UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

Facultad de Ingeniería

Escuela Profesional de Ingeniería en Informática y Sistemas

AGENTE INTELIGENTE ARTIFICIAL CON RAZONAMIENTO DE SENTIDO COMÚN A TRAVÉS DE COMPUTACIÓN EN LA NUBE

TESIS

Presentada por:

Bach. Ledvir Antonio Ventura Acosta

Para optar el Título Profesional de:

INGENIERO EN INFORMÁTICA Y SISTEMAS

TACNA – PERÚ

2024

Universidad Nacional Jorge Basadre Grohmann

Facultad de Ingeniería

Escuela Profesional de Ingeniería en Informática y Sistemas

Acta de sustentación de Título Profesional

En Laboratorio "A" de la Escuela Profesional de Ingeniería en Informática y Sistemas, siendo las 09:40 horas del día 19 de julio del 2024, y cumpliendo lo señalado en el Reglamento de Grados y Títulos de la Facultad de Ingeniería, se reunió el Jurado Calificador integrado por los docentes:

- Mag. Luis Johnson Paúl Mori Sosa	Presidente
- Dra. Ana Silvia Cori Morón	Secretario
- MSc. Hugo Manuel Barraza Vizcarra	Vocal

Designados mediante R.F. Nº 08866-2024-FAIN/UNJBG, para evaluar la Tesis: 'Agente inteligente artificial con razonamiento de sentido común a través de computación en la nube', presentada por el bachiller Ledvir Antonio Ventura Acosta, para optar el Título Profesional Ingeniero en Informática y Sistemas. Le asesoró, Dr. Erbert Francisco Osco Mamani (R.F. N°07440-2022-FAIN/UNJBG).

Dicho acto de sustentación se desarrolló en dos etapas: exposición y absolución de preguntas; procediéndose luego a la evaluación por parte de los miembros del Jurado.

Habiendo absuelto las preguntas que le fueron formuladas por los miembros del Jurado Calificador, y de conformidad con las respectivas disposiciones reglamentarias, procedieron a deliberar y calificar, declarándolo APROBADO por unanimidad, con el calificativo numérico de 15 (QUINCE) y cualitativo de bueno.

Ante la ausencia del presidente del jurado por una reunión propia de sus funciones como Director ESIS, asumió el Mag. Luis Johnson Paul Mori Sosa como accesitario.

Siendo las 10:40 horas del día 19 de julio del 2024, los miembros del Jurado Calificador firman la presente Acta en señal de conformidad.

Mag. Luis Johnson Paúl Mori Sosa
Miembro del Jurado Calificador

Dra. Ana Silvia Cori Morón
Miembro del Jurado Calificador

MSc. Hugo Manuel Barraza Vizcarra
Miembro del Jurado Calificador

CERTIFICADO DE SIMILITUD

Yo, **ERBERT FRANCISCO OSCO MAMANI** en mi condición de asesor acreditado por la Resolución de Facultad N° 07440-2022-FAIN/UNJBG de la tesis, titulada:

"AGENTE INTELIGENTE ARTIFICIAL CON RAZONAMIENTO DE SENTIDO COMÚN A TRAVÉS DE COMPUTACIÓN EN LA NUBE"

Presentada por el **Bach. Ledvir Antonio Ventura Acosta**. Para optar el Título Profesional de Ingeniero en Informática y Sistemas.

Habiendo cumplido con lo establecido en el reglamento de originalidad y de similitud de trabajo de investigación y producción intelectual, considerando que según la revisión, evaluación y análisis realizado a través del **software de similitud textual** Tunitin. Cuenta con el nivel de similitud permitido cuyo porcentaje es 8 %. Por lo que, CERTIFICO LA SIMILARIDAD de la tesis enunciada líneas arriba, la cual está expedita para continuar con los trámites para la obtención del Título Profesional de Ingeniero en Informática y Sistemas, según corresponda consiguientemente la publicación en el repositorio institucional.

Huella digital

Dr. Erbert Francisco Osco Mamani
DNI: 00409196

Huella digital

Bach. Ledvir Antonio Ventura Acosta
DNI: 70672970

turnitin **Identificación de reporte de similitud:** oid:23228:363145773

NOMBRE DEL TRABAJO
Informe final de tesis.pdf

AUTOR
Ledvir Ventura Acosta

RECUENTO DE PALABRAS
18702 Words

RECUENTO DE CARACTERES
103757 Characters

RECUENTO DE PÁGINAS
91 Pages

TAMAÑO DEL ARCHIVO
965.9KB

FECHA DE ENTREGA
Jun 26, 2024 1:38 PM GMT-5

FECHA DEL INFORME
Jun 26, 2024 1:40 PM GMT-5

8% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base de datos.

- 7% Base de datos de Internet
- 0% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Crossref
- 4% Base de datos de trabajos entregados

Excluir del Reporte de Similitud

- Material bibliográfico
- Material citado
- Material citado
- Coincidencia baja (menos de 15 palabras)
- Bloques de texto excluidos manualmente

Dedicatoria

A mamá y papá.

Agradecimientos

A mi asesor Dr. Erbert Osco, por sus indicaciones para concretar esta investigación.

Índice

Índice de tablas

Índice de figuras

Resumen

El objetivo de esta investigación es verificar que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al punto de referencia (88,8 %), el diseño del estudio es no experimental y el nivel de investigación es descriptivo, la población de estudio consta de 720 problemas de razonamiento de sentido común planteadas por el desafío del esquema de Winograd y problemas de desambiguación de pronombres, la técnica de recolección de datos utilizada fue la observación, la estrategia de recolección de datos fue realizada con la métrica de coincidencia exacta a través de computación en la nube, el procedimiento de recolección de datos fue primero mediante traducción automática adaptar los problemas al español, luego adaptar los problemas a la estructura *Stanford Question Answering Dataset* (SQuAD) para finalmente aplicar la métrica de coincidencia exacta al conjunto de problemas para obtener los datos, el procedimiento estadístico utilizado fue chi-cuadrado prueba de bondad de ajuste. En los resultados de esta investigación se obtuvo una exactitud de 45,33 % en el desafío del esquema de Winograd y un 56,66 % en los problemas de desambiguación de pronombres. Se verificó, que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al punto de referencia.

Palabras clave: agente inteligente, razonamiento de sentido común, computación en la nube, procesamiento de lenguaje natural.

Abstract

The objective of this research is to verify that the accuracy of the artificial intelligent agent with common sense reasoning through cloud computing is lower than the benchmark (88,8 %), the study design is non-experimental and the level of research is descriptive, the study population consists of 720 common sense reasoning problems posed by the Winograd scheme challenge and pronoun disambiguation problems, the data collection technique used was observation, the data collection strategy was carried out with the exact match metric through cloud computing, the data collection procedure was first through automatic translation to adapt the problems to Spanish, then adapt the problems to the Stanford Question Answering Dataset (SQuAD) structure to finally apply the exact match metric to the set of problems to obtain the data, the statistical procedure used was chi-square goodness-of-fit test. In the results of this research, an accuracy of 45,33 % was obtained in the Winograd scheme challenge and 56,66 % in the pronoun disambiguation problems. It was verified that the accuracy of the artificial intelligent agent with common sense reasoning through cloud computing is lower than the benchmark.

Keywords: intelligent agent, common sense reasoning, cloud computing, natural language processing.

Introducción

Esta investigación presenta un agente inteligente artificial que utiliza razonamiento de sentido común a través de un modelo de lenguaje para entender el contexto de un problema y extraer la respuesta de una pregunta propuesta. Este agente inteligente artificial tiene la capacidad de entender el contexto detrás de los datos que analiza. A través de computación en la nube, se verificó que proveer a un agente inteligente de un modelo de lenguaje, le brinda la capacidad de entender casi el 50 % de lo que procesa en los desafíos presentados. Se tiene como propósito verificar el estado actual del razonamiento de sentido común en español frente al punto de referencia siendo este la investigación propuesta por Sakaguchi et al. (2021) puesto que es la investigación que presenta la mejor exactitud encontrada a la fecha de realizada esta investigación.

Esta investigación tiene la siguiente estructura en el capítulo 1 se presenta el planteamiento del problema, objetivos, hipótesis y el cuadro de operacionalización de variables, en el capítulo 2 se presentan los antecedentes del trabajo de investigación, bases teóricas y definiciones conceptuales, en el capítulo 3 se presenta el planteamiento metodológico, el tipo y diseño de investigación, la población y muestra, técnicas e instrumentos de recolección de datos y técnicas para el procesamiento de datos, en el capítulo 4 se presentan los resultados y la contrastación de hipótesis, en el capítulo 5 se presenta la discusión, el contraste con trabajos de investigación similares, finalmente se presentan las conclusiones, recomendaciones, referencias bibliográficas y anexos.

CAPÍTULO I
PLANTEAMIENTO DEL PROBLEMA

1.1 Antecedentes del problema a investigar

El objetivo de la inteligencia artificial es el desarrollo de máquinas inteligentes capaces de aprender, pensar y razonar de la misma manera que los seres humanos. Hoy en día, la mayoría de los agentes inteligentes son considerados inteligencia artificial débil capaces de aplicaciones particulares pero limitadas. Sin embargo, a través del desarrollo e investigación continuo, se busca la inteligencia artificial sólida, la cual es inteligencia de máquina igual o superior a la inteligencia humana.

Si bien aún estamos lejos de la inteligencia artificial sólida, los investigadores han invertido ampliamente en las bases de la inteligencia artificial, la cual consiste de diversos subcampos incluyendo *machine learning*, procesamiento de lenguaje natural, razonamiento de sentido común, visión por computador, robótica entre otros. Se han presentado grandes avances en los campos mencionados anteriormente a excepción del razonamiento de sentido común puesto que es muy complejo desarrollar agentes inteligentes capaces de pensar y razonar.

El test de Turing

La prueba de Turing, propuesta por Alan Turing (1950), fue diseñada para proveer una definición operacional de inteligencia. Una computadora pasa la prueba si un interrogador humano, luego de proponer algunas preguntas escritas, no es capaz de identificar si las respuestas escritas vienen de una persona o de una computadora. La computadora necesitará contar con las siguientes capacidades:

- Procesamiento de lenguaje natural.
- Representación de conocimiento.
- Razonamiento automatizado.
- Machine learning.

El test de Turing deliberadamente evita el contacto directo entre el interrogador y la computadora, porque la simulación física de una persona es innecesaria para mostrar inteligencia. Sin embargo, la llamada prueba de Turing incluye una señal de video para que el interrogador pueda probar las habilidades perceptuales del sujeto, así como

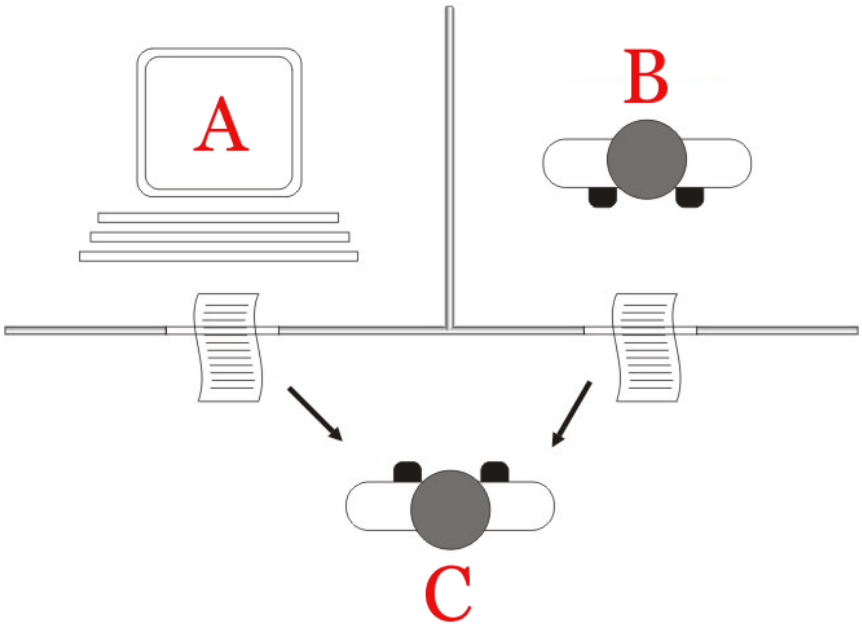
también la oportunidad del interrogador de pasarle objetos físicos, para pasar la prueba de Turing también se necesitaría:

- Visión por computador.
- Robótica.

Estas seis disciplinas componen la mayor parte de la inteligencia artificial (Russell y Norvig, 2010).

Figura 1

Prueba de Turing



Fuente: Pinar Saygin et at., 2000.

Los investigadores se han enfocado en el desarrollo de grandes bases de conocimiento para lograr razonamiento de sentido común en computadoras estas investigaciones se presentan a continuación:

Sap et al. (2019) en su investigación "*Atomic: An Atlas of Machine Commonsense for If-Then Reasoning*" presentan *Atomic*, un atlas de razonamiento de sentido común cotidiano, organizado a través de 877 mil descripciones textuales de conocimiento inferencial, comparado a los recursos existentes que se centran en el conocimiento taxonómico, *Atomic* se centra en el conocimiento inferencial organizado como relaciones de si-entonces.

Tandon et al. (2018) en su investigación "*Reasoning about Actions and State Changes by Injecting Commonsense Knowledge*" indican que comprender un texto de procedimiento, por ejemplo, un párrafo que describe la fotosíntesis, requiere acciones de

modelado y los cambios de estado que producen, de modo que se puedan responder preguntas sobre entidades en diferentes momentos. Aunque varios sistemas recientes han mostrado avances impresionantes en esta tarea, sus predicciones pueden ser globalmente inconsistentes o altamente improbables, en este artículo muestran cómo los efectos de acciones en el contexto de un párrafo pueden mejorar de dos maneras, primero al incorporar restricciones de sentido común globales y segundo al sesgar la lectura con preferencias de un conjunto de datos a gran escala.

Speer y Havasi (2013) en su investigación "*ConceptNet 5: A Large Semantic Network for Relational Knowledge*" presentan *ConceptNet*, un proyecto de representación del conocimiento, que provee un gran gráfico semántico que describe el conocimiento humano en general y cómo este es expresado en lenguaje natural. El alcance de *ConceptNet* incluye palabras y frases comunes en cualquier lenguaje humano escrito. Proporciona un amplio conjunto de conocimientos previos que una aplicación informática que trabaja con texto en lenguaje natural debería conocer.

Panton et al. (2006) en su investigación "*Common Sense Reasoning - From Cyc to Intelligent Assistant*" mencionan que la filosofía detrás del proyecto *Cyc* es que el software nunca alcanzará su máximo potencial hasta que este pueda reaccionar de manera flexible a una variedad de desafíos, además los sistemas no solo deben manejar tareas automáticamente, sino también anticipar activamente la necesidad de realizarlas. Un sistema que se basa en una gran base de conocimiento de propósito general puede potencialmente gestionar tareas que requieren conocimiento sobre el mundo o sentido común, el conocimiento que cada persona supone que sus vecinos también poseen.

Singh et al. (2002) en su investigación "*Open Mind Common Sense: Knowledge Acquisition from the General Public*" es un sistema de adquisición de conocimiento diseñado para adquirir conocimientos de sentido común del público en general a través de la web, describen y evalúan un sistema de campo, que permitió la construcción de una base de conocimientos de sentido común con 450 mil afirmaciones, el sistema adquiere hechos, descripciones e historias al permitir a los participantes construir y completar plantillas en lenguaje natural. Emplea desambiguación del sentido de las palabras y métodos para aclarar el conocimiento ingresado, inferencia analógica para proporcionar retroalimentación y permite a los participantes validar el conocimiento.

1.2 Descripción del problema

La inteligencia artificial viene siendo una fuerza sustancial en la evolución de aplicaciones de software, la amplia investigación e inversiones en inteligencia artificial junto con *big data* han contribuido a todas las áreas de negocios, incluyendo asistentes virtuales, autos autónomos, robótica, salud, seguridad, ventas entre otros. Todo esto mientras impacta simultáneamente como las personas viven sus vidas y se conectan, nuevas aplicaciones de inteligencia artificial y agentes inteligentes surgen a diario.

Si bien las posibilidades de la inteligencia artificial son realmente infinitas, su valor más importante surge cuando es una herramienta accesible, adaptable y flexible. Conforme big data continúa creciendo en disponibilidad, tamaño y variedad, las capacidades computacionales crecen al mismo tiempo.

