UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESINA DE LICENCIATURA

Programa de Apoyo al Egreso para Alumnos con Práctica Profesional Supervisada

TÍTULO: Un enfoque para la detección de pares atributo-valor en descripciones en lenguaje natural en el contexto de la oferta inmobiliaria

AUTOR/A: Bazzana Tanevitch, Luciana

DIRECTOR/A ACADÉMICO: Torres, Diego

DIRECTOR/A PROFESIONAL: Del Río, Juan Pablo

CODIRECTOR/A ACADÉMICO: Fernández, Alejandro

RESUMEN

La información estructurada es un recurso muy valioso para la construcción de sistemas de información. La transformación de datos no estructurados en datos estructurados puede ser automatizada, sin embargo, el procesamiento de textos humanos requiere el uso de técnicas de Procesamiento de Lenguaje Natural (NLP). Este estudio tiene como objetivo explorar y evaluar diversos enfoques para la extracción automática de pares atributo-valor a partir de descripciones de avisos inmobiliarios. El propósito final es enriquecer la base de datos del Observatorio de Valores del Suelo de la Provincia de Buenos Aires. El observatorio busca resolver la carencia de información en esta materia y facilitar la participación pública en la valorización de inmuebles.

Palabras Claves

Natural Language Processing, Extracción Atributo-Valor, Observatorio de Valores del Suelo

Trabajos Realizados

Se evalúan tres enfoques que utilizan NLP para la extracción de pares atributo-valor a partir de textos del dominio inmobiliario.

Matching basado en reglas implicó la declaración de patrones sintácticos para la extracción de variables y sus valores.

El entrenamiento de un modelo de reconocimiento de entidades requirió definir una estrategia para la anotación de datos manualmente, tarea realizada en conjunto con expertos en el dominio.

Se utilizaron modelos pre-entrenados basados en la arquitectura transformers en dos de sus aplicaciones: Question-Answering y conversacional.

La performance se mide utilizando las métricas de precisión, recall y f1-score, teniendo en cuenta los valores obtenidos por cada una de las variables que se desea detectar.

Conclusiones

El enfoque de matching basado en reglas no requiere entrenamiento previo ni etiquetado manual de datos. Sin embargo, su debilidad radica en su sesgo a los datos y su vulnerabilidad a la variabilidad del lenguaje natural, requiriendo ajustes constantes en los patrones. Por otro lado, entrenar un modelo NER es costoso al necesitar grandes volúmenes de datos anotados. A pesar de obtener excelentes resultados, se espera mejoría con un entrenamiento adicional. En contraste, los modelos basados en arquitectura transformers están pre-entrenados en un inmenso conjunto de datos, siendo GPT-3 el que ofrece mayor performance en concordancia con ser el que más datos utilizó en su entrenamiento. Los enfoques aplicados proporcionan una solución al problema de extracción de pares atributo-valor para el enriquecimiento del observatorio.

Trabajos Futuros

Queda como trabajo futuro alinear los pares extraídos acorde a la ontología que estructura la base de conocimiento donde se almacenan los datos del observatorio. Además, se pueden elaborar reglas para mejorar la performance de los enfoques aplicados, y así obtener resultados superiores.

Fecha de la presentación: MARZO 2024

Agradecimientos

A mi mamá, un pilar fundamental en mi vida. Su inquebrantable apoyo, contención y amor han sido la fuerza impulsora detrás de cada logro en mi vida.

A mi director, por alentarme y guiarme en los momentos de duda.

Al centro de investigación LIFIA, por brindarme un espacio donde descubrí mi vocación por la investigación.

Finalmente, mi sincero agradecimiento a cada persona que me brindó su apoyo a lo largo de este trayecto.

Un enfoque para la detección de pares atributo-valor en descripciones en lenguaje natural en el contexto de la oferta inmobiliaria

Luciana Tanevitch[0000−0002−5322−9314]

LIFIA, CICPBA-Facultad de Informática, UNLP, Argentina
{luciana.tanevitch}@lifia.info.unlp.edu.ar

Resumen La información estructurada es un recurso muy valioso para la construcción de sistemas de información. La transformación de datos no estructurados en datos estructurados puede ser automatizada, sin embargo, el procesamiento de textos humanos requiere el uso de técnicas de Procesamiento de Lenguaje Natural (NLP). Este estudio tiene como objetivo explorar y evaluar diversos enfoques para la extracción automática de pares atributo-valor a partir de descripciones de avisos inmobiliarios. El propósito final es enriquecer la base de datos del Observatorio de Valores del Suelo de la Provincia de Buenos Aires. El Observatorio busca resolver la carencia de información en esta materia y facilitar la participación pública en la valorización de inmuebles.

Palabras clave: Natural Language Processing · Extracción Atributo-Valor · Observatorio de Valores del Suelo

1. Introducción

Cada día, miles de usuarios publican productos en venta, comparan precios en diferentes portales y realizan pedidos. Tener datos estructurados es fundamental para la construcción de sistemas de recomendación, categorización y comparación de productos, ya que permiten automatizar procesos. Como la mayoría de la información disponible en la Web no está estructurada, es necesario emplear técnicas que posibiliten la extracción y organización de estos datos. En este sentido, las ontologías y grafos de conocimiento son herramientas útiles que permiten modelar la información en un formato fácilmente interpretable por máquinas [25]. Los portales inmobiliarios no son ajenos a esta característica de la Web, ya que generalmente no poseen metadatos para la extracción automática de datos. Los avisos publicados en estos sitios suelen incluir información tabulada donde se resumen ciertas características del inmueble y también una descripción libre que permite al anunciante agregar más información. Este trabajo se desarrolla en el marco del proyecto "Observatorio de valores del suelo e instrumentos de financiamiento del desarrollo urbano", que involucra la construcción de un Observatorio Inmobiliario (OI) a partir de técnicas de recolección automáticas en la Web. El OI es una iniciativa de articulación del sistema científico-técnico y

el sector público provincial, cuyo propósito es relevar, sistematizar y producir información georreferenciada de valores inmobiliarios y promover el desarrollo de instrumentos de gestión de suelo urbano y participación pública en la valorización inmobiliaria [3].

La construcción del OI implica la creación de una base de conocimiento estructurada mediante una ontología del dominio inmobiliario que modela avisos, inmuebles y sus características, entre otros. Para alimentar la base de conocimiento se utilizan técnicas para la extracción de características de inmuebles a partir de los avisos publicados en portales inmobiliarios. Los avisos incluyen información tabulada, tal como se ve en la Figura 1, que puede extraerse mediante *web scrapers* acorde a lo desarrollado en [13]. Pero los textos escritos en lenguaje natural implican cierta dificultad para las máquinas, ya que éstas no tienen la capacidad de interpretarlo de manera directa. En particular, la descripción de un inmueble puede contener información relevante que no está presente de manera tabulada y que es deseable extraer. Para permitir a las máquinas interpretar el lenguaje humano existen las técnicas de Procesamiento de Lenguaje Natural (NLP) [15].

