud de cada método se calculan considerando los sujetos correctamente extraídos y vinculados a Wikidata como verdaderos positivos. Los errores se dividen en sujetos no extraídos (falsos negativos) y sujetos extraídos con QID incorrecto (falsos positivos).

6.2 Evaluación del enfoque de detección de alucinaciones factuales

El análisis se centró en la capacidad del sistema para generar predicciones correctas, distinguiendo entre la detección efectiva de alucinaciones (*verdaderos positivos*), la identificación de afirmaciones factuales correctas (*verdaderos negativos*) y las clasificaciones erróneas.

Un aspecto crítico de nuestro *pipeline* es la gestión de la incertidumbre a través de las clasificaciones "neutras". Estas corresponden a casos en los que el sistema no dispone de la información necesaria en la fuente de conocimiento para validar el hecho o en los que el modelo NLI no detecta una relación semántica clara. Para evaluar el rendimiento del enfoque en su tarea fundamental (distinguir entre afirmaciones con alucinación y sin ella), los resultados presentados en la Tabla 5 se calculan excluyendo estos casos neutros. De este modo, nos centramos en medir la eficacia del sistema cuando este cuenta con los datos suficientes para emitir un juicio.

Como se observa en la Tabla 5, el rendimiento es notablemente positivo. Al no considerar los neutros, el enfoque propuesto alcanza una exactitud del 72 % para ambas estrategias a la hora de extraer los sujetos, además de una alta precisión. Sin embargo, se observó empíricamente que las predicciones incorrectas solían estar asociadas a puntuaciones de confianza bajas por parte del modelo NLI. Por ello, se aplicó un umbral del 70 % para considerar válida una clasificación. Bajo este criterio, si ninguna etiqueta supera dicho umbral, el ejemplo se descarta y se contabiliza como neutro por incertidumbre.

Al aplicar este umbral, se produce un incremento en la precisión (superando el 94 % en el mejor caso) y, de manera aparente, en la cobertura relativa mostrada en la Tabla 5. Es importante aclarar que este aumento en la cobertura se debe a que el enfoque propuesto, al ser más selectivo, opera sobre un subconjunto de datos donde la extracción de hechos y la inferencia son mucho más robustas. Al filtrar el ruido y la ambigüedad, el sistema reduce drásticamente los falsos negativos dentro del grupo evaluado, logrando un mejor balance entre precisión y cobertura, como refleja el F1-Score máximo de 0.7536 para el modelo basado en *SpaCy*.

6.3 Estudio comparativo preliminar

Con el fin de situar el rendimiento de nuestra propuesta frente al estado de la cuestión, se ha llevado a cabo una comparativa con *FactCC* (Kryściński et al., 2019), un modelo de referencia en la detección de inconsistencias factuales, al que se ha considerado como sistema base (*baseline*). Dado que *FactCC* realiza una clasificación binaria, se ajustó la salida de nuestro modelo omitiendo la probabilidad de la etiqueta neutra y seleccionando la mayor probabilidad entre las dos etiquetas restantes. Los resultados se presentan en la Tabla 6.

Como se observa en la comparativa, nuestro enfoque mejora sustancialmente los resultados de la línea base en el contexto del castellano, logrando una precisión cercana al 90 % frente al 75.72 % de *FactCC*. Esta dife-

Modelo	Estrategia	Precisión	Cobertura	F1-Score	Exactitud
FactCC (Baseline)	–	0.7572	0.2818	0.4107	0.4717
SpaCy (Propuesta)	Mayor probabilidad	0.8966	0.3380	0.4909	0.5460
GPT-4o (Propuesta)	Mayor probabilidad	**0.8973**	**0.3651**	**0.5190**	**0.5591**

Tabla 6: Comparativa de rendimiento de nuestro enfoque frente a FactCC para el subconjunto de oraciones procesadas.

rencia sugiere que la descomposición modular del texto en hechos estructurados antes del análisis NLI permite una discriminación mucho más fina de las alucinaciones que los modelos que analizan el texto de forma monolítica, validando la arquitectura propuesta para el procesamiento de información en español.

7 Conclusión y trabajo futuro

En este artículo se ha propuesto un enfoque novedoso para detectar alucinaciones factuales en castellano. El enfoque se basa en el uso de modelos NLI y se ha diseñado e implementado de manera modular, lo que le da una flexibilidad para poder adaptarlo fácilmente a otras fuentes de conocimiento, modelos NLI y/o idiomas en el futuro.