Según Henderson (2020) indica que existe una falta de madurez en razonamiento de sentido común en inteligencia artificial hoy en día. Esta falta de razonamiento de sentido común limita la capacidad de los agentes inteligentes a entender realmente el contexto completo detrás de los datos que analiza y limita sus valores en la mayoría de los casos que requieren aprendizaje no supervisado.

Este problema es uno de los factores críticos que impiden a la inteligencia artificial alcanzar su máximo potencial a través de una amplia gama de aplicaciones.

The Allen Institute for Artificial Intelligence desarrolló el proyecto Alexandria, una iniciativa de investigación de sentido común en inteligencia artificial. El sentido común representa uno de los problemas más difíciles y fundamentales en inteligencia artificial.

Según Oren Etzioni (Lerman, 2018) exdirector ejecutivo de *AI2* menciona que, a pesar de los recientes logros de la inteligencia artificial, el sentido común, que es trivialmente fácil para las personas, es notablemente difícil para la inteligencia artificial.

Los principales pasos incluyen la introducción de mediciones estándar para las capacidades de sentido común de un sistema de inteligencia artificial, desarrollar métodos novedosos de *crowdsourcing* para adquirir conocimientos de sentido común de las personas a una escala sin precedentes; y desarrollar aplicaciones que utilicen el sentido común para mejorar el rendimiento en una amplia gama de desafíos prácticos de inteligencia artificial.

Gary Marcus (2018) menciona que el sentido común es la condición previa para la inteligencia general; hasta que lleguemos allí estaremos atrapados con una inteligencia artificial limitada que rara vez es robusta y nunca tan flexible como los seres humanos.

Dado que actualmente no existe ninguna investigación previa sobre razonamiento de sentido común en inteligencia artificial en español, se plantea verificar el estado actual del mismo en la presente investigación.

1.3 Formulación del problema

1.3.1 Problema general

¿La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al punto de referencia?

1.3.2 Problemas específicos

a. ¿La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al punto de referencia del desafío del esquema de Winograd?
b. ¿La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al punto de referencia de los problemas de desambiguación de pronombres?

1.4 Objetivos de la investigación

1.4.1 Objetivo general

Verificar que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 88,8 %.

1.4.2 Objetivos específicos

a. Verificar que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 90,1 % en el desafío del esquema de Winograd.
b. Verificar que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 87,5 % en los problemas de desambiguación de pronombres.

1.5 Justificación e importancia de la investigación

Lograr razonamiento de sentido común es un factor importante en el desarrollo de máquinas inteligentes capaces de aprender, pensar y razonar. Los agentes inteligentes con razonamiento de sentido común tendrán la capacidad de explicar sus conclusiones, esto

puede mejorar la interacción humano computador, así como también resolver ambigüedades con procesamiento de lenguaje natural.

Si bien existen diversos desafíos tales como: *CommonsenseQA*, el desafío del esquema de Winograd, *Winogrande*, *COPA* entre otros, se decidió utilizar el desafío del esquema de Winograd porque ofrece una forma precisa y directa de evaluar la comprensión del lenguaje natural en agentes inteligentes, siendo este desafío utilizado por todos los antecedentes del presente trabajo de investigación.

1.6 Alcances y limitaciones

No se cuentan con investigaciones de razonamiento de sentido común en español por lo que en esta investigación no existen antecedentes locales ni nacionales a la fecha de presentación de la presente investigación.

Dado que el razonamiento de sentido común abarca muchas áreas de la inteligencia artificial como procesamiento de lenguaje natural, audio y visión por computador entre otros, solo se abordará el tema mediante el procesamiento de lenguaje natural y en la subtarea de respuesta a preguntas, se trabajará con un modelo de lenguaje del tipo extractivo.

Hay muchos desafíos en la accesibilidad de herramientas basadas en inteligencia artificial y limitaciones en cómo las personas pueden integrar estas tecnologías a sus vidas. Desafortunadamente, desarrollar razonamiento de sentido común es muy complejo debido a los diferentes desafíos que se presentan al tratar de enseñar a las computadoras pensamiento racional.

1.7 Viabilidad del estudio

Se contó con la plataforma de *Google Colab* la cual provee acceso gratuito a recursos computacionales en la nube, que incluyen unidades de procesamiento gráfico (GPUs),

Se realizó una revisión de los modelos de lenguaje pre entrenados disponibles y afinados en la subtarea de respuesta a preguntas disponibles en español encontrándose un pequeño grupo de los mismos.

El desafío del esquema de Winograd y los problemas de desambiguación de pronombres tienen como idioma original el idioma inglés, con la finalidad de viabilizar el estudio, se adaptaron ambos desafíos al idioma español mediante traducción automática.

1.8 Formulación de hipótesis

1.8.1 Hipótesis general

H_0: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube no es menor al 88,8 %.

H_1: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 88,8 %.

1.8.2 Hipótesis específicas

Hipótesis especifica 1

H_0: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube no es menor al 90,1 % en el desafío del esquema de Winograd.

H_1: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 90,1 % en el desafío del esquema de Winograd.

Hipótesis especifica 2

H_0: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube no es menor al 87,5 % en los problemas de desambiguación de pronombres.

H_1: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 87,5 % en los problemas de desambiguación de pronombres.

1.9 Variables

Variable 1: Agente inteligente artificial.

Variable 2: Razonamiento de sentido común.

1.10 Operacionalización de variables

Tabla 1

Operacionalización de variables

Variable 1	**Dimensión**	**Indicador**	**Valor final**	**Tipo de variable**
Agente inteligente artificial	Rendimiento	• Latencia. • Preguntas por segundo. • Tiempo total.	Segundos	Numérica
Variable 2	**Dimensión**	**Indicador**	**Valor final**	**Tipo de variable**
Razonamiento de sentido común	• Desafío del esquema de Winograd. • Problemas de desambiguación de pronombres.	Coincidencia exacta	Porcentaje	Numérica

Fuente: Elaboración propia.

CAPÍTULO II
MARCO TEÓRICO

2.1 Antecedentes del trabajo de investigación

A nivel internacional existen diversas investigaciones que buscan resolver el desafío del esquema de Winograd y los problemas de desambiguación de pronombres, dichas investigaciones se presentan a continuación:

Sakaguchi et al. (2021) investigadores de *The Allen Institute for Artificial Intelligence* y de la Universidad de Washington, en Washington, Estados Unidos, en su investigación "*WinoGrande: An Adversarial Winograd Schema Challenge at Scale*", presentan *WinoGrande*, un conjunto de datos a gran escala de 44 mil problemas, inspirados en el desafío del esquema de Winograd original, pero ajustado para mejorar tanto la escala y la dificultad del conjunto de datos. Los pasos clave de la construcción del conjunto de datos consisten en colaboración colectiva a gran escala, seguida de una reducción sistemática del sesgo utilizando un algoritmo *AfLite* novedoso. Sus experimentos demuestran que los modelos de estado de arte tienen una exactitud menor en *WinoGrande* en comparación con los seres humanos, confirmando que el alto desempeño en el desafío del esquema de Winograd original fue inflado por sesgos en el conjunto de datos. Además, reportan nuevos resultados de estado de arte.

He et al. (2019) investigadores de Microsoft en Hong Kong, China, presentaron en su investigación "*A Hybrid Neural Network Model for Commonsense Reasoning*", proponen un modelo de red neuronal híbrida para el razonamiento de sentido común. Una red neuronal híbrida consta de dos modelos de componentes, un modelo de lenguaje enmascarado y un modelo de similitud semántica, los cuales comparten un codificador contextual basado en *BERT*, pero usan diferentes capas de entrada y salida específicas del modelo. La red neuronal híbrida obtiene nuevos resultados en dos tareas de razonamiento de sentido común, el desafío del esquema de Winograd al 75,1 % y los problemas de desambiguación de pronombres al 90 %.

Kocijan et al. (2019) investigadores de la Universidad de Oxford, en Londres, Reino Unido, en su investigación "*WikiCREM: A Large Unsupervised Corpus for*

Coreference Resolution", presentan *WikiCREM*, un conjunto de datos preciso a gran escala de desambiguación de pronombres. Utilizan un enfoque basado en un modelo de lenguaje para la resolución de pronombres en combinación con su conjunto de datos de *WikiCREM*. Comparan una serie de modelos en una colección de diversos y desafiantes problemas de resolución de correferencia, donde igualan o superan los enfoques de investigaciones previas en 6 de 7 conjuntos de datos.

Wang et al. (2019) investigadores de Microsoft en Minnesota, Estados Unidos, en su investigación "*Unsupervised Deep Structured Semantic Models for Commonsense Reasoning*" proponen dos modelos de redes neuronales basados en un *framework* de modelos semánticos estructurados profundos para abordar dos tareas clásicas de razonamiento de sentido común, los desafíos del esquema de Winograd y la desambiguación de pronombres. Su evaluación muestra que los modelos propuestos capturan efectivamente la información contextual en la oración y la información de correferencia entre pronombres y sustantivos, logrando así una mejora significativa.

Trinh y Le (2018) investigadores de *Google Brain*, en Estados Unidos, en su investigación "*A Simple Method for Commonsense Reasoning*", presentan un método simple para el razonamiento de sentido común con redes neuronales, utilizando aprendizaje no supervisado. La clave de su método es el uso de modelos de lenguaje, entrenados en una cantidad masiva de datos no etiquetados, para calificar preguntas de opción múltiple planteadas por pruebas de razonamiento de sentido común. Tanto en desambiguación de pronombres como en los desafíos del esquema de Winograd, sus modelos superan a los métodos anteriores, sin utilizar costosas bases de conocimiento ni características diseñadas a mano.

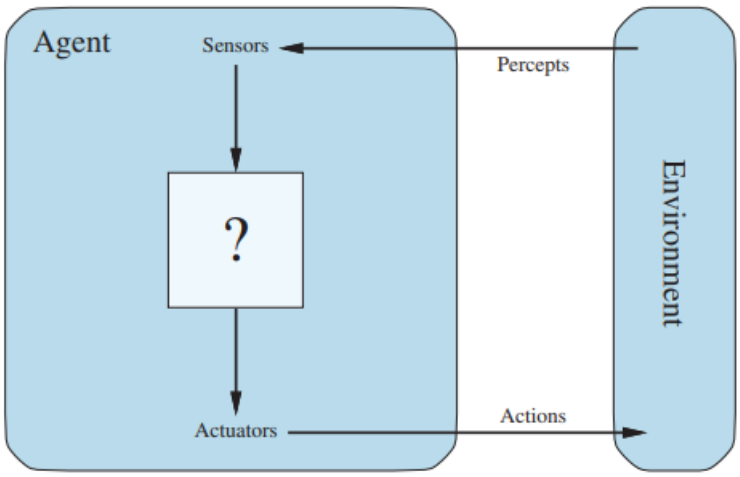
2.2 Bases teóricas

2.2.1 Agente inteligente artificial

Russell y Norvig (2010) definen a un agente inteligente como aquel que puede percibir su entorno a través de sensores y actuar en ese entorno a través de sus actuadores. Esta idea es ilustrada en la figura 2. Un agente humano tiene ojos, oídos y otros órganos como sensores y también cuenta con manos y piernas como actuadores. Un agente robótico puede tener cámaras como sensor y varios motores como actuadores. Un agente de software recibe presiones de teclas, archivos y paquetes de red como datos de entrada y actúa en el entorno al mostrar esto en la pantalla, al escribir archivos y al enviar paquetes de red.

Figura 2

Agente interactuando con su entorno



Fuente: Russell y Norvig, 2010.

Inteligencia artificial

La inteligencia artificial comprende una gran variedad de subcampos, que van desde lo general a lo específico, como por ejemplo jugar ajedrez, probar teoremas matemáticos, escribir poesía, manejar un auto o diagnosticar enfermedades. La inteligencia artificial es relevante en cualquier tarea intelectual, es un campo universal (Russell y Norvig, 2010).

Rich y Knight (1991) definen a la inteligencia artificial como el estudio de cómo hacer que las computadoras hagan cosas, que, por el momento, las personas son mejores.

Procesamiento de lenguaje natural

Según Russell y Norvig (2010) los seres humanos se diferencian de otras especies por su capacidad de lenguaje. Alrededor de hace más de 100 mil años, los seres humanos aprendieron a hablar, y cerca de más de 7 mil años aprendieron a escribir. Aunque los monos, delfines y otros animales han mostrado vocabularios de cientos de señales, solo los humanos pueden comunicarse confiablemente en un sin número de diferentes mensajes cualitativos en cualquier temática usando señales discretas.

Hay otros atributos que son únicamente humanos: ninguna otra especie usa ropa, crea arte, o mira televisión. Pero cuando Alan Turing propuso su prueba, él se basó en el lenguaje, no en arte o televisión. Hay dos razones principales por las que se quiere que los agentes inteligentes sean capaces de procesar el lenguaje natural primero, para comunicarse con los seres humanos y segundo para adquirir información de lenguaje escrito.

Existen millones de páginas de información en la web, casi todas en lenguaje natural, un agente que quiere adquirir conocimiento necesita entender los ambiguos y desordenados idiomas que utilizan los humanos. Para examinar el problema desde un punto de vista de tareas de búsqueda de información, clasificadores de texto, recuperación de información y extracción de información.

Modelos de lenguaje

El uso de modelos de lenguaje es decir modelos que predicen la distribución de probabilidad de expresiones de lenguaje. Un lenguaje puede ser definido como conjunto de cadenas. Los lenguajes naturales, como español o inglés, no pueden ser caracterizados como conjuntos definitivos de oraciones. Los lenguajes naturales son difíciles de tratar porque son muy grandes y cambian constantemente. Nuestros modelos de lenguaje son, a lo mejor una aproximación (Russell y Norvig, 2010).

Bert

BERT (Devlin et al. 2019) es un modelo de transformadores previamente entrenado en un gran corpus de datos en inglés de forma auto supervisada. Esto significa que fue entrenado previamente sólo con textos sin procesar, sin que ningún ser humano los etiquete de ninguna manera con un proceso automático para generar entradas y etiquetas a partir de esos textos. Precisamente fue pre entrenado con dos objetivos:

- Modelado de lenguaje enmascarado: tomando una oración, el modelo enmascara aleatoriamente 15% de las palabras en la entrada, luego ejecuta toda la oración enmascarada a través del modelo y tiene que predecir las palabras enmascaradas. Esto es diferente de las redes neuronales recurrentes tradicionales que usualmente ven las palabras una tras otra, o de modelos autorregresivos como *GPT* el cual enmascara internamente los *tokens* futuros. Esto permite al modelo aprender una representación bidireccional de la oración.
- Predicción de la siguiente oración: Los modelos concatenan dos oraciones enmascaradas como entradas durante el entrenamiento previo. A veces corresponden a frases que estaban a un lado de la otra en el texto original, a veces no. Luego, el modelo tiene que predecir si las dos oraciones se suceden o no.

De esta manera, el modelo aprende una representación interna del idioma inglés que puede luego ser usada para extraer características útiles para tareas posteriores, si se tiene

un conjunto de datos de oraciones etiquetadas, por ejemplo, se puede entrenar un clasificador estándar usando las características producidas por el modelo BERT como entradas.

Deberta

DeBERTa mejora los modelos BERT y *RoBERTa* utilizando atención desenredada y un decodificador de máscara mejorado. Con esas dos mejoras, DeBERTa supera a RoBERTa en la mayoría de tareas de comprensión de lenguaje natural con 80 GB de datos de entrenamiento.

En DeBERTa *V3* (He et al. 2023) mejoró aún más la eficiencia de DeBERTa utilizando el pre entrenamiento de *ELECTRA*. Comparado con DeBERTa, la versión 3 mejora significativamente el rendimiento del modelo en tareas posteriores.

mDeBERTa es la versión multilingüe de DeBERTa la cual utiliza la misma estructura que DeBERTa y fue entrenada con datos multilingües *CC100*. El modelo base de mDeBERTa V3 viene con 12 capas y un tamaño oculto de 768. Tiene 86 millones de parámetros con un vocabulario que contiene 250, 000 *tokens* los cuales introducen 190 millones de parámetros en la capa de incrustación.

Stanford Question Answering Dataset

Stanford Question Answering Dataset (SQuAD), es un conjunto de datos de comprensión lectora el cual consiste en más de 100 mil preguntas planteadas por trabajadores colectivos en un conjunto de artículos de Wikipedia, donde la respuesta a cada pregunta es un segmento de texto del pasaje de lectura correspondiente.

La comprensión lectora o la habilidad de leer texto y luego responder preguntas sobre el mismo, es una tarea desafiante para las computadoras, requiere tanto entendimiento del lenguaje natural y conocimiento sobre el mundo.