Los primeros sistemas de extracción de atributos usaban reglas y patrones construidos a mano, basándose en la estructura sintáctica, dependencias y palabras específicas. Luego, el avance en aprendizaje automático permitió descubrir nuevos horizontes [12]. La extracción de pares atributo-valor implica identificar y asociar características específicas que describen un objeto con sus correspondientes valores en conjuntos de datos.

Superficie total	440 m^2
Superficie cubierta	260 m^2
Ambientes	4
Dormitorios	2
Baños	1

Figura 1. Información tabulada de un inmueble

El objetivo de este trabajo es hallar un enfoque para la extracción de pares atributo - valor a partir de textos escritos en lenguaje natural, en particular para el enriquecimiento y verificación de los datos del OI. Por ejemplo, dado un inmueble del OI con atributo «`dirección: Buenos Aires 4564`»y una descripción libre, se desea extraer de la descripción un valor para ese atributo que permita verificar la información disponible, o añadir mayor exactitud. Se evaluará el rendimiento de tres enfoques distintos basados en NLP para la extracción de

características en el dominio inmobiliario: *rule-based matching*, modelos basados sobre la arquitectura *transformers* y un modelo de reconocimiento de entidades.

El trabajo se organiza de la siguiente manera. La Sección 2 presenta una revisión de trabajos relacionados al presente, teniendo en cuenta el NLP para la extracción de atributos en el dominio del comercio electrónico. La Sección 3 desarrolla el contexto del OVS, la herramienta en la cual se enmarca el trabajo y en la cual se realizará el aporte. La Sección 5 presenta al NLP como una solución al problema de interpretación automática de textos. La Sección 6 presenta tres enfoques posibles para la extracción de características en textos mediante el uso de técnicas de NLP. En la Sección 7 se desarrolla la aplicación de cada uno de los enfoques al OVS, cuya performance se evaluará acorde a las métricas definidas en la Sección 8. En la Sección 9 se exponen los resultados de cada uno de los enfoques, y finalmente, en la Sección 10 se le da un cierre el trabajo exponiendo fortalezas y debilidades de los enfoques.

2. Trabajos relacionados

La extracción de pares atributo-valor es una tarea relevante para la normalización de datos y la construcción de bases de datos. Utilizando NLP es posible procesar textos no estructurados para la extracción de características que permitan poblar una base de conocimiento estructurada por una ontología. Una amplia variedad de trabajos se encargan de abordar el problema de extracción de pares atributo-valor utilizando diferentes técnicas. Anantharangachar et. al. [4] escribieron patrones para la extracción de características a partir de textos para poblar una ontología. Mediante el uso de NLP, definen patrones para la detección de valores para los atributos definidos por la ontología que aparezcan en el texto. Pham y Pham [17] construyen un sistema basado en reglas para la extracción de características a partir de anuncios de portales inmobiliarios vietnamitas.

Blandón y Zapata [6] presentan una serie de patrones sintácticos implementados con la herramienta GATE de NLP, para poblar una ontología automáticamente. Linková y Gurský [16] proponen un método para la extracción de atributos y sus valores a partir de descripciones de productos con el fin de tener datos estructurados. Para los datos de tipo booleano, sugieren que la aparición de la característica en el texto implica que tiene el valor verdadero. Los datos numéricos son detectados con patrones de valor y unidad de medida. Las cadenas son detectadas con *matching* exacto y luego se busca el valor en el texto acorde a los valores esperados conocidos para ese atributo. Para este último caso sugieren el uso de herramientas incluidas en NLP como una mejora a su propuesta. Baur et. al. [5] evalúan distintos modelos de aprendizaje automático para la valuación de inmuebles a partir de la extracción de características en descripciones escritas en lenguaje natural, utilizando NLP. Huynh et. al. [14] etiquetaron datos manualmente para el entrenamiento de un modelo de reconocimiento de entidades en anuncios inmobiliarios. Sabeh et. al. [21] construyeron una herramienta llamada CAVE que permite corregir atributos existentes y enriquecer la

información disponible de avisos en plataformas e-commerce. CAVE se basa en el enfoque Question-Answering, para lo cual cada atributo es tratado como una pregunta. Su modelo está entrenado en un corpus específico de e-commerce construido a partir de títulos y descripciones de la plataforma Amazon. De manera similar, Wang et. at. [26] proponen AVEQA, un modelo de Question-Answering construido sobre BERT para la identificación de extraer pares atributo-valor a partir de descripciones de productos. Además, su enfoque tiene la capacidad de clasificar preguntas irrespondibles a partir del contexto.

Probst et. al. [18] entrenaron un modelo para extraer pares atributo-valor en descripciones de productos con el objetivo de realizar *data augmentation* en bases de datos de este dominio. Se basan en el algoritmo co-EM con Naïve Bayes para la clasificación de los conceptos identificados en atributo o valor. Para relacionar el atributo con su valor correspondiente usaron un *dependency parser*. Finalmente, la intervención humana permite corregir los resultados arrojados por el modelo. IDEALO [1] es un software de comparación de precios de productos online. Este trabajo utiliza soluciones basadas en BERT como una mejora frente a las basadas en reglas, para extraer automáticamente atributos numéricos de las descripciones de productos para enriquecer la base de conocimiento existente. Brinkmann et. al. [7] utiliza ChatGPT para la extracción de atributos y valores de descripciones de productos. Para esto compara el rendimiento frente a diferentes diseños de input, teniendo la posibilidad de que el modelo responda "no lo sé" en caso que la respuesta no esté presente en el contexto. Estos diseños de input pueden ser similares a los utilizados en Question-Answering para responder preguntas simples, o bien puede indicarse la tarea de extracción en un formato determinado para la generación de la respuesta. El trabajo además compara este enfoque con Question-Answering y Named Entity Recognition.