La evaluación preliminar del enfoque sobre un conjunto de 2000 oraciones demuestra que es capaz de detectar alucinaciones evidentes o de las que dispone la información necesaria para clasificarlas, alcanzando una precisión alrededor del 90 %.

Aunque los resultados obtenidos han sido en general buenos, el enfoque presenta algunas limitaciones que se pretender abordar como trabajo futuro. En primer lugar, nos planteamos integrar un módulo de resolución de la correferencia que permita mejorar la capacidad del enfoque. También queremos incorporar nuevas fuentes de conocimiento, así como ampliar la plantilla de hechos factuales, para aumentar la cobertura del enfoque. Además, como línea de investigación prioritaria, nos planteamos realizar una evaluación intrínseca y exhaustiva de cada componente del *pipeline*. Dado que el sistema depende de una sucesión de tareas complejas (segmentación, extracción, traducción y clasificación), es fundamental cuantificar la propagación del error y determinar el impacto específico de cada módulo en el desempeño global. Este análisis permitirá identificar con precisión los puntos de fallo, ya sean derivados de una vinculación incorrecta con Wikidata o de ambigüedades en la inferencia lógica, y optimizar la robustez del sistema frente a errores en cascada. Asimismo, pretendemos evaluar el impacto de utilizar modelos lingüísticos más extensos en la arquitectura de SpaCy, para determinar si la mayor densidad de parámetros mejora significativamente la extracción de hechos. Por otro lado, a medio-largo plazo, queremos explorar arquitecturas basadas en RAG que permitirían construir una base de conocimiento íntegramente en un idioma específico, evitando problemas de traducción y posibilitando la inclusión de hechos más específicos o recientes. Finalmente, no descartamos el desarrollo de un prototipo que pueda ser probado y validado por usuarios, con el objetivo de desarrollar y ofrecer una herramienta accesible que permita evaluar la fiabilidad del contenido textual en castellano.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Educación, Formación Profesional y Deportes a través de la beca de colaboración (num. solicitud: 25CO1/000382); por el Ministerio de Ciencia, Innovación y Universidades, a través de los proyectos "SAFEWORDS: Language Anonymization with Ethical and Legal Safeguards through NLP" (AIA2025-163322-C63), el proyecto "Mecánica cuántica para la comprensión y generación del lenguaje" (PID2024-160791OB-I00) financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, y el proyecto "Generación Consciente de Textos" (PID2021-123956OB-I00), financiado por el MCIN/AEI/10.13039/501100011033 y el Fondo Europeo de Desarrollo Regional (FEDER), bajo el lema "Una manera de hacer Europa", UE; por el Proyecto Desarrollo de Modelos ALIA en el marco del Plan Nacional de Tecnologías de Lenguaje -ENIA 2024 del Ministerio para la Transformación Digital

y de la Función Pública y PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024; y por la Fundación Ramón Areces a través del proyecto Criterios de Evaluación para Corpus de Calidad en Inteligencia Artificial (CRITERIA), desarrollado en el marco del II Concurso Nacional para la adjudicación de Ayudas a la Investigación en Humanidades 2025, bajo el tema "Humanidades Digitales" (Referencia: FRAHUMANIDADES25-01).

Bibliografía

Badathala, N., A. Saxena, y P. Bhattacharyya. 2023. NLI to the rescue: Mapping entailment classes to hallucination categories in abstractive summarization. En J. D. Pawar y S. Lalitha Devi, editores, *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, páginas 120–132, Goa University, Goa, India, Diciembre. NLP Association of India (NLPAI).

Berber Sardinha, T. 2024. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1):100083.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. v. Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, y P. Liang. 2022. On the opportunities and risks of foundation models. (arXiv:2108.07258).

Bowman, S. R., G. Angeli, C. Potts, y C. D. Manning. 2015. A large annotated corpus for learning natural language inference. (arXiv:1508.05326).

Chen, X., D. Song, H. Gui, C. Wang, N. Zhang, Y. Jiang, F. Huang, C. Lv, D. Zhang, y H. Chen. 2024. FactCHD: Benchmarking fact-conflicting hallucination detection. (arXiv:2310.12086).

Cossio, M. 2025. A comprehensive taxonomy of hallucinations in large language models. (arXiv:2508.01781).

Durmus, E., H. He, y M. Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Dušek, O. y Z. Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. En B. Davis Y. Graham J. Kelleher, y Y. Sripada, editores, *Proceedings of the 13th International Conference on Natural Language Generation*, páginas 131–137. Association for Computational Linguistics.