Para abordar la necesidad de un conjunto de datos grande y de alta calidad presentan SQuAD donde la respuesta para cada pregunta es un segmento del texto, o una etiqueta, del correspondiente pasaje. SQuAD contiene 107,785 pares de preguntas y respuestas de 536 artículos.

Los pasajes de lectura en SQuAD provienen de artículos de Wikipedia de alta calidad y cubren una diversa gama de temas en una variedad de dominios, desde celebridades musicales a conceptos abstractos. Un pasaje es un párrafo de un artículo y tienen un tamaño variable. Cada pasaje en SQuAD va acompañado de preguntas de comprensión lectora. Estas preguntas están basadas en el contenido del pasaje y pueden

ser contestadas al leer el pasaje. Finalmente, para cada pregunta cuenta con una o más respuestas.

2.2.2 Razonamiento de sentido común

Davis y Marcus (2015) indican que la inteligencia artificial ha visto grandes avances de muchos tipos recientemente, pero que hay un área crítica donde el progreso ha sido extremadamente lento: sentido común ordinario. Proponen las siguientes interrogantes: ¿Quién es más alto, el príncipe William o su bebe el príncipe George? ¿Se puede hacer una ensalada de una camisa? ¿Si perforamos un alfiler en una zanahoria hacemos un agujero en la zanahoria o en el alfiler? Este tipo de preguntas pueden parecer tontas, pero muchas tareas inteligentes, como comprender textos, visión por computador, planeamiento y razonamiento científico requieren el mismo tipo de conocimiento del mundo real y habilidades de razonamiento. Por ejemplo, si uno ve a una persona de 1,80m cargando a otra de 60 cm en sus brazos y nos dicen que son padre e hijo, no es necesario preguntar quién es quién. Si se necesita hacer una ensalada y no hay lechuga, no perdemos el tiempo considerando improvisar tomando una camisa y cortándola. Si se lee el texto, "Yo perforé un alfiler en una zanahoria, cuando lo saqué, tenía un agujero" no necesitamos considerar la posibilidad de que nos referimos al alfiler.

El razonamiento de sentido común es la rama de la inteligencia artificial que apunta a proveer a los agentes inteligentes la capacidad de aprender, hacerse lógicos mientras almacenan y aplican conocimiento en la misma forma que los seres humanos.

El razonamiento de sentido común es importante en muchas tareas de inteligencia artificial, desde comprensión de texto hasta visión por computador, planeamiento y razonamiento.

La importancia del conocimiento sobre el mundo real para el procesamiento de lenguaje natural y en particular para la desambiguación de pronombres de todos los tipos, fue discutida por Bar-Hillel (1960), en el contexto de traducción de máquina. Aunque algunas ambigüedades pueden ser resueltas usando reglas simples que son comparativamente fáciles de adquirir, una fracción sustancial sólo puede ser resuelta utilizando la comprensión sobre el mundo.

Un ejemplo conocido de Terry Winograd (1972) es un par de oraciones "El ayuntamiento negó el permiso a los manifestantes porque ellos temían violencia" vs. "El ayuntamiento negó el permiso a los manifestantes porque ellos abogaban por la violencia".

Desafío del esquema de Winograd

Levesque et al. (2012) presentan una alternativa a la prueba de Turing que tiene algunas ventajas conceptuales y prácticas. Un esquema de Winograd es un par de oraciones que difieren en solo una o dos palabras y que contienen ambigüedad referencial que es resuelta en direcciones opuestas en dos oraciones. Han compilado una colección de esquemas de Winograd, diseñados para que la respuesta correcta sea obvia para el lector humano, pero no puede ser fácilmente encontrada utilizando restricciones seleccionadas, el desafío del esquema de Winograd es presentado con una colección de una oración por cada par y requiere alcanzar precisión a nivel humano al elegir la desambiguación correcta.

El problema con Turing según Levesque et al. (2012) es que tiene algunos aspectos preocupantes, primero, el rol central del engaño. Si consideramos el caso de un agente inteligente tratando de pasar la prueba. Debe conversar con un interrogador y no solo mostrar sus capacidades, sino engañarlo y hacerle pensar que está hablando con una persona. Pero para imitar a una persona bien sin ser evasivo, la máquina necesitará asumir una identidad falsa. Una máquina debería ser capaz de mostrarnos que está pensando sin tener que pretender ser alguien o tener alguna propiedad que no tiene.

El desafío del esquema de Winograd es una prueba de comprensión lectora que incluye preguntas binarias. Aquí dos ejemplos:

- El trofeo no cabe en la maleta café porque es demasiado grande. ¿Qué es demasiado grande? respuesta 0: el trofeo - respuesta 1: la maleta.
- Joan se aseguró de agradecer a Susan por toda la ayuda que le había brindado. ¿Quién le había brindado la ayuda? respuesta 0: Joan - respuesta 1: Susan

Suponemos que las respuestas correctas aquí son obvias en cada pregunta, se cuenta con cuatro características.

1. Dos grupos son mencionados en una oración por frases nominales. Pueden ser dos hombres, dos mujeres, dos objetos inanimados o dos grupos de personas u objetos.
2. Un pronombre o adjetivo posesivo es usado en la oración en referencia a uno de los grupos, pero este también es adecuado para el segundo grupo.
3. La pregunta consiste en determinar el referente del pronombre o adjetivo posesivo. La respuesta 0 es siempre el primer grupo mencionado en la oración y la respuesta 1 es el segundo grupo.

4. Hay una palabra que aparece en la oración y posiblemente en la pregunta. Cuando esta es reemplazada por otra palabra, todo sigue teniendo perfecto sentido, pero la respuesta cambia.

Problemas de desambiguación de pronombres

Los problemas de desambiguación de pronombres son problemas complejos de resolución de correferencia, fueron tomados directamente o modificados de ejemplos encontrados en la literatura, bibliografías, biografías, ensayos, análisis de noticias, historias periodísticas o han sido construidos por los autores. (Morgenstern et al., 2016). Algunos ejemplos de problemas de desambiguación de pronombres:

1. La señora March le dio a la madre té y cereal, mientras vestía al pequeño bebé con tanta ternura como si hubiera sido suyo. Vestía: Sra. March / la madre. Como si hubiera sido: te/bebe.
2. Tom le entregó los planos que había cogido y mientras su compañero los extendía sobre sus rodillas, camino hacia el patio. Sus rodillas: Tom/compañero.
3. Una fría tarde de mayo, la profesora de inglés nos invitó a Marjorie y a mí a su habitación. Su habitación: la profesora de inglés /Marjorie.
4. Marino cayó con estrépito y quedó aturdido en el suelo. Castello instantáneamente se arrodilló a su lado y levantó su cabeza. Su cabeza: Mariano/Castello.

Lo siguiente puede observarse de estos ejemplos: primero un problema de desambiguación de pronombre puede ser tomado directamente de texto (ejemplo 3 es tomado de la autobiografía de Vera Brittain, testamento de juventud) o pueden ser modificados (ejemplos 1, 2 y 4 fueron modificados ligeramente de las novelas Mujercitas, Tom Swift y su dirigible y La ciudad pirata: un cuento argelino). Un problema de desambiguación de pronombres puede consistir en más de una oración, como en el ejemplo 4. En la práctica, raramente se usarán problemas de desambiguación de pronombres que contengan más de tres oraciones. Puede haber múltiples pronombres y por lo tanto múltiples ambigüedades en una oración, como en el ejemplo 1. En la práctica, solo se tienen un número limitado de casos de múltiples problemas de desambiguación de pronombres basados en una sola oración o conjunto de oraciones, ya que malinterpretar un solo texto podría reducir significativamente la puntuación si es la base de múltiples problemas de desambiguación de pronombres.

2.2.3 Computación en la nube

Según Ray (2017) la computación en la nube es la disponibilidad en demanda de recursos computacionales, especialmente el almacenamiento de datos en la nube y el poder computacional sin la administración activa del usuario, tiene como objetivo empoderar a los proveedores de servicios al facilitar la administración eficiente de recursos en centros de datos, proporcionan servicios que se adecuan a las necesidades de los usuarios.

2.3 Definiciones conceptuales

Aprendizaje no supervisado

Un tipo de aprendizaje auto organizado que ayuda a encontrar patrones previamente desconocidos en conjuntos de datos sin etiquetas preexistentes (Hinton y Sejnowski, 1999).

Base de conocimiento

Una base de conocimiento es una tecnología utilizada para almacenar información estructurada y no estructurada compleja utilizada por un sistema informático (Russell y Norvig, 2010).

Big data

Término utilizado para referirse a conjuntos de datos que son muy grandes o complejos para que el software de aplicación de procesamiento de datos tradicional pueda manejarlos adecuadamente (Breur, 2016).

Deep learning

Parte de una familia más amplia de métodos de *machine learning* basados en redes neuronales y representaciones de aprendizaje, se refiere al uso de múltiples capas en la red (LeCun et al., 2015).

Machine learning

Es un campo de estudio en inteligencia artificial que consiste en el desarrollo y estudio de algoritmos y modelos estadísticos que utilizan los sistemas informáticos para realizar una tarea específica de manera eficaz (Russell y Norvig, 2010).

Modelo de lenguaje extractivo

Dada una pregunta y un contexto, extrae la respuesta a la pregunta basándose en la información proporcionada en el contexto.

Modelo de lenguaje generativo

Completa una indicación brindada por el usuario con texto generado automáticamente.

CAPÍTULO III
MARCO METODOLÓGICO

3.1 Planteamiento metodológico

El tipo de investigación es básica o pura, se fundamenta en un argumento teórico y su objetivo fundamental es verificar el conocimiento, es observacional ya que no hay intervención por parte del investigador, según el control de la medición de la variable de estudio es prospectivo ya que los datos son primarios es decir recolectados por el propio investigador, según el número de mediciones sobre la variable de estudio es transversal ya que solo se realizó una medición.

El nivel de investigación es descriptivo ya que solo se busca especificar las características importantes del fenómeno de estudio (Supo, 2015).

El diseño de investigación es no experimental ya que este estudio se realiza sin manipular las variables deliberadamente, según Hernández et al. (2014) en estos estudios no se manipulan intencionalmente las variables para observar los efectos en otras variables.

3.2 Población y muestra

La población de esta investigación consta de dos pruebas de razonamiento de sentido común, el primero el desafío del esquema de Winograd que contiene 600 problemas y los problemas de desambiguación de pronombres que constan de 120 problemas ambos haciendo un total de 720 problemas de razonamiento de sentido común.

Ambas pruebas fueron construidas para probar el razonamiento de sentido común de un agente inteligente, todas las preguntas de estos desafíos son obvias de resolver para los seres humanos utilizando sentido común, pero difíciles de resolver para los agentes inteligentes, en este caso se realizó un muestreo por conveniencia según Supo (2015) se trata de un muestreo no probabilístico deliberado y sin procedimientos específicos.

Tabla 2

Población y muestra

Población y muestra	Cantidad	Muestra
Problemas de razonamiento de sentido común	720	360
Total	720	360

Fuente: Elaboración propia.

3.3 Equipos y materiales

Equipos

Para esta investigación se utilizó una laptop para acceder a la plataforma de *Google Colab* la cual cuenta con las siguientes características:

- Windows 11.
- Intel Core i7-9750H @ 2.60GHz.
- Nvidia GeForce RTX 2060.
- 16 GB de memoria RAM.
- 512 GB de memoria de almacenamiento.

La plataforma de Google Colab la cual cuenta con las siguientes características en la nube:

- Intel Xeon @ 2.20GHz.
- Nvidia T4 GPU.
- 13 GB de memoria RAM.
- 108 GB de memoria de almacenamiento.

Materiales

Los materiales utilizados se listan a continuación:

- Papel bond A4.
- Tinta de impresión.

3.4 Procedimiento de las pruebas experimentales

Para esta investigación no aplica un procedimiento de las pruebas experimentales puesto que es un estudio de nivel descriptivo no experimental.

3.5 Técnicas de recolección de datos

Observación

Técnica de recolección de datos que consiste en el registro de comportamientos y situaciones observables a través de un conjunto de un conjunto de categorías (Hernández, 2014).

Mediante la observación se recolectaron las respuestas del agente inteligente para luego ser almacenadas y tratadas con el software SPSS.

Instrumento de medición

En la tarea de respuesta a preguntas extractivo existen dos métricas, las cuales son: coincidencia exacta y *F1-Score*, para esta investigación se optó utilizar la métrica de coincidencia exacta dado que nos otorga una exactitud absoluta para los desafíos presentados mientras que la métrica de *F1-score* presenta tolerancia a errores.

Métrica de coincidencia exacta

Para verificar la exactitud de un agente inteligente en un determinado desafío, se utilizó como instrumento de medición a la métrica de coincidencia exacta, la cual mide el porcentaje de predicciones que genera el agente que coinciden con cualquiera de las respuestas propuestas por el conjunto de desafíos, dando como valor final la exactitud del agente inteligente en porcentaje, el enlace al código fuente del instrumento se encuentra en la página 2391 de la investigación de Rajpurkar et al. (2016).

Validación del instrumento

La validez se refiere al grado en que un instrumento mide la variable que se busca medir (Hernández et al., 2014).

La métrica de coincidencia exacta tiene una alta validez en contextos donde se requiere una respuesta exacta, pero puede tener una validez limitada en tareas donde se permiten variaciones en las respuestas correctas.

En cuanto a la validez de contenido, la métrica de coincidencia exacta tiene una alta validez de contenido cuando la tarea requiere respuestas que deben coincidir exactamente con una respuesta predefinida. Esto es común en escenarios como la corrección automatizada de exámenes donde se espera que la respuesta del estudiante sea exactamente igual a la respuesta correcta.

Respecto a la validez de criterio, la métrica de coincidencia exacta tiene una buena validez de criterio si el criterio es estrictamente la exactitud de las respuestas. Por

ejemplo, si se usa en un entorno donde solo la respuesta exacta es aceptable, la coincidencia exacta es válida.

En cuanto a la validez de constructo, la métrica de coincidencia exacta es válida en cuanto al constructo si el objetivo es medir la capacidad de un agente para replicar una respuesta exacta.

El instrumento fue tomado de la investigación de Rajpurkar et al. (2016).

Confiabilidad del instrumento

La confiabilidad de un instrumento de medición es el grado en que su aplicación reiterada al mismo individuo u objeto se obtendrán resultados semejantes, así como también resultados coherentes y consistentes (Hernández et al., 2014).

La métrica de coincidencia exacta es altamente confiable, ya que proporciona resultados consistentes y exactos en sus mediciones.

En cuanto a consistencia, la métrica de coincidencia exacta es altamente confiable en términos de consistencia. Si se utiliza para evaluar las mismas respuestas en diferentes momentos o por diferentes evaluadores, los resultados deberían ser consistentes siempre que la implementación de la métrica sea correcta. La confiabilidad es alta porque no depende de juicios subjetivos; una respuesta es o no es una coincidencia exacta, lo cual es fácilmente verificable de manera automatizada.

Respecto a la precisión de medición, la métrica de coincidencia exacta ofrece una medición exacta en términos de ser un valor binario (0 o 1), lo que facilita su interpretación y replicación. No hay ambigüedad en su evaluación, lo que contribuye a su alta confiabilidad.

Sin embargo, esta precisión en la medición puede llevar a la exclusión de respuestas correctas, pero no exactas, lo que podría considerarse una limitación en la confiabilidad si el sistema o la tarea permite variaciones en las respuestas.

El instrumento fue tomado de la investigación de Rajpurkar et al. (2016).

3.6 Técnicas para el procesamiento de datos

Para el procesamiento de datos se utilizó el software *IBM SPSS Statistics 29* en su versión de prueba para la obtención de resultados.

Prueba de normalidad

Se utilizó la prueba de normalidad de *Kolmogorov-Smirnov* para verificar la distribución de la variable aleatoria.

Prueba de hipótesis

Según Supo (2014) los procedimientos para la prueba de hipótesis, también conocido como el ritual de la significancia estadística son cinco y fueron planteados por Fisher; está compuesta de los siguientes pasos:

1. Plantear el sistema de hipótesis.
   Se plantea H_0 y H_1
2. Establecer el nivel de significancia.
   $\alpha = 0,05$ o 5 % de error.
3. Elegir la prueba estadística.
   Para esta investigación se utilizó la prueba chi-cuadrado prueba de bondad de ajuste.
4. Dar lectura al p-valor calculado.
5. Tomar una decisión estadística.