3. El Observatorio de Valores del Suelo

Un Observatorio Inmobiliario (OI) es un sistema que almacena datos del mercado inmobiliario en una base georreferenciada. Un OI proporciona información actualizada y precisa sobre tendencias, precios, oferta y demanda de inmuebles en una ubicación específica. Esto permite a profesionales del sector tomar decisiones informadas sobre transacciones inmobiliarias, inversiones y desarrollo de proyectos en función de la situación actual del mercado. En este sentido dos organismos de la Provincia de Buenos Aires, OPISU y CIC, presentaron el proyecto "Observatorio de valores del suelo e instrumentos de financiamiento del desarrollo urbano" donde el laboratorio LIFIA participa en su construcción. El objetivo general de este proyecto es generar información pública para contribuir al financiamiento urbano mediante instrumentos de recuperación de plusvalías urbanas [10]. El Observatorio de Valores del Suelo (OVS) tiene como objetivo cuantificar la valorización inmobiliaria producida por las acciones del Estado, y a partir de esos datos se busca aportar a la formulación e implementación de políticas destinadas a mejorar las condiciones habitacionales en sectores populares. [2].

La base de conocimiento del OVS se construyó a partir de la extracción automática de características a partir de avisos inmobiliarios en diferentes plataformas de e-commerce, almacenándose en formato RDF bajo una ontología inmobiliaria que le da soporte [13]. La importancia de la estructuración de los datos está relacionada a la automatización de procesos. Si la información se encuentra disponible de manera estructurada, es posible crear procesos que utilicen esos datos sin necesidad de intervención humana. Por esta razón, resulta imprescindible la tarea de reconocer atributos y valores que permitan completar o verificar la información almacenada.

3.1. Lenguaje Natural en Avisos Inmobiliarios

Mientras que para información tabulada los datos son procesados automáticamente, cuando se trata de textos escritos en lenguaje natural la tarea reviste mayor complejidad ya que se debe hallar una mención que represente un atributo de un inmueble en el texto, y luego buscar su valor asociado. La Figura 2 muestra una instancia de un inmueble de la base de conocimiento. Algunos elementos tales como las medidas de las habitaciones y baño, y las coordenadas fueron extraídos a partir de información tabulada. La descripción es almacenada de manera textual acorde al aviso publicado. Allí aparecen características que no están estructuradas tales como dirección, medidas del terreno, valor de expensas y tasa ABL.

4. Problema

A partir de un conjunto de descripciones de avisos inmobiliarios extraídos de la base de conocimiento que dispone el organismo, un equipo de expertos en el dominio decide las variables que desean detectar, en función del impacto que éstas representan para el OI. Se define como variable a un atributo que caracteriza a un inmueble en la vida real. Las variables se definen por el equipo de expertos en el dominio acorde a sus necesidades, detallando por cada una la descripción de lo que ésta representa, las formas de aparición más frecuentes de esa variable en el texto, el tipo de dato (numérico, texto, booleano), y qué es lo que necesitan identificar de cada variable. El equipo de expertos en el dominio definió ocho variables de máxima prioridad. Esto significa que si bien hay más variables que pueden ser analizadas, las primeras ocho son las de mayor interés en esta etapa del proyecto. A continuación se describe cada una de ellas.

- **Dirección.** El OI cuenta con un campo *address* y uno *description*. Se desea detectar incongruencias o completar el campo *address* a partir de la descripción cuando no tenga un valor asignado. La dirección puede estar presente en las descripciones en diferentes formatos: (1) calle y altura (Independencia 1239), (2) calle e intersección (Independencia e Industria), (3) calle y entre calles (Independencia e/ Industria y Edgar Aschieri), (4) por nombre de barrio y/o lote en el cual se ubica la oferta al interior de un condominio (Barrio

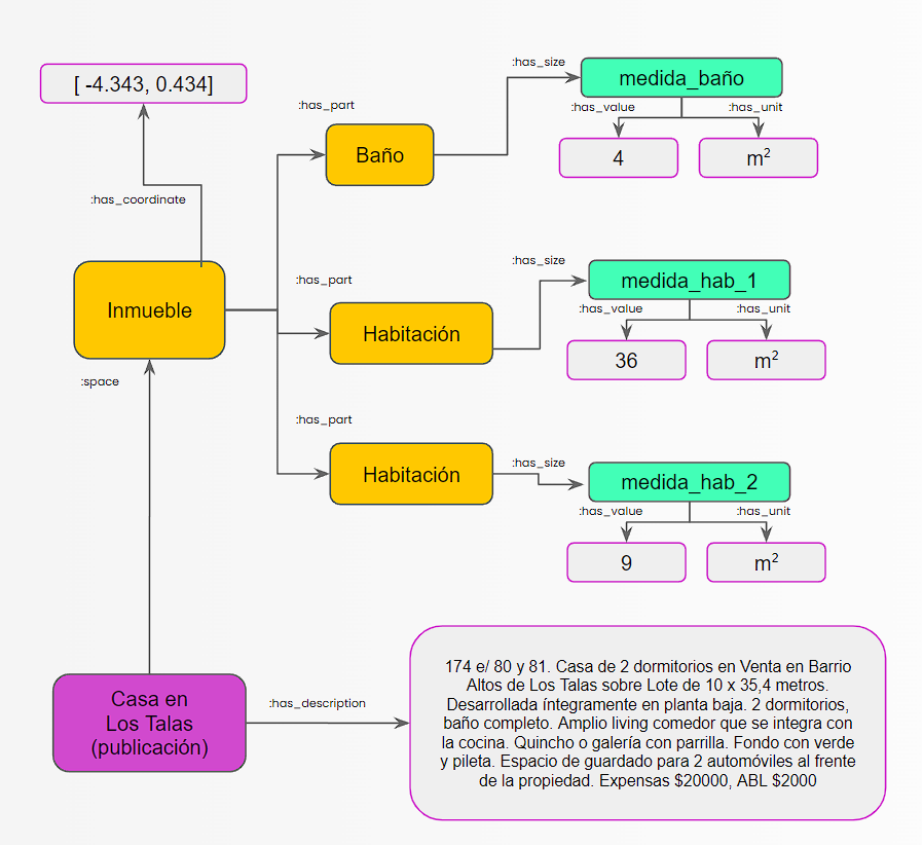

Figura 2. Instancia de un inmueble

Grand Bell, Lote 57), (5) otros tipo de situaciones que no se reflejan en las categorías anteriores (Ruta 15 km 12).