Farnschläder, T. 2025. Alucinación AI: Una guía con ejemplos. DataCamp Blog.

Hanley, H. W. A. y Z. Durumeric. 2024. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. (arXiv:2305.09820).

Heo, S., S. Son, y H. Park. 2025. Halucheck: Explainable and verifiable automation for detecting hallucinations in llm responses. *Expert Systems with Applications*, 272:126712.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, y P. Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Kang, H., T. Blevins, y L. Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation. *arXiv preprint arXiv:2402.10496*.

Kryściński, W., B. McCann, C. Xiong, y R. Socher. 2019. Evaluating the factual consistency of abstractive text summarization. (arXiv:1910.12840).

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, y V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Macpherson, F. y D. Platchias. 2024. *Hallucination: Philosophy and Psychology*. MIT Press.

Maestre, M. M., I. Martínez-Murillo, T. J. Martin, B. Navarro-Colorado, A. Ferrández, A. S. Cueto, y E. Lloret. 2025. Roadmap for natural language generation: Challenges and insights. *Procesamiento del Lenguaje Natural*, 74(0):67–79.

Ouyang, J. 2025. TreeCut: A synthetic unanswerable math word problem dataset for LLM hallucination evaluation. (arXiv:2502.13442).

Sepúlveda-Torres, R., I. Martínez-Murillo, E. Saquete, E. Lloret, y M. Palomar. 2025. To write or not to write as a machine? that's the question. *IEEE Transactions on Big Data*, 11(3):1042–1053.

Singhal, A., V. Shao, G. Sun, R. Ding, J. Lu, y K. Zhu. 2024. A comparative study of translation bias and accuracy in multilingual large language models for cross-language claim verification. (arXiv:2410.10303).

Thomson, C. y E. Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. En B. Davis Y. Graham J. Kelleher, y Y. Sripada, editores, *Proceedings of the 13th International Conference on Natural Language Generation*, páginas 158–168. Association for Computational Linguistics.

Tiedemann, J., M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato Y. Scherrer, R. Vazquez, y S. Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Tiedemann, J. y S. Thottingal. 2020. OPUS-MT — Building open translation services for the World. En *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

van Deemter, K. 2024. The pitfalls of defining hallucination. *Computational Linguistics*, 50(2):807–816.

Wang, A., K. Cho, y M. Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. (arXiv:2004.04228).

Weidinger, L., J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, y I. Gabriel. 2021. Ethical and social risks of harm from language models. (arXiv:2112.04359).

Xu, F., X. Hu, Z. Yu, L. Lin, X. Zhang, Y. Zhang, W. Zhou, J. Gu, y X. Wan. 2025. Had: Hallucination detection language models based on a comprehensive hallucination taxonomy. (arXiv:2510.19318).

Zhang, H., S. Anjum, H. Fan, W. Zheng, Y. Huang, y Y. Feng. 2025. Poly-fever: A multilingual fact verification benchmark for hallucination detection in large language models. (arXiv:2503.16541).

A Anexo 1: Prompt de extracción de sujetos (GPT-4o)

A continuación se detalla el *prompt* de sistema diseñado para la extracción de sujetos. Este texto se configuró para obtener una salida limpia que permitiera la vinculación directa con la API de Wikidata:

> *Dada una oración en español, extrae el sujeto principal que debe usarse para buscar en Wikidata. Sigue estas reglas:*
>
> 1. *Si el sujeto es una persona (como actores, políticos, etc.), devuelve su nombre completo, sin*

artículos ni descripciones.
Ejemplo: Ryan Gosling ha estado en un país de África → Ryan Gosling

2. *Si el sujeto es un cargo o título (como "el Papa", "el emperador"), devuelve el nombre de la persona si aparece. Si solo se menciona el cargo, ignora la oración.*
   Ejemplo: El Papa Francisco visitó Brasil → Papa Francisco
   Ejemplo: El Papa visitó Brasil → (ignorar)
3. *Si el sujeto es una obra (película, serie, libro...), devuelve solo el título, sin añadidos como "(película)".*
   Ejemplo: Los Diez Mandamientos es una película épica → Los Diez Mandamientos
4. *Si el sujeto es un territorio, país o región, y no hay ninguna persona como sujeto, devuelve solo el nombre del lugar.*
   Ejemplo: La taiga de España es verde → España
5. *Si hay tanto persona como lugar, prioriza a la persona.*
   Ejemplo: Ryan Gosling ha estado en un país de África → Ryan Gosling

Devuelve solo el sujeto extraído, sin explicaciones ni texto adicional.