CAPÍTULO IV
RESULTADOS

4.1 Descripción de las pruebas experimentales

Para verificar la exactitud del agente inteligente, se adaptaron los siguientes desafíos de razonamiento de sentido común: el desafío del esquema de Winograd (Levesque et al., 2012) y los problemas de desambiguación de pronombres (Morgenstern et al., 2016) al español mediante traducción automática, luego se adaptaron ambos desafíos al formato de *SQuAD* (Rajpurkar, 2016) que cuenta con los siguientes argumentos:

- *question_column = "question"*: el nombre de la columna que contiene la pregunta en el conjunto de datos.
- *context_column = "context"*: el nombre de la columna de contiene el contexto.
- *id_column = "id"*: el nombre de la columna que contiene el campo de identificación de la pregunta y la respuesta.
- *label_column = "answers"*: el nombre de la columna que contiene las respuestas.

Antes de ejecutar el código siguiente se debe cargar el script de los desafíos adaptado a la plataforma de *Google Colab*, luego se ejecutó el siguiente código:

```
from datasets import load_dataset
from evaluate import evaluator
task_evaluator = evaluator("question-answering")
data = load_dataset("/content/dataset.py", split="validation")
eval_results = task_evaluator.compute(
 model_or_pipeline="timpal0l/mdeberta-v3-base-squad2",
 data=data,
 metric="squad",
)
```

Donde:

model_or_pipeline: Es el modelo utilizado en esta investigación.

data: Especifica el desafío en el cual vamos a ejecutar la medición.

split: Define qué división de *dataset* cargar.

metric: Especifica la métrica que utilizamos en el evaluador.

Se utilizó el modelo de lenguaje *DeBERTaV3* (He et al., 2022) afinado en la tarea de respuesta a preguntas con el conjunto de datos *SQuAD 2.0* (Rajpurkar et al., 2018) el cual es un conjunto de datos de comprensión lectora, que consiste en preguntas propuestas por muchos trabajadores en un conjunto de artículos de Wikipedia, donde la respuesta a cada pregunta es un extracto de texto o una etiqueta correspondiente al contexto, para una información detallada sobre el proceso de *fine-tuning* del modelo ver el anexo 4.

Finalmente, los datos obtenidos incluyen la exactitud de la agente inteligente expresada en porcentaje, cabe mencionar que también se analizaron los datos manualmente obteniendo la misma exactitud en ambos casos.

4.2 Presentación y análisis de los resultados

Presentación de resultados

Para verificar la exactitud del agente inteligente artificial se verificó como se desempeña en dos desafíos: El desafío del esquema de Winograd y los problemas de desambiguación de pronombres. Para realizar esta verificación, se utilizó la métrica de coincidencia exacta para proporcionar respuestas a ambos desafíos que incluyen texto.

Resultados estadística descriptiva

Agente inteligente artificial

En la tabla 3, se presentan los resultados obtenidos para la variable agente inteligente artificial.

Dimensión: Desempeño

Tabla 3

Frecuencia de la dimensión desempeño

Latencia	0,462 (s)
Preguntas por segundo	2,16 (s)
Tiempo total	166,5 (s)

Fuente: Elaboración propia.

Razonamiento de sentido común

A continuación, se presentan los resultados obtenidos para la variable de interés razonamiento de sentido común tras aplicar la métrica de coincidencia exacta. Los resultados se expresan en tablas de frecuencia.

Dimensión 1: Desafío del esquema de Winograd

En la tabla 4 se verificó la capacidad del agente inteligente para responder preguntas planteadas por el desafío del esquema de Winograd. Se obtuvieron cuantas preguntas fueron contestadas correctamente e incorrectamente.

Tabla 4

Exactitud en el desafío del esquema de Winograd

	Frecuencia	Porcentaje (%)
Respuesta correcta	136	45,3
Respuesta incorrecta	164	54,7
Total	300	100

Fuente: Elaboración propia.

Dimensión 2: Problemas de desambiguación de pronombres

En la tabla 5, para verificar la capacidad del agente inteligente para utilizar razonamiento de sentido común, se le formuló al agente inteligente artificial una serie de preguntas sobre problemas de desambiguación de pronombres.

Tabla 5

Exactitud en los problemas de desambiguación de pronombres

	Frecuencia	Porcentaje (%)
Respuesta correcta	34	56,7
Respuesta incorrecta	26	43,3
Total	60	100

Fuente: Elaboración propia.

4.3 Contrastación de hipótesis

Prueba de normalidad

En la tabla 6, se utilizó la prueba de Kolmogórov-Smirnov para identificar si la distribución de la variable de interés cuenta con distribución normal.

Tabla 6

Prueba de Kolmogórov-Smirnov

Estadístico de prueba	0,355
Sig. asin.	0,001

Fuente: Elaboración propia.

1. Planteamiento de Hipótesis

H_0: La distribución de la variable aleatoria no es diferente a la distribución normal.

H_1: La distribución de la variable aleatoria es diferente a la distribución normal.

2. Nivel de significancia

$\alpha = 0,05$ o 5 %.

3. Prueba estadística

Kolmogórov-Smirnov.

4. Valor de P: 0,001

Con una probabilidad de error de 0,1 % la distribución de la variable aleatoria es diferente a la distribución normal.

5. Toma de decisiones

La distribución de la variable aleatoria es diferente a la distribución normal.

Dado que los datos no provienen de una distribución normal no se pudo aplicar una prueba paramétrica por lo tanto se utilizó una prueba no paramétrica en este caso, chi-cuadrado prueba de bondad de ajuste como se aprecia en la tabla 7 y 8.

Tabla 7

Tabla de frecuencias en el desafío del esquema de Winograd

	N observado	**N esperado**	**Residuo**
Correcto	136	270	-134
Incorrecto	164	30	134
Total	300		

Fuente: Elaboración propia.

Donde: El N observado es el número de respuestas correctas e incorrectas del agente inteligente presentando en esta investigación y el N esperado es el número de respuestas correctas e incorrectas del punto de referencia que se tiene para esta investigación.

Tabla 8

Prueba estadística para el desafío del esquema de Winograd

	Exactitud
Chi-cuadrado	677,017
G. l.	1
Sig. asin.	0,001

Fuente: Elaboración propia.

1. Planteamiento de Hipótesis

H_0: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube no es menor al 90,1 % en el desafío del esquema de Winograd.

H_1: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 90,1 % en el desafío del esquema de Winograd.

2. Nivel de significancia

$\alpha = 0,05$ o 5 %.

3. Prueba estadística

Chi-cuadrado prueba de bondad de ajuste.

4. **Valor de P:** 0,001

En la tabla 8, con una probabilidad de error de 0,1 % la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 90,1 % en el desafío del esquema de Winograd.

5. Toma de decisiones

La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor a 90,1 % en el desafío del esquema de Winograd.

Tabla 9

Frecuencia de los problemas de desambiguación de pronombres

	N observado	**N esperado**	**Residuo**
Correcto	34	52	-18
Incorrecto	26	8	18
Total	60		

Fuente: Elaboración propia.

Donde: El N observado es el número de respuestas correctas e incorrectas del agente inteligente presentando en esta investigación, y el N esperado es el número de respuestas correctas e incorrectas del punto de referencia que se tiene para esta investigación.

Tabla 10

Prueba estadística para los problemas de desambiguación de pronombres

	Exactitud
Chi-cuadrado	52,152
G. l.	1
Sig. asin.	0,001

Fuente: Elaboración propia.

1. Planteamiento de Hipótesis

H_0: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube no es menor al 87,5 % en los problemas de desambiguación de pronombres.

H_1: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 87,5 % en los problemas de desambiguación de pronombres.

2. Nivel de significancia

$\alpha = 0,05$ o 5 %.

3. Prueba estadística

Chi-cuadrado prueba de bondad de ajuste.

4. Valor de P = 0,001

En la tabla 10 se indica que con una probabilidad de error de 0,1 % la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 87,5 % en los problemas de desambiguación de pronombres.

5. Toma de decisiones

La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 87,5 % en los problemas de desambiguación de pronombres.

Tabla 11

Frecuencia de ambos desafíos

	N observado	N esperado	Residuo
Correcto	170	320	-150
Incorrecto	190	40	150
Total	360		

Fuente: Elaboración propia.

Donde: El N observado es el número de respuestas correctas e incorrectas del agente inteligente presentando en esta investigación, y el N esperado es el número de respuestas correctas e incorrectas del punto de referencia que se tiene para esta investigación.

Tabla 12

Prueba estadística de ambos desafíos

	Exactitud
Chi-cuadrado	625,740
G. l.	1
Sig. asin.	0,001

Fuente: Elaboración propia.

1. Planteamiento de Hipótesis

H_0: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube no es menor al 88,8 %.

H_1: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 88,8 %.

2. Nivel de significancia

$\alpha = 0,05$ o 5 %.

3. Prueba estadística

Chi-cuadrado prueba de bondad de ajuste.

4. Valor de P = 0,001

Como se indica en la tabla 12 con una probabilidad de error de 0,1 % la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 88,8 %.

5. Toma de decisiones

La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 88,8 %.

CAPÍTULO V
DISCUSIÓN

5.1 Pruebas de validación del modelo experimental

La métrica coincidencia exacta es una medida estándar utilizada para verificar la exactitud de modelos de lenguaje afinados en la tarea de respuesta a preguntas.

En el caso de la coincidencia exacta, esto implica evaluar si la métrica refleja adecuadamente la capacidad del agente inteligente para comprender y responder correctamente a preguntas basadas en un contexto.

Así como también implica que la métrica debe proporcionar resultados consistentes cuando se aplica en diferentes contextos o con diferentes conjuntos de datos.

La métrica de coincidencia exacta tiene tanto validez como confiabilidad cuando se utiliza para verificar la exactitud de agentes inteligentes en la tarea de respuesta a preguntas. Su diseño y aplicación están bien fundamentados en la teoría y la práctica del procesamiento de lenguaje natural, lo que la convierte en una herramienta eficaz para medir la exactitud de estos modelos (Rajpurkar, et al. 2016).

5.2 Aplicación de la tecnología encontrada

La aplicación puede ser muy amplia y variada, abarcando múltiples dominios y escenarios como por ejemplo asistentes virtuales, comercio, educación, investigación, salud entre otros.

Un ejemplo de uso en educación podría ser una aplicación educativa que permita a los estudiantes hacer preguntas sobre cualquier tema y recibir respuestas detalladas extraídas de sus libros y otros recursos educativos presentados por sus docentes.

La aplicación de un modelo de lenguaje afinado en respuesta a preguntas es amplia y diversa. Desde asistentes virtuales y educación hasta investigación y salud, esta tecnología puede mejorar significativamente la eficiencia, precisión y accesibilidad de la información en múltiples dominios.

5.3 Contraste con trabajos de investigación similares

Se verificó que la exactitud del agente inteligente artificial con razonamiento de sentido común en español es menor al punto de referencia, en ambos desafíos se encontró

que la exactitud del agente en el desafío del esquema de Winograd es de 45,33 % y en los problemas de desambiguación de pronombres es de 56,66 %.

A continuación, se discuten los resultados obtenidos con los estudios previos:

Sakaguchi et al. (2021) en su investigación "*WinoGrande: An Adversarial Winograd Schema Challenge at Scale*" presentan *WinoGrande* un conjunto de datos a gran escala de 44 mil problemas, inspirados por el desafío del esquema de Winograd original, pero ajustados para mejorar la escala y la dificultad del conjunto de datos. Los resultados demuestran que la exactitud en el desafío del esquema de Winograd y los problemas de desambiguación de pronombres, con el modelo de lenguaje RoBERTa, alcanzan una exactitud del 90,1 % y 87,5 % respectivamente convirtiéndose en el punto de referencia de esta investigación superando en más de 33,44 % en el desafío del esquema de Winograd y un 30,84 % en los problemas de desambiguación de pronombres a la investigación presentada.

He et al. (2019) en su investigación "*A Hybrid Neural Network Model for Commonsense Reasoning*", proponen un modelo de red neuronal híbrida para el razonamiento de sentido común. Una red neuronal híbrida consiste en dos modelos, un modelo de lenguaje enmascarado y un modelo con similitud semántica, los cuales comparten un codificador contextual basado en el modelo *BERT*, en su investigación obtienen los siguientes resultados en el desafío del esquema de Winograd 75,1 % y en los problemas de desambiguación de pronombres 90 %, superando así en un 29,77 % y 33,34 % respectivamente a la presente investigación.

Kocijan et al. (2019) en su investigación "WikiCREM: A Large Unsupervised Corpus for Coreference Resolution", presentan *WikiCREM*, un conjunto de datos preciso a gran escala de desambiguación de pronombres. Utilizan un enfoque basado en un modelo de lenguaje para la resolución de pronombres en combinación con su conjunto de datos de WikiCREM, en su investigación presentan una exactitud del 71,8 % en los desafíos del esquema de Winograd y un 86,7 % de exactitud el los problemas de desambiguación de pronombres, superando así en 26,5 % y 30 % respectivamente a la presente investigación.

Wang et al. (2019) en su investigación "*Unsupervised Deep Structured Semantic Models for Commonsense Reasoning*" explora el aprendizaje de razonamiento de sentido común de grandes cantidades de texto sin formato mediante el aprendizaje no supervisado. Proponiendo dos modelos de redes neuronales basados en modelos

semánticos estructurales profundos para abordar el desafío del esquema de Winograd en y los problemas de desambiguación de pronombres en los cuales obtiene 62,4 % y 78,3 % respectivamente superando así en 17,07 % en el desafío del esquema de Winograd y 21,64 % en los problemas de desambiguación de pronombres a la presente investigación.

Trinh y Le (2018) en su investigación "*A Simple Method for Commonsense Reasoning*", presentan un método simple para el razonamiento de sentido común con redes neuronales, utilizando aprendizaje no supervisado. La clave de su método es el uso de modelos de lenguaje, entrenados en cantidades masivas de datos no etiquetados para responder a los problemas propuestos por las pruebas de razonamiento de sentido común. Sus modelos obtienen una exactitud de 63,7 % en los desafíos del esquema de Winograd y 70 % en los problemas de desambiguación de pronombres superando así en un 18,37 % y 13,34 % respectivamente a la presente investigación.

El aporte a la línea de investigación es conocer el estado actual del razonamiento de sentido común en español en un agente inteligente artificial, a continuación, se presenta la tabla resumen de contraste con los trabajos de investigación similares:

Tabla 13

Contraste con trabajos de investigación similares

	Modelo de lenguaje	**Conjunto de datos**	**Desafío del esquema de Winograd**	**Problemas de desambiguación de pronombres**
Trinh y Le (2018)	LSTM Personalizado	LM 1 Billion	63,7 %	70 %
Wang et al. (2019)	DSSM	Gutenbert, 1 Billion Word	62,4 %	78,3 %
Kocijan et al. (2019)	BERT	WikiCREM	71,8 %	86,7 %
He et al. (2019)	BERT	WSCR	75,1 %	90,1 %
Sakaguchi et al. (2021)	RoBERTa	Winogrande	90,1 %	87,5 %
Presente investigación	mDeBERTa	SQuAD	45,3 %	56,6 %

Fuente: Elaboración propia.

Conclusiones

1. Se verificó que la exactitud del agente inteligente artificial con razonamiento de sentido común es menor al 88,8 % obteniendo una exactitud promedio de 51 %.
2. Se verificó que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor con 45,33 % de exactitud en el desafío del esquema de Winograd frente al punto de referencia de 90,1 %.
3. Se verificó que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor con 56,66 % de exactitud en los problemas de desambiguación de pronombres frente al punto de referencia de 87,5 %.

Recomendaciones

1. Se plantea como propósito de un siguiente estudio, comparar la exactitud de un agente inteligente artificial extractivo frente a un agente inteligente artificial generativo, para conocer cuál de los dos agentes tiene una mayor exactitud frente a los desafíos planteados en esta investigación.
2. Una posible acción para mitigar el problema sería generar un conjunto de datos de entrenamiento de razonamiento de sentido común para poder afinar a un agente inteligente con los mismos y medir los resultados.
3. Aún no se puede dar una solución definitiva al problema de esta línea de investigación, dado que el tema de razonamiento de sentido común es muy amplio en esta investigación solo se abordó esta línea desde la rama del procesamiento de lenguaje natural, se recomienda explorar otros acercamientos no sólo por el lado de procesamiento de lenguaje natural sino también por otras ramas de la inteligencia artificial, como visión por computador y el análisis de sonido.

Referencias bibliográficas

Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in computers*, 1, 91-163.

Breur, T. (2016). Statistical Power Analysis and the contemporary "crisis" in social sciences. *Journal of Marketing Analytics*, 4(2-3), 61-65.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9), 92-103.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

He, P., Gao, J., & Chen, W. (2022, September). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *In The Eleventh International Conference on Learning Representations.*

He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

He, P., Liu, X., Chen, W., & Gao, J. (2019, November). A Hybrid Neural Network Model for Commonsense Reasoning. *In Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, (pp. 13-21).