- **Factor de Ocupación Total (FOT).** El factor de ocupación total, FOT, es un indicador del potencial constructivo del terreno en altura. Se requiere identificar el valor numérico de FOT que Figura en la descripción, ya que la enunciación del mismo en los avisos es un indicador de zonas con alta dinámica inmobiliaria y de que el oferente intenta capitalizar una normativa urbana que permite un potencial constructivo mayor que en otros lugares. El FOT puede aparecer escrito de distintas maneras (texto completo con o sin mayúsculas y, más frecuentemente con sigla: FOT, F.O.T, Fot, fot, F.o.t, etc.). Puede estar expresado con un punto o una coma. Por otra parte, en algunas ocasiones este indicador puede ser variable o tener más de un valor alternativo para un mismo inmueble según ciertas circunstancias, motivo por el cual deberían guardarse como variables estructuradas cada una de las posibilidades o valores que presenta este indicador.
- **Lote irregular.** La variable hace referencia a la forma del lote. Se considera un lote regular cuando es rectangular o cuadrado y cuando no cumple esa condición se considera que es un lote irregular, pudiendo tener diferentes formas. Los lotes irregulares suelen tener menor precio que los regulares dado que complejizan el aprovechamiento constructivo de los productos inmobiliarios que se desarrollan sobre el lote. Si bien las descripciones suelen enunciar la palabra "irregular", bien podría referirse a una forma que no es rectangular (por ejemplo: forma de martillo, triangular, etc.). Es una variable booleana ya que el terreno puede ser o no irregular.
- **Medidas del terreno.** Hace referencia a las longitudes de frente y fondo del lote donde se emplaza el inmueble.
- **Esquina.** Indica si el lote se encuentra en una esquina. Es una variable booleana, ya que el lote puede estar o no en una esquina.
- **Barrio.** Se refiere al nombre del barrio en el que se encuentra el inmueble, exclusivamente tratándose de clubes de campo, countries y barrios privados.
- **Cantidad de frentes.** La variable indica la cantidad de frentes de un inmueble. Esta situación se aclara expresamente cuando un lote posee salida a más de una calle. Se expresa generalmente por un texto que enuncie una cantidad numérica (dos, 2, doble) y la palabra "frentes".
- **Pileta.** Indica si el lote posee o no piscina.

5. El Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (NLP) es una rama de la Inteligencia Artificial que permite a las máquinas interpretar el lenguaje humano. El NLP puede clasificarse en Comprensión de Lenguaje Natural (NLU) y Generación de Lenguaje Natural (NLG). NLU permite a las máquinas entender textos escritos en lenguaje natural mediante la extracción de conceptos, entidades, emociones, etc. NLG es una técnica que permite a las máquinas generar textos que sean entendibles y tengan sentido para los humanos [15]. El NLP tiene un amplio uso

en técnicas para la extracción de respuestas de un texto, bots conversacionales y detección de entidades. En este trabajo se analizarán diferentes enfoques que utilizan NLP en su construcción.

6. Enfoques de Extracción

6.1. *Matching* Basado en Reglas

El NLP permite a las máquinas procesar textos. En este sentido, cada unidad lingüística recibe el nombre de token y cada token tiene asociadas etiquetas gramaticales y de dependencias sintácticas, que valen dentro del texto donde aparecen. Gracias a estas etiquetas se pueden definir patrones que son útiles para identificar automáticamente secuencias en un texto. Estos patrones se pueden construir como listas de diccionarios, donde cada diccionario describe las características de un token. En este contexto, los patrones no solo reconocen secuencias de caracteres, como lo harían las expresiones regulares, sino que además incorporan la riqueza lingüística para identificar coincidencias basadas en atributos que posea el token, como por ejemplo características sintácticas, morfológicas, categorías gramaticales, etc.

Partiendo del dominio inmobiliario en el que se enmarca este trabajo, un patrón para identificar direcciones en el formato `nombre calle` al `altura` (ej. Independencia al 4600) puede ser [{'POS': 'PROPN', 'OP': '+'}, {'LOWER': 'al', 'OP': '?'}, {'LIKE_NUM': True}]. El POS 'PROPN' es para representar sustantivos propios. Las calles que llevan nombre entran en esta categoría. El operando '+' implica la presencia de uno o más tokens continuos de este tipo, por ejemplo 'Salta', 'Buenos Aires', 'Domingo Faustino Sarmiento'. Luego del nombre de la calle puede aparecer opcionalmente la contracción 'al' y finalmente el número. De esta manera ese patrón permite reconocer direcciones como 'Salta al 1300', 'Salta 1356', 'Buenos Aires al 4000', etc. Sin embargo, este patrón no es adecuado para identificar direcciones con calles numéricas como 'Av. 7 al 2500', porque '7' no es un sustantivo propio sino un número. Entonces debería crearse un nuevo patrón [{'POS': 'PROPN', 'OP': '+'}, {'LIKE_NUM': True} {'LOWER': 'al', 'OP': '?'}, {'LIKE_NUM': True}]. Ahora cubriríamos direcciones como 'Av. 7 al 2500', 'Av. 7 2535', 'Diag. 73 3504'. Pero surge un nuevo caso: 'Av. 7 n° 2535' no sería identificado por el patrón. Para que lo sea, se requiere modificar el patrón para que en vez de 'al' matchee con 'n°', 'nro', 'num' y todas sus variantes. De esta manera, los patrones son creados acorde al conjunto de datos sobre el que se trabaja, prestando atención a las formas sintácticas que toma cada variable.

6.2. Modelo de Reconocimiento de Entidades (NER)

NER es una técnica de NLP que permite extraer conceptos y categorizarlos según etiquetas predefinidas [22]. La implementación se realiza generalmente con modelos de aprendizaje supervisado, requiriendo datos etiquetados. Si bien

los datos etiquetados a menudo se obtienen de conjuntos de datos disponibles en la Web, en casos donde la disponibilidad de datos en un dominio e idioma específico es limitada, es necesario crear un conjunto de datos etiquetado para el entrenamiento del modelo. Para crear un conjunto de datos de entrenamiento para un modelo NER es necesario: i) definir las etiquetas a utilizar, ii) seleccionar el conjunto de datos a etiquetar y iii) elaborar la estrategia de etiquetado para que la anotación sea homogénea, es decir que haya consistencia en la forma en que se aplican las etiquetas a lo largo de todo el conjunto de datos. Esto significa que si se decide que cierta información corresponde a una etiqueta específica, esa decisión debe aplicarse de manera uniforme cada vez que se encuentre información similar en el conjunto de datos. Esta homogeneidad en la anotación es crucial para entrenar un modelo de aprendizaje automático de manera efectiva, ya que permite al modelo aprender patrones consistentes a partir de los datos etiquetados.

6.3. Transformadores

Los transformadores son una arquitectura de redes neuronales eficiente para el manejo de secuencias de datos basado en el mecanismo de atención que permite capturar de manera óptima las dependencias entre palabras [24]. BERT [11] es un modelo pre-entrenado implementado con transformadores que tiene la capacidad de ser adaptado fácilmente para diferentes tareas de procesamiento de lenguaje natural, como Question-Answering (QA). Los modelos QA permiten generar la respuesta a una pregunta a partir de un contexto dado. Un punto a tener en cuenta en QA es la posibilidad de que la respuesta a la pregunta no esté en el contexto, y en cuyo caso el modelo debe tener la capacidad de responder "no lo sé".

Otra tecnología en pleno auge son los modelos conversacionales basados en mecanismos de atención. La arquitectura de `transformers` tiene buenos resultados para esta aplicación [19]. GPT-3 [8] es un modelo conversacional pre-entrenado con la capacidad de aprender a partir de la interacción mediante *reinforcement learning*. Una fortaleza de GPT-3 en comparación con los otros enfoques radica en su capacidad para interpretar texto y generar respuestas. En situaciones que involucran variables booleanas, donde en otros enfoques se asocia la presencia del atributo con un valor positivo, GPT-3 tiene la capacidad de generar una respuesta con un valor booleano basado en el procesamiento del texto recibido.

7. Aplicando los Enfoques de Extracción al OI

7.1. *Matching* Basado en Reglas

Tal como se explicó en la Sección 6.1, definir patrones que logren la cobertura de la mayoría de los casos requiere en primer lugar conocer la manera en la que se escribe frecuentemente cada una de las variables. Para el desarrollo de este

enfoque se utiliza la herramienta SpaCy[1], con su modelo de tamaño grande en lenguaje español[2]. Dado que la dirección tiene múltiples formatos posibles, se escribieron varios patrones para cada uno de esos formatos. Luego, todos ellos serán utilizados en conjunto para detectar direcciones. Dado que el FOT se asocia generalmente a un valor numérico, se utilizó un patrón de dependencias para hallar el modificador numérico del token que representa el FOT. Además, para detectar aquellos casos donde el FOT toma varios valores, se definió un patrón que busca la palabra "FOT" escrita en alguna de las maneras posibles, y la presencia de algún indicador de variación (por ejemplo, "fot residencial"). En el caso de las dimensiones del lote, se pretende encontrar coincidencias en el formato `numero x numero`, con o sin unidades de medida. El reconocimiento del nombre de un barrio cerrado involucra una mayor complejidad. Se buscará la mención de la palabra "barrio" seguida de una secuencia de tokens con POS PROPN, dado que representaría el nombre del barrio. Para variables booleanas (como detectar la irregularidad del terreno, detectar si está ubicado en una esquina y la presencia o no de pileta en el lote) la presencia de ciertas palabras claves relacionadas al atributo en el texto se consideran valores positivos. Por ejemplo, la presencia de la palabra "esquina" en el texto implicaría que el lote está ubicado en una esquina.

El Cuadro 1 resume los patrones para cada variable, y ejemplos que esos patrones pueden reconocer.

7.2. Modelo de Reconocimiento de Entidades (NER)

Acorde a la Sección 6.2, para entrenar un modelo NER es necesario generar un conjunto de datos etiquetados y para ello se proponen tres tareas que se describen en profundidad a continuación.

i) Definición de las etiquetas a utilizar

- **Dirección.** Dada la diversidad de formatos presentados para la dirección, se define una etiqueta por cada una de esas categorías. (1) DIR_CALLE_ALTURA, (2) DIR_INTERSECCION, (3) DIR_ENTRE, (4) DIR_LOTE, (5) DIR_OTROS.
- **FOT.** Se define la etiqueta FOT para identificar el o los valores de FOT presentes.
- **Lote irregular.** Se define la etiqueta IRREGULAR para la detección de tokens que hagan referencia a la irregularidad del terreno así como formas no rectangulares.
- **Medidas del terreno.** Se define la etiqueta DIMENSIONES para identificar las medidas de frente y fondo del lote.
- **Esquina.** Se define la etiqueta ESQUINA para la detección de tokens que hagan referencia a la ubicación del lote en una esquina.

[1] https://spacy.io
[2] https://spacy.io/models/es#es_core_news_lg

Variable	Patrón	Ejemplos de *matching*
Dirección (calle y altura)	[{'LOWER': {'IN': ['calle', 'avenida', 'av', 'diagonal', 'diag']}, 'OP':'?'}, {'TEXT': '.', 'OP':'?'}, {'POS': 'PROPN', 'OP': '+'}, {'LOWER': 'n'}, {'TEXT': '°'}, {'LIKE_NUM': True}] [{'LOWER': {'IN': ['calle', 'avenida', 'av', 'diagonal', 'diag']}, 'OP':'?'}, {'TEXT': '.', 'OP':'?'}, {'POS': 'PROPN', 'OP': '+'}, {'LOWER': 'al', 'OP':'?'}, {'LIKE_NUM': True}]	Av. Manuel Belgrano al 6200 Diag. 73 nro° 3450 Alberti 3359
Dirección (intersección)	[{'LOWER': {'IN': ['calle', 'avenida', 'av', 'diagonal','diag']}, 'OP':'?'}, {'TEXT': '.', 'OP':'?'}, {'POS': {'IN': ['PROPN', 'NUM']}, 'OP': '+'}, {'LOWER': {'IN': ['y', 'esquina', 'esq', 'e']}}, {'LOWER': {'IN': 'calle', 'avenida', 'av', 'diagonal','diag']}, 'OP':'?'}, {'TEXT': '.', 'OP':'?'}, {'POS': {'IN': ['PROPN', 'NUM']}, 'OP': '+'},	calle Alsina y Av. San Martín calle 19 y calle 36 7 y 50
Dirección (entre calles)	[{'LOWER': {'IN': 'calle', 'avenida', 'av', 'diagonal','diag']}, 'OP':'?'}, {'TEXT': '.', 'OP':'?'}, {'POS': {'IN': ['PROPN', 'NUM']}, 'OP': '+'}, {'LOWER': {'IN': ['e/', 'entre', 'e']}}, {'LOWER': {'IN': 'calle', 'avenida', 'av', 'diagonal','diag']}, 'OP':'?'}, {'TEXT': '.', 'OP':'?'}, {'POS': {'IN': ['PROPN', 'NUM']}, 'OP': '+'}, {'LOWER': 'y'}, {'LOWER': {'IN': 'calle', 'avenida', 'av', 'diagonal','diag']}, 'OP':'?'}, {'TEXT': '.', 'OP':'?'}, {'POS': {'IN': ['PROPN', 'NUM']}, 'OP': '+'},	calle 7 e/ 71 y 72 calle 618 e/ 6 bis y 7
Dirección (lote en barrio cerrado)	[{'LOWER': 'lote'}, {'POS': 'IN': ['NUM','PROPN'] }]	lote 24, lote A22
FOT	[{'RIGHT_ID': 'fot', 'RIGHT_ATTRS': {'LOWER': {'IN':['fot', 'f.o.t']}}}, {'LEFT_ID': 'fot', 'REL_OP': '>', 'RIGHT_ID': 'num', 'RIGHT_ATTRS': {'DEP': 'nummod'}} [{'LOWER': {'IN':['fot', 'f.o.t']}}, {'LOWER': {'IN':['res', 'residencial', 'comercial', 'com', 'industrial']}, 'OP': '?'}, {'IS_PUNCT': True, 'OP':'?'}, {'LIKE_NUM': True}]	FOT 1 F.O.T 3,6 FOT residencial: 2.5 FOT comercial: 3
Irregular	[{'LEMMA': 'irregular'},	Terreno irregular de ...
Medidas del terreno	[{'LIKE_NUM': True}, {'LOWER': {'IN': ['mts','m','metros']}, 'OP':'?'}, {'LOWER': 'x'}, {'LIKE_NUM': True}, {'LOWER': {'IN': ['mts','m','metros']}, 'OP':'?'}, {'LOWER': 'x', 'OP':'?'}, {'LIKE_NUM': True, 'OP':'?'}, {'LOWER': {'IN': ['mts','m','metros']}, 'OP':'?'}, {'LOWER': 'x', 'OP':'?'}, {'LIKE_NUM': True, 'OP':'?'}, {'LOWER': {'IN': ['mts','m','metros']}, 'OP':'?'},	8.66 x 26 8.66 x 26 m 17,32 mts x 26 mts
Esquina	[{'LOWER': 'esquina'},	Lote en importante esquina...
Barrio	[{'LOWER': 'barrio'}, {'POS':'PROPN', 'OP': '+"}	Barrio Grand Bell
Frentes	[{'RIGHT_ID': 'frentes', 'RIGHT_ATTRS': {'LOWER': {'IN': ['frentes', 'frente']}}}, {'LEFT_ID': 'frentes', 'REL_OP': '>', 'RIGHT_ID': 'num', 'RIGHT_ATTRS': {'DEP': 'nummod'}}]	2 frentes tres frentres
Pileta	[{'LOWER': {'IN': ['piscina', 'pileta']}},	El lote posee pileta