Henderson, A. (2020). *Common Sense Reasoning in Autonomous Artificial Intelligent Agents Through Mobile Computing.* [Tesis de maestria, Universidad de Harvard].

Hernández, R. F. (2014). *Metodología de la investigación Vol. 6.* México: McGRAW-HILL.

Hinton, G., & Sejnowski, T. J. (1999). *Unsupervised learning: foundations of neural computation.* MIT press.

Kocijan, V., Camburu, O. M., Cretu, A. M., Yordanov, Y., Blunsom, P., & Lukasiewicz, T. (2019, November). WikiCREM: A Large Unsupervised Corpus for Coreference Resolution. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), (págs. 4303-4312).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*, (7553), 436-444.

Lerman, R. (28 de Febrero de 2018). *The Sattle Times*. Obtenido de The Sattle Times: https://www.seattletimes.com/business/technology/paul-allen-invests-125-million-to-teach-computers-common-sense/

Levesque, H. J., Davis, E., & Morgenstern, L. (2012, May). The winograd schema challenge. *In Thirteenth international conference on the principles of knowledge representation and reasoning.*

McCormick, C. (2020). *Question Answering with a Fine-Tuned BERT*. Obtenido de https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/

Morgenstern, L., Davis, E., & Ortiz, C. L. (2016). Planning, executing, and evaluating the winograd schema challenge. *AI Magazine, 37(1)*, 50-54.

Panton, K., Matuszek, C., Lenat, D., Schneider, D., Witbrock, M., Siegel, N., & Shepard, B. (2006). Common sense reasoning–from Cyc to intelligent assistant. *In Ambient Intelligence in Everyday Life: Foreword by Emile Aarts*, (pp. 1-31). Berlin, Heidelberg: Springer Berlin Heidelberg.

Pinar Saygin, A., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and machines, 10*, (4), 463-518.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (pp. 784-789).

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, November). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (págs. 2383-2392).

Ray, P. P. (2017). An introduction to dew computing: definition, concept and implications. *IEEE Access, 6*, 723-737.

Rich, E., & Knight, K. (1991). *Artificial Intelligence* (second edition ed.). McGraw-Hill.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence a modern approach.* London.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9), 99-106.

Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., & Choi, Y. (July de 2019). Atomic: An atlas of machine commonsense for if-then reasoning. *In Proceedings of the AAAI conference on artificial intelligence*, *Vol. 33, No. 01*, págs. 3027-3035.

Singh, P. L. (2002). Open mind common sense: Knowledge acquisition from the general public. *In On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE: Confederated International Conferences CoopIS, DOA, and ODBASE 2002 Proceedings*, (págs. 1223-1237). Springer Berlin Heidelberg.

Speer, R., & Havasi, C. (2013). ConceptNet 5: A large semantic network for relational knowledge. *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, 161-176.

Supo, J. (2014). *Cómo probar una hipótesis.* BIOESTADISTICO EIRL.

Supo, J. (2015). *Cómo empezar una tesis.* Bioestadístico Eirl.

Tandon, N., Dalvi, B., Grus, J., Yih, W. T., Bosselut, A., & Clark, P. (2018). Reasoning about Actions and State Changes by Injecting Commonsense Knowledge. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (págs. 57-66).

Trinh, T. H., & Le, Q. V. (2018). A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59, 433–460.

Wang, S., Zhang, S., Shen, Y., Liu, X., Liu, J., Gao, J., & Jiang, J. (2019, June). Unsupervised Deep Structured Semantic Models for Commonsense Reasoning. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Volume 1*, págs. 882-891.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, 3(1), 1-191.

Anexos

Anexo 1 Matriz de consistencia.

Agente inteligente artificial con razonamiento de sentido común a través de computación en la nube

<table>
<tr><th>Problema</th><th>Objetivo</th><th>Hipótesis</th><th colspan="2">Variables</th></tr>
<tr><td>Problema general</td><td>Objetivo general</td><td>Hipótesis general</td><td colspan="2">V1: Agente inteligente artificial</td></tr>
<tr><td rowspan="2">¿La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al punto de referencia?</td><td rowspan="2">Verificar que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 88,8 %.</td><td rowspan="2">H₁: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 88,8 %.</td><td>Dimensión</td><td>Indicador</td></tr>
<tr><td>Desempeño</td><td>• Latencia.
• Problemas por segundo.
• Tiempo total.</td></tr>
<tr><td>Problemas específicos</td><td>Objetivos específicos</td><td>Hipótesis específicas</td><td colspan="2">V2: Razonamiento de sentido común</td></tr>
<tr><td rowspan="2">a) ¿La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al punto de referencia en el desafío del esquema de Winograd?
b) ¿La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al punto de referencia en los problemas de desambiguación de pronombres?</td><td rowspan="2">a) Verificar que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 90,1 % en el desafío del esquema de Winograd.
b) Verificar que la exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 87,5 % en los problemas de desambiguación de pronombres.</td><td rowspan="2">H₁: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 90,1 % en el desafío del esquema de Winograd.
H₁: La exactitud del agente inteligente artificial con razonamiento de sentido común a través de computación en la nube es menor al 87,5 % en los problemas de desambiguación de pronombres.</td><td>Dimensión</td><td>Indicador</td></tr>
<tr><td>• Desafío del esquema de Winograd.
• Problemas de desambiguación de pronombres.</td><td>Coincidencia exacta</td></tr>
<tr><td>Tipo de investigación</td><td>Básica o pura</td><td>Nivel de investigación</td><td colspan="2">Descriptivo</td></tr>
<tr><td>Diseño de investigación</td><td>No experimental</td><td>Población - Muestra</td><td colspan="2">720 - 360</td></tr>
</table>

Anexo 2

Ficha técnica de la métrica de coincidencia exacta

Identificación del instrumento	
Nombre de la métrica	Coincidencia exacta.
Tipo de métrica	Métrica de evaluación de exactitud en procesamiento de lenguaje natural.
Áreas de aplicación	Tarea de respuesta a preguntas.
Descripción de la métrica	
Propósito	La métrica de coincidencia exacta se utiliza para medir la capacidad de un modelo de inteligencia artificial para extraer respuestas que coincidan exactamente con las respuestas de referencia, sin variaciones en palabras, orden de palabras o puntuación.
Aplicaciones	Es utilizada principalmente en tareas de respuesta a preguntas y otras áreas del procesamiento de lenguaje natural donde se requiere que la salida generada por el modelo sea idéntica a una respuesta de referencia predefinida.
Especificaciones técnicas	
Método de cálculo	La métrica coincidencia exacta asigna un valor binario (1 o 0) a cada par de respuesta generada y respuesta de referencia. Fórmula: $$EM = \frac{1}{N}\sum_{i=1}^{N}\delta(Pred_i = Ref_i)$$ Donde: • N es el número total de ejemplos evaluados.

	• $\delta(Pred_i = Ref_i)$ es una función que retorna 1 si la predicción ($Pred_i$) coincide exactamente con la respuesta de referencia (Ref_i) y 0 si no coincide.
Valor de salida	Un valor entre 0 y 1, donde 1 indica que todas las respuestas generadas coinciden exactamente con las respuestas de referencia.
Condiciones de uso	
Entrada	Respuesta generada por el modelo y respuesta de referencia predefinida.
Salida	Un valor de coincidencia exacta que indica la exactitud de las respuestas generadas.
Limitaciones	
• La métrica de coincidencia exacta no considera respuestas que sean correctas pero formuladas de manera diferente, lo que puede ser limitante en tareas donde múltiples formulaciones son aceptables. • No distingue entre errores menores (por ejemplo, diferencias de sinónimos o errores gramaticales menores) y errores graves.	
Características y beneficios	
Sencillez	La métrica es fácil de implementar y entender, proporcionando una evaluación clara y directa.
Exactitud	Evalúa de manera precisa si las respuestas coinciden exactamente con las referencias, útil en tareas donde solo una respuesta específica es válida.
Objetividad	Ofrece una evaluación objetiva basada en la igualdad exacta, sin ambigüedades en la interpretación de los resultados.

Mantenimiento y actualización	
Estándares de la industria	Utilizada en conjunto con otros estándares de evaluación en competiciones y puntos de referencia de procesamiento de lenguaje natural, como SQuAD y otros conjuntos de datos de respuesta a preguntas.
Cumplimiento	Cumple con las prácticas comunes en la evaluación de modelos de procesamiento de lenguaje natural y es ampliamente reconocida en la comunidad académica y en la industria.
Ejemplos de uso	
Conjunto de datos de preguntas y respuestas de Stanford	Utilizado para evaluar el desempeño de modelos en la tarea de respuesta a preguntas basadas en texto.
Evaluación comparativa	Comparación de diferentes modelos en términos de su capacidad para generar respuestas exactas en competiciones como la de SQuAD.
Soporte técnico	
Documentación	Disponible en la documentación de la mayoría de las bibliotecas de procesamiento de lenguaje natural, como Hugging Face Transformers, *TensorFlow*, etc.
Implementación	Ejemplos y guías para la implementación de la métrica en Python, disponibles en plataformas de desarrollo y foros de la comunidad de procesamiento de lenguaje natural.
Referencias	Doc. técnica: Rajpurkar et al. (2016)

Fuente: Elaboración propia.

Anexo 3

Desafío del esquema de Winograd

Nº	Contexto/Pregunta	ALTERNATIVAS	
		A	B
1	Los concejales negaron el permiso a los manifestantes porque [temían/propugnaban] la violencia. ¿Quiénes [temían/propugnaban] la violencia?	Los concejales	Los manifestantes
2	El trofeo no cabe en la maleta marrón porque es demasiado [pequeña/grande]. ¿Qué es demasiado [pequeña/grande]?	La maleta	El trofeo
3	Joan se aseguró de agradecer a Susan por toda la ayuda que había [brindado/recibido]. ¿Quién había [brindado/recibido] ayuda?	Susan	Joan
4	Paul trató de llamar a George por teléfono, pero no [tuvo éxito/estaba disponible]. ¿Quién no [tuvo exitoso/estaba disponible]?	Paul	George
5	El abogado le hizo una pregunta al testigo, pero este se mostró reacio a [responderla/repetirla]. ¿Quién se mostró reacio a [responder/repetir] la pregunta?	El testigo	El abogado
6	El camión de reparto pasó zumbando junto al autobús escolar porque iba muy [rápido/lento]. ¿Qué iba tan [rápido/lento]?	El camión	El autobús
7	Frank se sintió [reivindicado/aplastado] cuando su rival de toda la vida, Bill, reveló que él era el ganador de la	Frank	Bill

	competencia. ¿Quién fue el ganador de la competencia?		
8	El hombre no podía levantar a su hijo porque [estaba muy débil/pesaba mucho]. ¿Quién era [débil/pesado]?	El hombre	El hijo
9	La bola grande atravesó la mesa porque estaba hecha de [acero/espuma de poliestireno]. ¿Qué estaba hecha de [acero/espuma de poliestireno]?	La bola	La mesa
10	John no podía ver el escenario con Billy delante de él porque es muy [bajo/alto]. ¿Quién es tan [bajo/alto]?	John	Billy
11	Tom arrojó su mochila a Ray después de que llegó [a la parte superior/al pie] de las escaleras. ¿Quién llegó [a la parte superior/al pie] de las escaleras?	Tom	Ray
12	Aunque corrieron más o menos a la misma velocidad, Sue le ganó a Sally porque tuvo un [buen/mal] comienzo. ¿Quién tuvo un [buen/mal] comienzo?	Sue	Sally
13	La escultura rodó del estante porque no estaba [anclada/nivelado]. ¿Qué no estaba [anclada/nivelado]?	La escultura	El estante
14	El dibujo de Sam estaba colgado justo encima del de Tina y se veía mucho mejor con otro [debajo/encima]. ¿Cuál se veía mejor?	El dibujo de Sam	El dibujo de Tina
15	A Anna le fue mucho [mejor/peor] que su buena amiga Lucy en el examen porque había estudiado mucho. ¿Quién estudió mucho?	Anna	Lucy

16	Los bomberos llegaron [después/antes] que la policía porque venían de muy lejos. ¿Quién vino desde muy lejos?	Los bomberos	La policía
17	Frank estaba molesto con Tom porque la tostadora que le había [comprado/vendido] no funcionaba. ¿Quién había [comprado/vendido] la tostadora?	Frank	Tom
18	Jim [le gritó/consoló] a Kevin porque estaba muy molesto. ¿Quién estaba molesto?	Jim	Kevin
19	El saco de papas se había colocado [encima/debajo] de la bolsa de harina, por lo que había que moverlo/a primero. ¿Qué había que mover primero?	El saco de papas	La bolsa de harina
20	Pete envidia a Martin [porque/aunque] tiene mucho éxito. ¿Quién tiene mucho éxito?	Martin	Pete
21	Estaba tratando de equilibrar la botella boca abajo sobre la mesa, pero no pude hacerlo porque estaba muy [pesada en la parte superior/desnivelada]. ¿Qué [era pesada en la parte superior/estaba desnivelada]?	La botella	La mesa
22	Extiende el mantel sobre la mesa para [protegerla/exhibirlo]. ¿Para [proteger/exhibir] qué?	La mesa	El mantel
23	Los estudiantes mayores estaban intimidando a los más jóvenes, así que los [rescatamos/castigamos]. ¿A quiénes [rescatamos/castigamos]?	Los estudiantes más jóvenes	Los estudiantes mayores

24	Vertí agua de la botella en la taza hasta [llenarla/que estuvo vacía]. ¿Qué estaba [llena/vacía]?	La taza	La botella
25	Susan sabe todo sobre los problemas personales de Ann porque es [entrometida/indiscreta]. ¿Quién es [entrometida/indiscreta]?	Susan	Ann
26	Sid le explicó su teoría a Mark pero no pudo [convencerlo/entenderlo]. ¿Quién no [convenció/entendió] a quién?	Sid no convenció a Mark	Mark no comprendió a Sid
27	Susan sabía que el hijo de Ann había tenido un accidente automovilístico, [así que/porque ella] se lo contó. ¿Quién le contó a la otra sobre el accidente?	Susan	Ann
28	El tío de Joe aún puede vencerlo en el tenis, a pesar de que es/tiene 30 años [mayor/menos]. ¿Quién es [mayor/más joven]?	El tío de Joe	Joe
29	La policía salió de la casa y entró en el garaje, [donde/luego] encontraron/de - encontrar el arma homicida. ¿Dónde encontraron el arma homicida?	En el garaje	En la casa
30	La pintura en la sala de estar de Mark muestra un roble. Está a la derecha de [la librería/una casa]. ¿Qué hay a la derecha de [la librería/una casa]?	La pintura	El árbol
31	Hay un hueco en la pared. Puedes ver el jardín [a través/detrás] de este/a. ¿Puedes ver el jardín [a través/detrás] de qué?	El hueco	La pared

32	El drenaje está obstruido con cabello. Tiene que ser [limpiado/removido]. ¿Qué se tiene que [limpiar/remover]?	El desagüe	El cabello
33	Mi reunión empezaba a las 4:00 y necesitaba tomar el tren a las 4:30, así que no había mucho tiempo. Afortunadamente, [fue corto/se retrasó], así que funcionó. ¿Qué [fue corto/se retrasó]?	La reunión	El tren
34	Hay un pilar entre el escenario y yo, y no puedo [verlo/ver a su alrededor]. ¿[Qué no puedo ver/Alrededor de que no puedo ver]?	El escenario	El pilar
35	Emitieron un anuncio, pero entró un metro en la estación y no pude [escucharlo/oírlo]. ¿Qué no pude [escuchar/oír]?	El anuncio	El metro
36	En medio del concierto al aire libre, la lluvia comenzó a caer, [y/pero] continuó hasta las 10. ¿Qué continuó hasta las 10?	La lluvia	El concierto
37	Usé un trapo viejo para limpiar el cuchillo y luego lo [puse en el cajón/tiré a la basura]. ¿Qué [puse en el cajón/tiré a la basura]?	El cuchillo	El trapo
38	Ann le preguntó a Mary a qué hora cierra la biblioteca, [pero/porque] se le había olvidado. ¿Quién se había olvidado?	Mary	Ann
39	Saqué la botella de agua de la mochila para [que fuera más ligera/tenerla a mano]. ¿Qué [sería más ligera/tendría a mano]?	La mochila	La botella

40	No pude poner la olla en el estante porque [estaba demasiado alto/era demasiado grande]. ¿Qué [estaba demasiado alto/era demasiado grande]?	El estante	La olla
41	Estoy seguro de que mi mapa mostrará este edificio; es muy [famoso/bueno]. ¿Qué es [famoso/bueno]?	El edificio	El mapa
42	Bob pagó la educación universitaria de Charlie. [Es muy generoso/Él está muy agradecido]. ¿Quién es [generoso/agradecido]?	Bob	Charlie
43	Bob pagó la educación universitaria de Charlie, pero ahora Charlie actúa como si nunca hubiera sucedido. [Está muy herido/Es muy desagradecido]. ¿Quién [está herido/es desagradecido]?	Bob	Charlie
44	Bob estaba jugando a las cartas con Adam y estaba muy por delante. Si Adam no hubiera tenido una repentina racha de buena suerte, habría [ganado/perdido]. ¿Quién habría [ganado/perdido]?	Bob	Adam
45	Adam no puede dejar el trabajo aquí hasta que llegue Bob para reemplazarlo. Si Bob se hubiera ido de casa al trabajo a tiempo, [estaría aquí a esta hora/ya se había ido]. ¿Quién [estaría aquí/se habría ido]?	Bob	Adán
46	Si el estafador hubiera logrado engañar a Sam, habría [obtenido/perdido] mucho dinero. ¿Quién habría [obtenido/perdido] el dinero?	El estafador	Sam

47	Era una tarde de verano y el perro estaba sentado en medio del césped. Después de un rato, se levantó y se movió a un lugar debajo del árbol, porque [hacía calor/estaba más fresco]. ¿Qué estaba [caliente/más fresco]?	El perro	El lugar debajo del árbol
48	El gato yacía junto a la ratonera esperando al ratón, pero era/estaba demasiado [cauteloso/impaciente]. ¿Qué era demasiado [cauteloso/impaciente]?	El ratón	El gato
49	Anne dio a luz a una niña el mes pasado. Es una [mujer/bebé] muy encantadora. ¿Quién es una [mujer/bebé] muy encantadora?	Anne	La hija de Anne
50	Alice trató frenéticamente de evitar que su hija [hablara/ladrara] en la fiesta, lo que nos dejó preguntándonos por qué se estaba comportando de manera tan extraña. ¿Quién se estaba comportando de manera extraña?	Alice	La hija de Alice
51	Vi a Jim gritándole a un tipo con uniforme militar y una enorme barba roja. No sé [quién era/por qué], pero parecía muy infeliz. ¿Quién parecía muy infeliz?	El tipo del uniforme	Jim
52	El pez se comió al gusano. Estaba [sabroso/hambriento]. ¿Qué estaba [sabroso/hambriento]?	El gusano	El pez
53	Estaba tratando de abrir la cerradura con la llave, pero alguien había llenado el ojo de la cerradura con chicle y no podía [entrar/sacarlo]. ¿Qué no podía [entrar/sacar]?	La llave	El chicle

54	El perro persiguió al gato, que subió corriendo a un árbol. Esperó en [arriba/abajo]. ¿Cuál esperó [arriba/abajo]?	El gato	El perro
55	En la tormenta, el árbol se cayó y se estrelló contra el techo de mi casa. Ahora, tengo que conseguirlo [sacarlo/repararlo]. ¿Qué tiene que [sacar/reparar]?	El árbol	El techo
56	El cliente entró al banco y apuñaló a uno de los cajeros. Inmediatamente fue llevado a la [sala de emergencias/comisaría]. ¿Quién fue llevado a la [sala de emergencias/comisaría]?	El cajero	El cliente
57	John estaba investigando en la biblioteca cuando escuchó a un hombre tarareando y silbando. [Estaba muy molesto/Era muy molesto]. ¿Quién [estaba molesto/era molesto]?	John	El hombre tarareando
58	John estaba trotando por el parque cuando vio a un hombre haciendo malabares con sandías. [Quedo muy impresionado/Fue muy impresionante]. ¿Quién [quedó impresionado/fue impresionante]?	John	El malabarista
59	Bob colapsó en la acera. Pronto vio que Carl venía a ayudar. Estaba muy [enfermo/preocupado]. ¿Quién estaba [enfermo/preocupado]?	Bob	Carl
60	Sam y Amy están apasionadamente enamorados, pero los padres de Amy no están contentos porque son	Los padres de Amy	Sam y Amy

	[esnobs/tienen quince años]. ¿Quiénes [son esnobs/tienen quince]?		
61	Mark le dijo a Pete muchas mentiras sobre sí mismo, que Pete incluyó en su libro. Debería haber sido más [sincero/escéptico]. ¿Quién debería haber sido más [sincero/escéptico]?	Mark	Pete
62	Joe vendió su casa y compró una nueva a unas pocas millas de distancia. Se [mudará/mudará a ella] el jueves. ¿[De/A] qué casa se mudará?	La casa vieja	La casa nueva
63	Muchas personas comienzan a leer los libros de Paul y no pueden dejar de leerlos. [Están cautivados/Son populares] porque Paul escribe muy bien. ¿Quiénes o qué son [cautivados /populares]?	Los lectores	Los libros
64	Mary sacó su flauta y tocó una de sus piezas favoritas. La ha [amado/tenido] desde que era una niña. ¿Qué ha [amado/tenido] María desde que era niña?	La pieza	La flauta
65	Sam acercó una silla al piano, pero estaba rota/o, así que tuvo que [ponerse de pie/cantar en su lugar]. ¿Qué estaba rota/o?	La silla	El piano
66	Como estaba lloviendo, llevé el periódico [sobre/en] mi mochila para mantenerla/o seca/o. ¿Qué estaba tratando de mantener seca/o?	La mochila	El periódico
67	Sara tomó prestado el libro de la biblioteca porque lo necesita para un	El libro	El articulo

	artículo en el que está trabajando. Lo [lee/escribe] cuando llega a casa del trabajo. ¿Qué [lee/escribe] Sara cuando llega a casa del trabajo?		
68	Esta mañana, Joey construyó un castillo de arena en la playa y puso una bandera de juguete en la torre más alta, pero esta tarde [una brisa/la marea] la/o derribó. ¿Qué derribó [la brisa/marea]?	La bandera	El castillo de arena
69	Jane llamó a la puerta de Susan, pero no hubo respuesta. Ella estaba [fuera/decepcionada]. ¿Quién estaba [fuera/decepcionada]?	Susan	Jane
70	Jane llamó a la puerta y Susan abrió. La invitó a [salir/pasar]. ¿Quién invitó a quién?	Jane invitó a Susan	Susan invitó a Jane
71	Sam tomó clases de francés de Adam, porque [estaba ansioso/era conocido] por hablarlo con fluidez. ¿Quién [estaba ansioso/era conocido] por hablar francés con fluidez?	Sam	Adam
72	El camino al lago estaba bloqueado, así que no pudimos [alcanzarlo/usarlo]. ¿Qué no pudimos [alcanzar/usar]?	El lago	El camino
73	El sol estuvo cubierto por una espesa nube durante toda la mañana, pero afortunadamente, cuando comenzó el picnic, ya [se había ido/estaba afuera]. ¿Qué [se había ido/estaba fuera]?	La nube	El sol
74	Fuimos al lago, porque se había visto un tiburón en la playa del océano, por lo que era un lugar [peligroso/más seguro]	La playa	El lago

	para nadar. ¿Cuál era un lugar [peligroso/más seguro] para nadar?		
75	Sam trató de pintar un cuadro de pastores con ovejas, pero terminaron pareciéndose más a [perros/golfistas]. ¿Qué parecían [perros/golfistas]?	Las ovejas	Los pastores
76	Mary metió a su hija Anne en la cama para que pudiera [dormir/trabajar]. ¿Quién va a [dormir/trabajar]?	Anne	Mary
77	Fred y Alice tenían abrigos muy cálidos, pero no eran/estaban [suficientes/preparados] para el frío de Alaska. ¿Quiénes o qué no [fue suficiente/estaban preparados] para el frío?	Los abrigos	Fred y Alice
78	Thomson visitó la tumba de Cooper en 1765. En esa fecha llevaba cinco años [muerto/viajando]. ¿Quién había estado [muerto/viajando] durante cinco años?	Cooper	Thomson
79	Jackson estuvo muy influenciado por Arnold, aunque vivió dos siglos [antes/después]. ¿Quién vivió [antes/después]?	Arnold	Jackson
80	La hija de Tom, Eva, está comprometida con el Dr. Stewart, quien es su pareja. Los dos [médicos/amantes] se conocen desde hace diez años. ¿Qué dos personas se conocen desde hace diez años?	Tom y el Dr. Stewart	Eva y el Dr. Stewart
81	No puedo cortar ese árbol con esa hacha; es demasiado [grueso/pequeña]. ¿Qué es demasiado [grueso/pequeña]?	El árbol	El hacha

82	Los zorros entran de noche y atacan a las gallinas. Tendré que [protegerlas/matarlos]. ¿Qué tengo que [proteger/matar]?	Las gallinas	Los zorros
83	Los zorros entran de noche y atacan a las gallinas. Se han vuelto/puesto muy [audaces/nerviosas]. ¿Qué se ha puesto [audaz/nerviosa]?	Los zorros	Las gallinas
84	Fred se cubrió los ojos con las manos, porque el viento estaba arrastrando arena. Los/as [abrió/bajó] cuando cesó el viento. ¿Qué [abrió/bajó] Fred?	Sus ojos	Sus manos
85	La actriz solía llamarse Terpsichore, pero lo cambió a Tina hace unos años, porque pensó que era [más fácil/demasiado difícil] de pronunciar. ¿Qué nombre fue [más fácil/demasiado difícil] de pronunciar?	Tina	Terpsichore
86	Fred miraba la televisión mientras George salía a comprar comestibles. Después de una hora [se levantó/volvió]. ¿Quién [se levantó/volvió]?	Fred	George
87	Se suponía que Fred tenía que hacer funcionar el lavavajillas, pero lo pospuso porque quería ver la televisión. Pero el programa resultó ser aburrido, así que cambió de opinión y lo/la [encendió/apagó]. ¿Qué [encendió/apagó] Fred?	El lavavajillas	La televisión
88	Fred es el único hombre vivo que recuerda a mi bisabuelo. [Es/Era] un hombre extraordinario. ¿Quién [es/era] un hombre extraordinario?	Fred	Mi bisabuelo

89	Fred es el único hombre vivo que aún recuerda a mi padre cuando era un bebé. Cuando Fred vio a mi padre por primera vez, tenía doce [años/meses]. ¿Quién tenía doce [años/meses]?	Fred	Mi padre
90	En julio, Kamtchatka declaró la guerra a Yakutsk. Dado que el ejército de Yakutsk estaba mucho mejor equipado y era diez veces más grande, [obtuvieron la victoria/fueron derrotados] en cuestión de semanas. ¿Quién [salió victorioso/fue derrotado]?	Yakutsk	Kamtchatka
91	Elizabeth trasladó su empresa de Esparta a Troya para ahorrar dinero en impuestos; los impuestos son mucho [más altos/más bajos] allí. ¿Dónde son [más altos/más bajos] los impuestos?	En Esparta	En Troya
92	Esther calcula que ahorrará costos de envío si construye su fábrica en Springfield en lugar de Franklin, porque [la mayoría/ninguno] de sus clientes vive allí. ¿En qué ciudad [vive la mayoría/no vive ninguno] de los clientes de Esther?	Springfield	Franklin
93	¡Mira! ¡Hay un [tiburón/pececillo] nadando justo debajo de ese pato! ¡Será mejor que se ponga a salvo rápido! ¿Qué necesita ponerse a salvo?	El pato	El pececillo
94	Hay demasiados ciervos en el parque, por lo que el servicio del parque trajo una pequeña manada de lobos. La población debería [aumentar/disminuir]	Los lobos	Los ciervos

	en los próximos años. ¿Qué población [aumentará/disminuirá]?		
95	Los arqueólogos han concluido que los humanos vivieron en Lauta hace 20.000 años. [Cazaban ciervos/Buscaron pruebas] en las orillas del río. ¿Quiénes [cazaban ciervos/buscaron pruebas]?	Los humanos prehistóricos	Los arqueólogos
96	Los científicos están estudiando tres especies de peces que recientemente se han encontrado viviendo en el Océano Índico. [Aparecieron/Empezaron] hace dos años. ¿Quién o qué [apareció/empezó] hace dos años?	Los peces	Los científicos
97	Los periodistas entrevistaron a las estrellas de la nueva película. Fueron muy [cooperativos/persistentes], por lo que la entrevista duró mucho tiempo. ¿Quiénes fueron [cooperativos/persistentes]?	Las estrellas	Los periodistas
98	La policía arrestó a todos los pandilleros. Estaban tratando de [dirigir/detener] el tráfico de drogas en el vecindario. ¿Quién estaba tratando de [dirigir/detener] el tráfico de drogas?	La pandilla	La policía
99	Guardé el pastel en el refrigerador. Tiene mucha/as [mantequilla/sobras]. ¿Qué tiene mucha/as [mantequilla/sobras]?	El pastel	El refrigerador
100	Sam se rompió ambos tobillos y camina con muletas. Pero dentro de un mes más o menos deberían [estar mejor/ser innecesarias]. ¿Qué deberían [estar mejor/ser innecesario]?	Los tobillos	Las muletas

101	Cuando los patrocinadores del proyecto de ley llegaron al ayuntamiento, se sorprendieron al descubrir que la sala estaba llena de opositores. Eran muy [mayoritarios/minoritarios]. ¿Quiénes estaban en [la mayoría/minoría]?	Los opositores	Los patrocinadores
102	A todos les encantaron las galletas de avena; solo a unas pocas personas les gustaron las galletas con chispas de chocolate. La próxima vez, deberíamos hacer [más/menos] de ellas. ¿Qué galleta deberíamos hacer [más/menos], la próxima vez?	Las galletas de avena	Las chispas de chocolate
103	Esperábamos colocar copias de nuestro boletín en todas las sillas del auditorio, pero simplemente [no había suficientes/había demasiadas]. ¿[Hay demasiados/No hay suficientes] de qué?	Sillas	Copias del boletín
104	Clavé un alfiler a través de una zanahoria. Cuando saqué el alfiler, [dejó/tenía] un agujero. ¿Qué [dejó/tenía] un agujero?	El alfiler	La zanahoria
105	No pude encontrar una cuchara, así que intenté usar un bolígrafo para remover mi café. Pero resultó ser una mala idea, porque se llenó de [tinta/café]. ¿Qué se llenó de [tinta/café]?	El café	El bolígrafo
106	Steve sigue el ejemplo de Fred en todo. Lo [admira/influye] enormemente. ¿Quién [admira a /influye en] quién?	Steve admira a Fred	Fred influye en Steve
107	La mesa no cabe por la puerta porque es demasiado [ancha/estrecha]. ¿Qué es demasiado [ancha/estrecha]?	La mesa	La puerta