Cuadro 1. Patrones para cada variable

- **Barrio.** Es un campo de tipo texto, para el cual se define la etiqueta NOMBRE_BARRIO.
- **Cantidad de frentes.** Se anota con la etiqueta CANT_FRENTES.
- **Pileta.** Se define la etiqueta PILETA para la detección de tokens que hagan referencia a la existencia de pileta en el lote.

El Cuadro 2 resume las etiquetas definidas para cada una de las variables que el equipo de expertos desea detectar.

Variable	**Etiqueta**
Dirección	(1) DIR_CALLE_ALTURA (2) DIR_INTERSECCION (3) DIR_ENTRE (4) DIR_LOTE (5) DIR_OTROS
Factor de Ocupación Total	FOT
Lote irregular	IRREGULAR
Medidas del terreno	DIMENSIONES
Esquina	ESQUINA
Barrio	NOMBRE_BARRIO
Cantidad de frentes	CANT_FRENTES
Pileta	PILETA

Cuadro 2. Resumen de etiquetas definidas para cada variable

ii) Selección del conjunto de datos a etiquetar Los datos se obtienen a partir de la base de conocimiento que dispone el organismo para el OI, acorde a lo detallado en la Sección 3, generando un documento de texto con descripciones de avisos inmobiliarios, que es el campo "description" del aviso en el grafo de conocimiento. Ese documento es fraccionado aleatoriamente en varios documentos más pequeños, para poder otorgarle una porción reducida del total a cada persona involucrada en el proceso de etiquetado.

iii) Elaboración de la estrategia de etiquetado Para llevar a cabo la anotación de descripciones de inmuebles, un grupo de expertos del dominio se reúne periódicamente en sesiones presenciales de aproximadamente dos horas. Cada persona cuenta con una computadora con navegador web y conexión a Internet. Antes de comenzar, se distribuyen archivos de datos únicos para cada persona, así como un archivo de etiquetas que es común para todos los participantes. Las etiquetas provistas son las definidas en el Cuadro 2. El etiquetado se realiza con la herramienta de anotación online NER Annotator for SpaCy[3], dado que se utiliza SpaCy para el entrenamiento del modelo. Cada persona debe importar su

[3] https://tecoholic.github.io/ner-annotator/

archivo de datos y luego el archivo de etiquetas. Las descripciones de inmuebles le serán mostradas en el entorno de a una, debiendo anotar en cada una de ellas las características que detecte. En la Figura 3 se puede ver una descripción de aviso inmobiliario en el entorno de la herramienta de anotaciones, con todas las etiquetas disponibles para el etiquetado y dos anotaciones realizadas correspondientes a dirección, y frente y fondo del lote. El proceso se repite hasta completar el documento, y una vez logrado la persona descarga el archivo que contiene las anotaciones, y lo envía por correo electrónico a una dirección indicada.

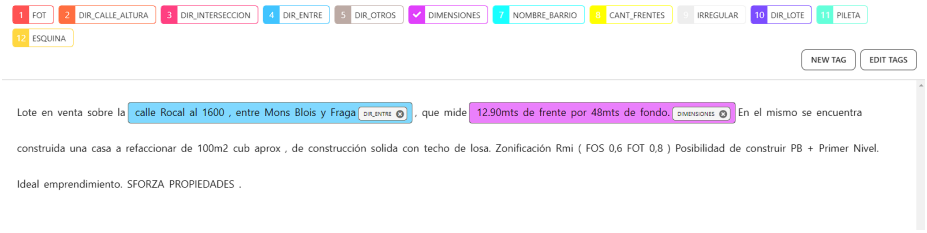

Figura 3. Etiquetado usando NER Annotator for SpaCy

Una vez generado el conjunto de datos etiquetados, se procede con el entrenamiento del modelo. En este trabajo se utiliza el modelo SpaCy de tamaño grande en lenguaje español. Una vez realizado el entrenamiento, el modelo tiene la capacidad de clasificar un nuevo ejemplo. El modelo se entrenó sobre 485 datos, que si bien es una cantidad muy pequeña de datos se satisface la cantidad mínima de ejemplos necesarios de cada una de las clases, de acuerdo a las métricas arrojadas por el verificador que posee la herramienta.