108	Grace estaba feliz de cambiarme su suéter por mi chaqueta. Ella piensa que le queda [genial/desaliñado]. ¿Qué le queda [genial/desaliñado] a Grace?	La chaqueta	El suéter
109	Bill piensa que llamar la atención sobre sí mismo fue de mala educación [con/por parte de] Bert. ¿Quién llamó la atención sobre sí mismo?	Bill	Bert
110	John [contrató a/se contrató a sí mismo con] Bill para que lo cuidara. ¿Quién cuida a quién?	Bill cuida a John	John cuida a Bill
111	John le [prometió/ordenó] a Bill que se iría/fuera, así que una hora después se fue. ¿Quién se fue?	Juan	Bill
112	La biografía de Sam Goodman del general espartano Jenófanes transmite un sentido vívido de las dificultades que enfrentó en su [infancia/investigación]. ¿Quién enfrentó dificultades?	Jenófanes	Sam
113	La madre de Emma había muerto hacía mucho tiempo, y su [lugar/educación] había sido [ocupado por/estado a cargo de] una excelente mujer como institutriz. ¿[El lugar/La educación] de quién había sido [ocupado/gestionada]?	La madre de Emma	Emma
114	Jane llamó a la puerta de Susan pero ella/_ no [respondió/obtuvo respuesta]. ¿Quién no [respondió/obtuvo respuesta]?	Susan	Jane
115	Joe pagó al detective después de [recibir/que entregó] el informe final del	Joe	El detective

	caso. ¿Quién [recibió/entregó] el informe final?		
116	Beth no se enfadó con Sally, que la había interrumpido, porque se detuvo y [contó hasta diez/se disculpó]. ¿Quién [contó hasta diez/se disculpó]?	Beth	Sally
117	Jim le hizo una seña al cantinero y le hizo un gesto hacia [su vaso vacío/la llave del baño]. ¿[El vaso vacío/La llave del baño] de quién?	Jim	El barman
118	Dan ocupó el asiento trasero mientras que Bill ocupaba el asiento delantero porque su "¡Pido!" fue [más rápido/lento]. ¿De quién fue el "Pido" [más rápido/lento]?	Bill	Dan
119	Tom le dijo "Jaque" a Ralph mientras [tomaba/movía] su alfil. ¿De quién fue el alfil que [tomó/movió] Tom?	Ralph	Tom
120	Cuando Andrea en el fumigador aéreo pasó por encima de Susan, pudo ver la [pista/equipo] de aterrizaje. ¿Quién pudo ver la [pista/equipo] de aterrizaje?	Andrea	Susan
121	Tom llevó a Ralph a la escuela para que no tuviera que [caminar/conducir solo]. ¿Quién no tendría que [caminar/conducir solo]?	Ralph	Tom
122	Bill le pasó el plato medio vacío a John porque [estaba lleno/tenía hambre]. ¿Quién [estaba lleno/tenía hambre]?	Bill	John
123	Bill le pasó el gameboy a John porque su turno [había terminado/era el	Bill	John

	siguiente]. ¿De quién [había terminado el/era el siguiente] turno?		
124	El hombre subió/cargo al niño [a su litera/sobre sus hombros]. ¿[La litera/Los hombros] de quién?	Del niño	Del hombre
125	[Palmeando/Estirando] su espalda, la mujer le sonrió a la niña. ¿La espalda de quién [palmeó/estiró] la mujer?	De la niña	De la mujer
126	Billy lloró porque Toby no [compartía/aceptaba] su juguete. ¿Quién era el dueño del juguete?	Toby	Billy
127	Lily le habló a Donna, rompiendo su [concentración/silencio]. ¿[La concentración/El silencio] de quién?	Donna	Lily
128	Cuando Tommy dejó caer su helado, Timmy se rio, por lo que el padre lo miró con [severidad/simpatía]. ¿Quién recibió la mirada de papá?	Timmy	Tommy
129	Mientras Ollie cargaba a Tommy por los largos y sinuosos escalones, [sus piernas colgaban/ le dolían sus piernas]. [¿Las piernas de quién colgaban? / ¿A quién le dolían las piernas?]	Tommy	Ollie
130	El padre llevó al niño dormido en sus/su [brazos/cuna]. ¿[Los brazos/La cuna] de quién?	El padre	El niño
131	La mujer sostuvo a la niña contra su [pecho/voluntad]. ¿[El pecho/La voluntad] de quién?	De la mujer	De la niña
132	Los padres de Pam llegaron a casa y la encontraron teniendo relaciones sexuales con su novio, Paul. Estaban	Pam y Paul	Los padres de Pam

	[avergonzados/furiosos] por eso. ¿Quiénes estaban [avergonzados/furiosos]?		
133	El Dr. Adams le informó a Kate que [tenía cáncer/se había jubilado] y le presentó varias opciones para el tratamiento futuro. ¿Quién [tenía cáncer/se había jubilado]?	Kate	Dr. Adams
134	Dan tuvo que impedir que Bill jugara con el pájaro herido. Es muy [compasivo/cruel]. ¿Quién es [compasivo/cruel]?	Dan	Bill
135	George consiguió entradas gratis para la obra, pero se las dio a Eric [porque/a pesar de que no] estaba [particularmente/particularmente] ansioso por verla. ¿Quién [estaba/no estaba] ansioso por ver la obra?	Eric	George
136	Jane le dio dulces a Joan porque [tenía/no tenía] hambre. ¿Quién [tenía/no tenía] hambre?	Joan	Jane
137	Traté de pintar un cuadro de un huerto, con limones en los limoneros, pero salieron más como [bombillas / postes de teléfono]. ¿Qué parecían [bombillas / postes de teléfono]?	Los limones	Los limoneros
138	James le pidió un favor a Robert, pero él [se negó/fue rechazado]. ¿Quién [se negó/fue rechazado]?	Robert	James
139	Kirilov cedió la presidencia a Shatov porque era [más/menos] popular. ¿Quién era [más/menos] popular?	Shatov	Kirilov

140	Emma no le pasó la pelota a Janie aunque ella [estaba libre/vio que estaba libre]. ¿Quién [estaba libre/vio que la otra jugadora estaba libre]?	Janie	Emma
141	Joe vio a su hermano esquiar en la televisión anoche, pero el tonto no [lo reconoció/no tenía puesto un abrigo]. ¿Quién es el tonto?	Joe	El hermano de Joe
142	Puse el [pesado libro/ala de mariposa] sobre la mesa y se rompió. ¿Qué se rompió?	La mesa	El ala de mariposa
143	Madonna despidió a su entrenadora porque [se acostó con/no podía soportar a] su novio. ¿Quién [se acostó con/no podía soportar] el novio de quién?	La entrenadora se acostó con el novio de Madonna	Madonna no soportaba al novio de la entrenadora
144	Carol creía que Rebecca [sospechaba / lamentaba] que-había/haber robado el reloj. ¿Quién es sospechosa de robar el reloj? / ¿Quién robó el reloj?	Carol	Rebecca
145	Este libro presentó a Shakespeare a [Ovidio/Goethe]; fue una gran influencia en su escritura. ¿La escritura de quién fue influenciada?	Shakespeare	Goethe
146	Este libro presentó a Shakespeare a [Ovidio/Goethe]; fue una excelente selección de sus escritos. ¿Una buena selección de escritos de quién?	Ovidio	Shakespeare
147	Alice buscó a su amiga Jade entre la multitud. Como siempre [tiene buena suerte/usa un turbante rojo], Alice la vio rápidamente. ¿Quién siempre [tiene buena suerte/lleva un turbante rojo]?	Alice	Jade

148	Durante un juego de etiquetas, Ethan [persiguió a/huyo de] Luke porque él era "eso". ¿Quién era "eso"?	Ethan	Luke
149	En la competencia de Loebner, los jueces no pudieron determinar qué encuestados eran los chatbots porque eran muy [avanzados/estúpidos]. ¿Quiénes eran tan [avanzados/estúpidos]?	Los chatbots	Los jueces
150	El usuario cambió su contraseña de "GrWQWu8JyC" a "luchas con torres de sauce" ya que era fácil de [recordar/olvidar]. ¿Qué era fácil de [recordar/olvidar]?	La contraseña "GrWQWu8JyC"	La contraseña " luchas con torres de sauce "

Fuente: Adaptado de Levesque et al. (2012).

Anexo 3

Problemas de desambiguación de pronombres

Nº	Contexto/Pregunta	Alternativas	
		A	B
1	Entonces papá calculó cuánto le debía el hombre a la tienda; a eso añadió la pensión del hombre en la cabaña de los cocineros. **Él** restó esa cantidad del salario del hombre y extendió su cheque.	Papá	El hombre
2	Siempre antes, Larry había ayudado a papá con su trabajo. Pero ahora no **podía** ayudarlo, porque papá dijo que su jefe en la compañía ferroviaria no querría que nadie más que él trabajara en la oficina.	Larry	Papá
3	Siempre antes, Larry había ayudado a papá con su trabajo. Pero ahora no podía **ayudarlo**, porque papá dijo que su jefe en la compañía ferroviaria no querría que nadie más que él trabajara en la oficina.	Larry	Papá
4	Siempre antes, Larry había ayudado a papá con su trabajo. Pero ahora no podía ayudarlo, porque papá dijo que **su** jefe en la compañía ferroviaria no querría que nadie más que él trabajara en la oficina.	Larry	Papá
5	El burro deseó que desapareciera una verruga en su pata trasera, y **lo** hizo.	Burro	Verruga
6	Cuando finalmente se calmaron un poco y llegaron a casa, el Sr. Farley puso la piedra mágica en una caja fuerte de hierro. Es posible que algún día quieran **usarla**, pero realmente por ahora, ¿qué más podrían desear?	Piedra mágica	Caja fuerte

7	Los Wainwright trataron al Sr. Crowley como un príncipe hasta que hizo su testamento a su favor; luego lo trataron como basura. La gente decía que murió solo para librarse de **sus** eternas molestias.	Wainwrights	Gente
8	Varias veces Henry había estado presente en las entrevistas que su padre había tenido con destacados detectives que deseaban su ayuda para resolver misterios desconcertantes, y esas ocasiones se destacaron como días especiales para **él**.	Henry	Padre
9	¿Qué hay de la vez que cortaste bulbos de tulipán en las hamburguesas porque pensabas que **eran** cebollas?	Bulbos de tulipán	Hamburguesas
10	Nadie se une a Facebook para estar triste y solo. Pero un nuevo estudio del psicólogo George Lincoln de la Universidad de Wisconsin sostiene que eso es exactamente **lo** que nos hace sentir.	Facebook	Estudio
11	Igualmente deslumbrante es C.K. Dexter Haven, un joven dandi pálido que sostiene un bastón con mango de jade, con un poodle dormido a **sus** pies.	Haven	Poodle
12	Lionel tiene cautivo a un científico, el Dr. Vardi, que ha inventado un dispositivo que vuelve invisibles a los animales; Lionel planea usarlo en Geoffrey y **enviarlo** a robar material nuclear de una bóveda del ejército.	Lionel	Dr. Vardi
13	Larry, un adolescente tímido, vive con su madre viuda en un proyecto de vivienda de Brooklyn. El padre de Larry, líder de una pandilla, fue asesinado a tiros; el discípulo de su padre, Antonio, toma a Larry bajo **su**	Larry	Antonio

	protección y rápidamente lo convierte en un traficante de drogas.		
14	La pradera de Dakota estaba tan cálida y brillante bajo el sol resplandeciente que no parecía posible que alguna vez **hubiera** sido barrida por los vientos y las nieves de ese duro invierno.	La pradera	El sol
15	No es fácil espaciar los ojales exactamente a la misma distancia, y es muy difícil cortarlos con precisión del tamaño correcto. El más pequeño deslizamiento de las tijeras hará que el agujero sea demasiado grande, e incluso un hilo sin cortar **lo** dejará demasiado pequeño.	El tamaño adecuado	El agujero
16	Los dueños se quedaron en la ciudad para administrar sus tiendas y vivían en las habitaciones detrás de **estas**.	Dueños	Tiendas
17	Incluso antes de llegar a la ciudad, pudieron escuchar un sonido como el estallido de maíz. Dora preguntó qué **era** y papá dijo que eran petardos.	Ciudad	Sonido
18	Todos los botones de la espalda del vestido a cuadros de Dora estaban abotonados de afuera hacia adentro. Maude debería haber pensado en abotonarla; pero no, había dejado que la pobrecita Dora hiciera lo mejor que **pudiera**, sola.	Dora	Maude
19	Bernard, que no le había dicho al funcionario del gobierno que **tenía** menos de 21 años cuando presentó un reclamo de vivienda, no consideró que hubiera hecho nada deshonesto. Aun así, cualquiera que supiera que tenía 19 años podría quitarle su reclamo.	Bernard	El funcionario del gobierno

20	Bernard, que no le había dicho al funcionario del gobierno que tenía menos de 21 años cuando presentó un reclamo de vivienda, no consideró que **hubiera** hecho nada deshonesto. Aun así, cualquiera que supiera que tenía 19 años podría quitarle su reclamo.	Bernard	El funcionario del gobierno
21	Bernard, que no le había dicho al funcionario del gobierno que tenía menos de 21 años cuando presentó un reclamo de vivienda, no consideró que hubiera hecho nada deshonesto. Aun así, cualquiera que supiera que tenía 19 años podría quitarle **su** reclamo.	Bernard	El funcionario del gobierno
22	Los políticos lejanos en Washington no podían conocer a los colonos por lo que **deben** hacer reglas para regularlos.	Políticos	Colonos
23	Los políticos lejanos en Washington no podían conocer a los colonos por lo que deben hacer reglas para **regularlos**.	Políticos	Colonos
24	Los hombres tenían derecho a que sus hijos trabajaran para ellos hasta **los** 21 años.	Hombres	Hijos
25	Dado que Chester dependía del tío Vernon, no **podía** casarse sin su aprobación.	Chester	Tío Vernon
26	Dado que Chester dependía del tío Vernon, no podía casarse sin **su** aprobación.	Chester	Tío Vernon
27	Me senté allí sintiéndome como un tipo sobre el que una vez había leído en un libro, que asesinó a otra persona y escondió el cuerpo debajo de la mesa del comedor, y luego tuvo que ser la vida y el alma de una cena, con **eso** ahí todo el tiempo.	Libro	Cuerpo
28	El Sr. Taylor era un hombre de temperamento inseguro y su tendencia general era pensar que	Sr. Taylor	David

	David era un pobre tonto y que cualquier paso que diera en cualquier dirección por su propia cuenta era solo otra prueba de **su** idiotez innata.		
29	Salí a buscar un poco de comida, más para pasar el tiempo que porque **quisiera**.	Comida	Tiempo
30	El Sr. Moncrieff visitó el lujoso apartamento de Chester en Nueva York, pensando que pertenecía a su hijo Edward. El resultado fue que el Sr. Moncrieff decidió cancelar la asignación de Edward con el argumento de que ya no **necesita** su apoyo financiero.	Sr. Moncrieff	Edward
31	El Sr. Moncrieff visitó el lujoso apartamento de Chester en Nueva York, pensando que pertenecía a su hijo Edward. El resultado fue que el Sr. Moncrieff decidió cancelar la asignación de Edward con el argumento de que ya no necesita **su** apoyo financiero.	Sr. Moncrieff	Chester
32	Mamá se acercó y se sentó junto a Alice. Suavemente **le** acarició el cabello y dejó que la niña llorara.	Mamá	Alicia
33	Mamá se acercó y se sentó junto a Alice. Suavemente le acarició **el** cabello y dejó que la niña llorara.	Mamá	Alicia
34	Alice estaba quitando el polvo de la sala y tratando de encontrar el botón que mamá había escondido. Hoy no hay tiempo para mirar fotos antiguas en **su** álbum de fotos favorito. Hoy tuvo que buscar un botón, así que puso el álbum en una silla sin siquiera abrirlo.	Alicia	Mamá

35	Alice estaba quitando el polvo de la sala y tratando de encontrar el botón que mamá había escondido. Hoy no hay tiempo para mirar fotos antiguas en su álbum de fotos favorito. Hoy **tuvo** que buscar un botón, así que puso el álbum en una silla sin siquiera abrirlo.	Alicia	Mamá
36	Alice estaba quitando el polvo de la sala y tratando de encontrar el botón que mamá había escondido. Hoy no hay tiempo para mirar fotos antiguas en su álbum de fotos favorito. Hoy tuvo que buscar un botón, así que puso el álbum en una silla sin siquiera **abrirlo**.	Sala de estar	Botón
37	Papá miró los rostros de los niños, tan desconcertados y tristes ahora. Ya era bastante malo que se **les** tuviera que negar tantas cosas porque no podía pagarlas.	Niños	Caras
38	Papá miró los rostros de los niños, tan desconcertados y tristes ahora. Ya era bastante malo que se les tuviera que negar tantas cosas porque no podía **pagarlas**.	Niños	Cosas
39	Todos los días, después de la cena, el Sr. Schmidt tomaba una larga siesta. Mark lo dejaba dormir durante una hora, luego lo despertaba, lo regañaba y lo ponía a trabajar. **Necesitaba** que terminara su trabajo, porque su trabajo era hermoso.	Sr. Schmidt	Mark
40	Todos los días, después de la cena, el Sr. Schmidt tomaba una larga siesta. Mark lo dejaba dormir durante una hora, luego lo despertaba, lo regañaba y lo ponía a trabajar.	Sr. Schmidt	Mark