7.3. Transformadores

Los modelos seleccionados para el enfoque QA son implementaciones basadas en BERT con soporte en lenguaje español:

1. mrm8488/bert-base-spanish-wwm-cased-finetuned-spa-squad2-es
2. timpal0l/mdeberta-v3-base-squad2
3. rvargas93/distill-bert-base-spanish-wwm-cased-finetuned-spa-squad2-es

Todos los modelos están publicados en la plataforma HuggingFace[4]. Cada una de las implementaciones difiere en algún punto. (1) BETO es una implementación de BERT entrenada en grandes conjuntos de datos en lenguaje español [9] (2) DeBERTa es un modelo que mejora los modelos BERT y RoBERTa utilizando *disentangled attention* y *enhanced mask decoder*. En particular mDeBERTa es la versión multilenguaje. (3) También es una implementación sobre

[4] https://huggingface.co

el modelo BETO pero tiene la particularidad de que utiliza técnicas de destilación (*distillation*) para reducir la complejidad del modelo original, manteniendo un rendimiento aceptable. Además, para garantizar la capacidad del modelo de responder "no lo sé" cuando la respuesta no esté presente, los modelos seleccionados fueron entrenados sobre el dataset SQuAD2 [20], que incluye preguntas irrespondibles dado un contexto.

Dado un conjunto de características deseables y un conjunto de descripciones de avisos inmobiliarios, se generan preguntas para la extracción del valor de cada característica tomando como contexto la descripción del inmueble a analizar:

- ¿Cuál es la dirección del inmueble?
- ¿Cuál es el valor del FOT?
- ¿El terreno es irregular?
- ¿Cuáles son las dimensiones del lote?
- ¿El lote está en una esquina?
- ¿En qué barrio privado está ubicado?
- ¿Cuántos frentes tiene el inmueble?
- ¿El inmueble tiene pileta?

La librería `transformers`[5] cuenta con una interfaz llamada `pipeline`[6] que permite abstraer la implementación y simplificar el uso de los modelos publicados en la plataforma HuggingFace. `Pipeline` recibe como entrada un conjunto de preguntas y un texto donde se buscarán las respuestas. Retorna una respuesta para cada pregunta, en caso que ésta se encuentre en el texto. Así, es posible utilizar los modelos sin profundizar en su funcionamiento. Se espera que los modelos tengan la capacidad de detectar los valores para cada característica, o se abstengan de dar una respuesta en caso que no esté presente.

8. Metodología de Evaluación

8.1. Datos

A partir de descripciones de avisos inmobiliarios extraídos del OI, se construye manualmente un conjunto de verdad (*ground truth*) sobre el cual se evaluará cada uno de los enfoques. El conjunto de datos posee 74 registros, correspondientes a 74 publicaciones reales de avisos inmobiliarios en las cuales se detallan los valores de cada uno de los atributos que están presentes en la descripción. Los avisos no necesariamente contienen menciones de todas las variables a detectar, aquellas que no estén presentes se marcan como vacío. Por cada atributo a detectar en cada descripción se anota el valor extraído del texto para el caso de valores numéricos y cadenas, y para las variables booleanas se anota con `true` en caso que esa característica esté presente. La construcción supervisada de este conjunto permite garantizar diversidad en la forma de escritura, es decir, para la variable FOT pueden hallarse anotaciones que indiquen "F.O.T 1, fot 2,5, FOT

[5] `https://huggingface.co/docs/transformers/index`
[6] `https://huggingface.co/docs/transformers/main_classes/pipelines`

4.1", y también el caso de variaciones por ejemplo "FOT comercial: 2, FOT residencial: 1.5". Se medirá el rendimiento de los enfoques sobre cada una de las variables definidas en la Sección 4.

8.2. Métricas

Las métricas a utilizar para evaluar los enfoques son: precisión, recall y f1-score.

La precisión permite conocer la capacidad del modelo de identificar pares atributo-valor correctamente y se calcula en base a la Fórmula 1, donde VP representa aquellos pares reconocidos correctamente por la herramienta y FP los casos en los que el modelo identificó pares atributo-valor que en realidad no estaban presentes.

El recall mide la capacidad del modelo de identificar la totalidad de los pares correctos, y se calcula acorde a 2, donde VP representa aquellos pares reconocidos correctamente por la herramienta y FN aquellas menciones que deberían haber sido reconocidas y no lo fueron.

El f1-score se calcula con 3 y es una medida armoniosa que combina la precisión y el recall.

$$\text{Precisión} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}} \tag{3}$$

Dado que algunas de las variables toman valores textuales, se utiliza *matching* parcial para evaluar la similitud entre la predicción obtenida y el valor esperado. El *matching* parcial es una técnica que permite hallar coincidencias de acuerdo a una cota determinada de similitud, y no marcando coincidencias exactas. Si la similitud entre la predicción y el esperado es superior a una cota fijada, el ejemplo se computa como exitoso. En caso contrario, se contará como fallido. De esta manera, VP son aquellos valores predichos que coinciden en más de 90 % con el valor esperado. FP implica una respuesta incorrecta, lo que puede ocurrir porque la similitud entre el esperado y el predicho es inferior al 90 %, o porque el modelo no arroja respuesta cuando se esperaba obtener una respuesta. FN implica que el modelo emite una respuesta, cuando se esperaba obtener una respuesta nula.

9. Resultados

9.1. *Matching* Basado en Reglas

Para la evaluación de resultados, se aplicaron en simultáneo todos los patrones definidos para cada una de las variables a detectar. En particular para

el caso de la dirección, que tiene varios patrones asociados, es posible hallar una superposición cuando hay una coincidencia con intersecciones y entre calles. Por ejemplo, para el caso que la dirección sea "avenida 51 e/ calle 7 y calle 8", el patrón DIR_INTERSECCION identificará "calle 7 y calle 8" mientras que DIR_ENTRE revelará "avenida 51 e/ calle 7 y calle 8". Por esta razón, se define como regla que frente a múltiples coincidencias se tomará como resultado el de mayor longitud, ya que eso implicaría la extracción de la dirección más completa posible. El Cuadro 3 resume los resultados de las métricas para cada una de las variables.

Variable	**Precisión**	**Recall**	**F1-Score**
Dirección	0.33	0.95	0.49
FOT	0.89	0.89	0.89
Lote irregular	1.0	0.57	0.72
Medidas del terreno	0.8	0.42	0.55
Esquina	1.0	1.0	1.0
Barrio	0.44	0.25	0.32
Cantidad de frentes	0.5	0.5	0.5
Pileta	1.0	0.95	0.97

Cuadro 3. Resultados por variable, enfoque *Matching* Basado en Reglas

9.2. Modelo de Reconocimiento de Entidades (NER)

Luego de que el modelo sea entrenado, se utiliza sobre el conjunto del *ground truth* para identificar los valores de cada variable. El Cuadro 4 muestra las métricas obtenidas para cada una de las variables.