	Necesitaba que terminara **su** trabajo, porque su trabajo era hermoso.		
41	Los letreros sobre las puertas de las tiendas tenían imágenes que indicaban qué trabajo se hacía en el interior. Aunque cada vez más personas estaban aprendiendo a leer, cada artesano todavía tenía letreros, no deseando perder a un posible patrón simplemente porque **era** analfabeto.	Artesano	Patrón
42	Mark quedó absorto en Blaze, el caballo blanco. **Tenía** miedo de que los trabajadores en Burlington Stables lo golpearan y lo intimidaran porque era tímido, por lo que asumió la alimentación y el cuidado del animal.	Mark	Blaze
43	Mark quedó absorto en Blaze, el caballo blanco. Tenía miedo de que los trabajadores en Burlington Stables lo golpearan y lo intimidaran porque era tímido, por lo que **asumió** la alimentación y el cuidado del animal.	Mark	Blaze
44	Mark estaba cerca de los talones del Sr. Singer. **Lo** escuchó llamar al capitán, prometiéndole, en la jerga que todos hablaban esa noche, que nada debería dañarse en el barco excepto solo las municiones, pero que el capitán y toda su tripulación harían mejor en quedarse en la cabina hasta que terminara el trabajo.	Mark	Sr. Singer
45	Mark estaba cerca de los talones del Sr. Singer. Lo **escuchó** llamar al capitán, prometiéndole, en la jerga que todos hablaban esa noche, que nada debería dañarse en el	Mark	Sr. Singer

	barco excepto solo las municiones, pero que el capitán y toda su tripulación harían mejor en quedarse en la cabina hasta que terminara el trabajo.		
46	Mark estaba cerca de los talones del Sr. Singer. Lo escuchó llamar al capitán, prometiéndole, en la jerga que todos hablaban esa noche, que nada debería dañarse en el barco excepto solo las municiones, pero que el capitán y toda **su** tripulación harían mejor en quedarse en la cabina hasta que terminara el trabajo.	Mark	Capitán
47	Mark escuchó los pies de Steve bajando la escalera. La puerta de la tienda se cerró tras **él**. Corrió a mirar por la ventana.	Mark	Steve
48	Mark escuchó los pies de Steve bajando la escalera. La puerta de la tienda se cerró tras él. **Corrió** a mirar por la ventana.	Mark	Steve
49	De una cosa Mark estaba seguro. Harry sabía mucho menos que **él**.	Mark	Harry
50	Así que Mark durmió. Era de día cuando despertó con la mano de Warren sobre **su** hombro.	Mark	Warren
51	Al darse la vuelta en su litera superior, Tatyana podía mirar por encima del borde y ver claramente a su madre. ¡Qué pequeña, erguida y rígida **yacía** en la litera de abajo! Tenía los ojos cerrados, pero Tatyana dudaba si dormía.	Tatyana	Madre
52	Al darse la vuelta en su litera superior, Tatyana podía mirar por encima del borde y ver claramente a su madre. ¡Qué pequeña,	Tatiana	Madre

	erguida y rígida yacía en la litera de abajo! Tenía los ojos cerrados, pero Tatyana dudaba si **dormía**.		
53	Cuando Tatyana llegó a la cabaña, su madre estaba durmiendo. **Tuvo** cuidado de no molestarla, se desvistió y volvió a subir a su litera.	Tatyana	Madre
54	Cuando Tatyana llegó a la cabaña, su madre estaba durmiendo. Tuvo cuidado de no **molestarla**, se desvistió y volvió a subir a su litera.	Tatyana	Madre
55	Cuando Tatyana llegó a la cabaña, su madre estaba durmiendo. Tuvo cuidado de no molestarla, se desvistió y volvió a subir a **su** litera.	Tatiana	Madre
56	Tatyana manejaba dos guitarras y una bolsa, y aún podía señalar a los Freeman: "¿No es bueno que **hayan** venido, mamá?"	Dos guitarras y una bolsa	Los Freeman
57	Tatyana sabía que a la abuela siempre le gustaba servir abundante comida a sus invitados. Ahora Tatyana observó cómo la abuela abrazaba a la pequeña madre de Tatyana en un amplio y escuálido abrazo y luego **la** empujaba hacia la mesa, le quitaba el chal de los hombros, la sentaba en el lugar de honor y decía simplemente: "Hay mucho".	Tatiana	Abuela
58	La mesa estaba repleta de comida y, en el suelo, junto a **esta**, había vasijas, cestas y un cubo de leche de cinco cuartos.	Mesa	Alimento
59	Grant trabajó duro para cosechar sus frijoles para que él y su familia tuvieran suficiente para comer ese invierno. Su amigo Henry le	Frijoles	Grant y su familia

	permitió **apilarlos** en su granero donde se secarían. Más tarde, él y Tatyana los descascararían y los cocinarían para sus cenas dominicales.		
60	Grant trabajó duro para cosechar sus frijoles para que él y su familia tuvieran suficiente para comer ese invierno. Su amigo Henry le permitió apilarlos en su granero donde se secarían. Más tarde, él y Tatyana los descascararían y los cocinarían para **sus** cenas dominicales.	Frijoles	El y Tatiana

Fuente: Adaptado de Morgenstern et al. (2016).

Anexo 4

Proceso de *fine-tuning* del modelo de lenguaje

Se detallará el proceso de *fine-tuning* para el modelo DeBERTaV3 (He et al., 2022) en el conjunto de datos SQuAD (Rajpurkar et al., 2016) para la respuesta a preguntas extractiva mediante la plataforma en la nube de Google Colab, los siguientes pasos fueron tomados de la documentación.

Primero hay que asegurarse de contar con las siguientes librerías instaladas:

```
pip install transformers datasets
```

Cargar el conjunto de datos SQuAD

Iniciamos cargando un subconjunto de datos de SQuAD más pequeño desde la librería *Datasets*. Esto da la oportunidad de experimentar y asegurarse que todo funciona antes de gastar más tiempo en el entrenamiento en el conjunto de datos completo.

```
from datasets import load_dataset
squad = load_dataset("squad", split="train[:5000]")
```

Se divide el conjunto de datos en conjuntos de entrenamiento y prueba con el método *train_test_split*:

```
squad = squad.train_test_split(test_size=0.2)
```

Para poder ver un ejemplo:

```
squad["train"][0]
```

En el cual se pueden visualizar los siguientes campos:

- *Answers*: la ubicación inicial del token de la respuesta y el texto de la respuesta.
- *Context*: información del contexto del cual el modelo necesita extraer la respuesta.
- *Question*: la pregunta que el modelo debe responder.

Preprocesamiento de datos

El siguiente paso es cargar el *tokenizador* de DeBERTaV3 para procesar los campos de pregunta y contexto:

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("microsoft/mdeberta-v3-base
")
```

Hay pocos pasos de preprocesamiento en particular para la tarea de respuesta a preguntas que se deben conocer:

1. Algunos ejemplos en el conjunto de datos pueden tener un contexto muy grande que excede el máximo tamaño de entrada del modelo. Para secuencias más largas, se trunca el contexto con la siguiente configuración: *truncation="only_second"*.

2. Luego, se traza la posición inicial y final de la respuesta en el contexto original al establecer la configuración: *return_offset_mapping=True*.
3. Con el trazado en mano, ahora se pueden encontrar los tokens iniciales y finales de la respuesta. Se utiliza el método *sequence_ids* para encontrar que parte del desplazamiento pertenece a la pregunta y cual al contexto.

A continuación, se muestra cómo se puede crear una función para truncar y trazar los tokens iniciales y finales de la respuesta al contexto:

```
def preprocess_function(examples):
    questions = [q.strip() for q in examples["question"]]
    inputs = tokenizer(
        questions,
        examples["context"],
        max_length=384,
        truncation="only_second",
        return_offsets_mapping=True,
        padding="max_length",
    )
    offset_mapping = inputs.pop("offset_mapping")
    answers = examples["answers"]
    start_positions = []
    end_positions = []
    for i, offset in enumerate(offset_mapping):
        answer = answers[i]
        start_char = answer["answer_start"][0]
        end_char = answer["answer_start"][0] + len(answer["text"][0])
        sequence_ids = inputs.sequence_ids(i)
        # Encuentra el inicio y el final del contexto.
        idx = 0
        while sequence_ids[idx] != 1:
            idx += 1
        context_start = idx
        while sequence_ids[idx] == 1:
            idx += 1
        context_end = idx - 1
        # Si la respuesta no está completamente dentro del contexto,
etiquétela (0, 0)
        if offset[context_start][0] > end_char or
offset[context_end][1] < start_char:
            start_positions.append(0)
            end_positions.append(0)
        else:
            # De lo contrario, son las posiciones de token inicial y
final.
```

```
            idx = context_start
            while idx <= context_end and offset[idx][0] <= start_char:
                idx += 1
            start_positions.append(idx - 1)
            idx = context_end
            while idx >= context_start and offset[idx][1] >= end_char:
                idx -= 1
            end_positions.append(idx + 1)
    inputs["start_positions"] = start_positions
    inputs["end_positions"] = end_positions
    return inputs
```

Para aplicar la función de preprocesamiento sobre todo el conjunto de datos se utiliza la función *map*. Se puede acelerar la función *map* al configurar *batched=True* para procesar múltiples elementos del conjunto de datos a la vez. Se elimina cualquier columna que no se necesita con el siguiente código:

```
tokenized_squad = squad.map(preprocess_function, batched=True,
remove_columns=squad["train"].column_names)
```

Ahora se crea un conjunto de ejemplos utilizando *DefaultDataCollator*:

```
from transformers import DefaultDataCollator
data_collator = DefaultDataCollator()
```

Entrenamiento

Ahora ya podemos afinar el modelo, cargamos el modelo *DeBERTaV3* con *AutoModelForQuestionAnswering*:

```
from transformers import AutoModelForQuestionAnswering,
TrainingArguments, Trainer
model =
AutoModelForQuestionAnswering.from_pretrained("microsoft/mdeberta-v3-
base")
```

En este punto, solo quedan tres pasos:

1. Definir *hyperparameters* en *TrainingArguments*. El único parámetro requerido es *output_dir* el cual especifica donde se va a guardar modelo.
2. Para pasarle los argumentos de entrenamiento a *Trainer* junto con el modelo, el conjunto de datos, *tokenizer* y *data collator*.
3. Se llama a función *train()* para afinar el modelo.

```
training_args = TrainingArguments(
    output_dir="mi_modelo_rp",
    evaluation_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
```

```
    num_train_epochs=3,
    weight_decay=0.01,
    push_to_hub=True,
)
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_squad["train"],
    eval_dataset=tokenized_squad["test"],
    tokenizer=tokenizer,
    data_collator=data_collator,
)
trainer.train()
```

Se adjunta el enlace a la plataforma de Google Colab para la evaluación del agente inteligente:

https://colab.research.google.com/drive/1wXFag3hTeFJM2vOAy1eCpJ13da7AnDe8?usp=sharing

En las páginas 24 y 25 se detalla el funcionamiento del mismo.

Anexo 5

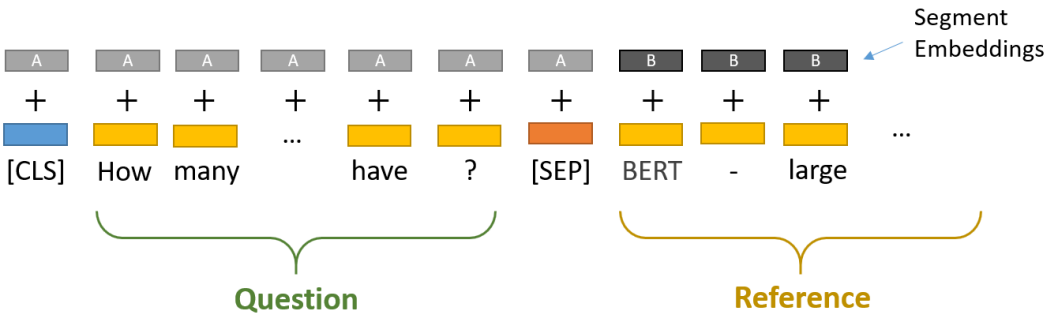
Cómo responde preguntas el modelo de lenguaje

Formato de entrada de BERT

Para la tarea de respuestas a preguntas en *BERT*, se empaquetan tanto la pregunta y el texto de referencia al input.

Figura 3

Formato de entrada de BERT



Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

Fuente: McCormick, 2020.

Las dos partes del texto son separadas por un token especial [SEP].

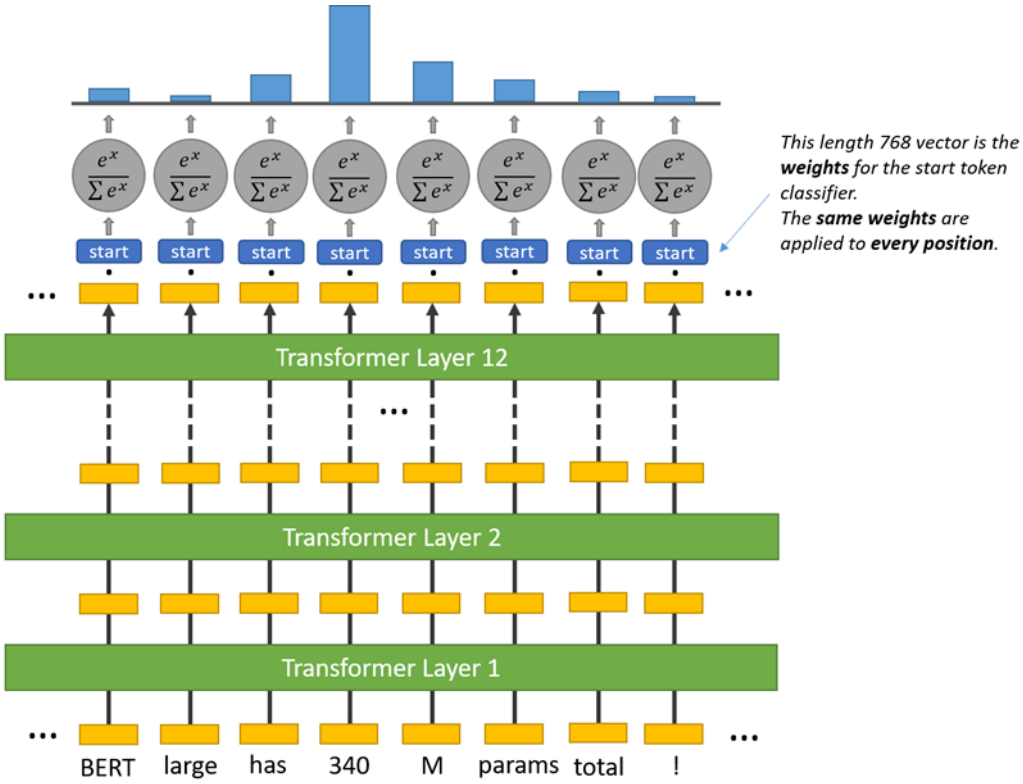
BERT también utiliza "*Segment Embeddings*" para diferenciar la pregunta del texto de referencia. Estos simplemente dos *embeddings* para el segmento A y B que BERT aprendió, y en el cual agrega un token de embeddings antes de alimentarlo en la capa de entrada.

Clasificadores de Token de inicio y fin

BERT necesita resaltar una etiqueta de texto que contiene la respuesta, esto se representa prediciendo cual token marca el inicio de la respuesta y cual token marca el final.

Figura 4

Clasificadores de tokens de inicio y fin



Fuente: Fuente: McCormick, 2020.

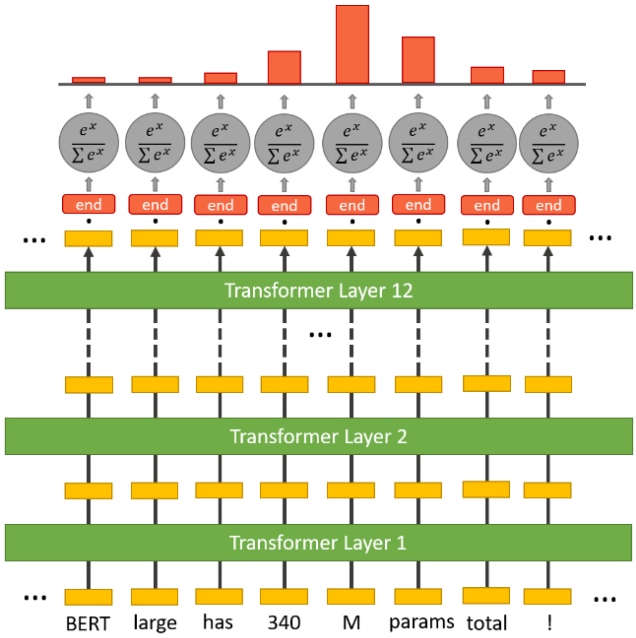
Para cada token en el texto, se ingresa su embedding final al inicio del clasificador de token. El clasificador de token de inicio solo tiene un solo conjunto de pesos el cual es aplica a cada palabra.

Luego de tomar el producto entre los embeddings de salida y los pesos iniciales, se aplica la activación *softmax* para producir una distribución de probabilidad sobre todas las palabras. La palabra que tenga la probabilidad más alta de ser el token inicial es la que se elige.

Se repite este proceso para el token final

Figura 5

Proceso para el token final



Fuente: McCormick, 2020.