Variable	**Precisión**	**Recall**	**F1-Score**
Dirección	0.79	0.63	0.7
FOT	0.96	0.87	0.91
Lote irregular	0.91	0.78	0.84
Medidas del terreno	0.78	0.62	0.69
Esquina	1.0	0.92	0.96
Barrio	1.0	0.55	0.7
Cantidad de frentes	0.7	0.7	0.7
Pileta	1.0	0.8	0.88

Cuadro 4. Resultados por variable, enfoque NER

9.3. Transformadores

El Cuadro 5 muestra los resultados de performance de cada uno de los modelos QA enumerados en la Sección 7.3. Cada columna muestra los resultados de una variable. En cada columna se resalta en rosa al modelo con mejor performance para esa variable.

modelo	Direccion	FOT	Irregular	Medidas	Esquina	Barrio	Frentes	Pilleta
BETO	p: 0.12 r: 0.66 f1: 0.20	p: 0.36 r: 0.65 f1: 0.46	p: 0.7 r: 0.5 f1: 0.58	p: 0.4 r: 0.81 f1: 0.53	p: 0.41 r: 0.92 f1: 0.57	p: 0.33 r: 0.2 f1:0.25	p: 0.11 r: 1 f1: 0.21	p: 0.56 r: 0.9 f1: 0.69
mDeBERTa	p: 0.30 r: 0.85 f1: 0.44	p: 0.67 r: 0.76 f1: 0.71	p: 0.57 r: 0.57 f1: 0.57	p: 0.5 r: 0.97 f1: 0.65	p: 0.43 r: 1.0 f1: 0.6	p: 0.47 r: 0.44 f1: 0.45	p: 0.01 r: 1 f1: 0.03	p: 0.86 r: 1 f1: 0.93
BETO + distilled	p: 0.22 r: 0.82 f1: 0.35	p: 0.23 r: 0.61 f1: 0.33	p: 0.81 r: 0.64 f1: 0.72	p: 0.34 r: 0.92 f1: 0.5	p: 0.29 r: 0.92 f1: 0.44	p: 0.33 r: 0.27 f1: 0.3	p: 0.1 r: 0.87 f1: 0.19	p: 0.56 r: 0.65 f1: 0.6

Cuadro 5. Resultados por variable, enfoque *transformers* usando modelos QA

Los resultados arrojados por GPT-3 se muestran en el Cuadro 6.

Variable	Precision	Recall	F1 Score
Dirección	0.58	0.96	0.72
FOT	0.93	1.0	0.96
Lote irregular	0.81	0.92	0.86
Medidas del terreno	0.74	1.0	0.85
Esquina	1.0	0.92	0.96
Barrio	0.77	0.73	0.75
Cantidad de frentes	1.0	0.63	0.77
Pileta	1.0	1.0	1.0

Cuadro 6. Resultados por variable, enfoque transformers usando GPT-3

9.4. Resultados generales

Para la identificación de direcciones, GPT-3 tuvo la mejor performance con un f1 de 0.72, seguido por NER con 0.7. En cuanto al FOT, GPT-3 obtuvo el mayor puntaje con un f1 de 0.96, seguido de NER con 0.91. Para identificar las medidas de frente y fondo del terreno, GPT-3 tuvo los mejores resultados con f1 de 0.85 seguido de NER con 0.69. En la identificación del nombre del barrio privado, GPT-3 obtuvo resultados superiores con un f1 de 0.75 seguido de NER con 0.7. La cantidad de frentes fue detectada con un f1 de 0.77 por GPT-3, seguido por NER con f1 de 0.7. Para variables booleanas como la irregularidad del terreno, la situación de esquina y la existencia de pileta, *rule-based matching*, NER

y QA resuelven la situación como un procesamiento de secuencias. Dado que no realizan inferencias sobre el resultado, sino que extraen la mención en el texto, si se asocia la presencia del atributo como una ocurrencia positiva, pueden ocurrir interpretaciones incorrectas. Por ejemplo, si el texto menciona "el inmueble tiene un shopping en la esquina", estos enfoques podrían identificar incorrectamente la ubicación en la esquina como positiva, aunque no sea el caso. El mejor puntaje f1 para detectar la irregularidad del terreno fue 0.86, obtenido por GPT-3. NER obtuvo el siguiente mejor f1, con valor 0.84. *Matching* Basado en Reglas obtuvo los mejores números para detectar si la propiedad está ubicada en una esquina. Finalmente, GPT-3 obtuvo 1.0 en f1 para identificar si un inmueble posee pileta o no, seguido de *rule-based matching* con 0.97.

Este trabajo evalúa tres enfoques para la extracción automática de pares atributo-valor, obteniendo resultados satisfactorios. Si bien en esta evaluación los mejores resultados fueron obtenidos por GPT-3, NER también demostró un desempeño excelente incluso con un conjunto de datos limitado. Estos resultados podrían ser superiores al entrenar el modelo sobre un conjunto de datos de gran tamaño. El Cuadro 7 muestra la comparación de resultados que cada modelo tuvo para cada variable, donde se resalta en rosa el de mayor performance para cada variable. El repositorio con la implementación y resultados está disponible públicamente[7].

Enfoque	**Dirección**	**FOT**	**Lote Irregular**	**Medidas**	**Esquina**	**Barrio**	**Cantidad Frentes**	**Pileta**
Rule based matching	P: 0.33 R: 0.95 F1: 0.49	P: 0.89 R: 0.89 F1: 0.89	P: 1.0 R: 0.57 F1: 0.72	P: 0.8 R: 0.42 F1: 0.55	P: 1.0 R: 1.0 F1: 1.0	P: 0.44 R: 0.25 F1: 0.32	P: 0.5 R: 0.5 F1: 0.5	P: 1.0 R: 0.95 F1: 0.97
NER	P: 0.79 R: 0.63 F1: 0.7	P: 0.96 R: 0.87 F1: 0.91	P: 0.91 R: 0.78 F1: 0.84	P: 0.78 R: 0.62 F1: 0.69	P: 1.0 R: 0.92 F1: 0.96	P: 1.0 R: 0.55 F1: 0.7	P: 0.7 R: 0.7 F1: 0.7	P: 1.0 R: 0.8 F1: 0.88
Transformers (BETO)	P: 0.12 R: 0.66 F1: 0.2	P: 0.36 R: 0.65 F1: 0.46	P: 0.7 R: 0.5 F1: 0.58	P: 0.4 R: 0.81 F1: 0.53	P: 0.41 R: 0.92 F1: 0.57	P: 0.33 R: 0.2 F1: 0.25	P: 0.11 R: 1.0 F1: 0.21	P: 0.56 R: 0.9 F1: 0.69
Transformers (mDeBERTa)	P: 0.3 R: 0.85 F1: 0.44	P: 0.67 R: 0.76 F1: 0.71	P: 0.57 R: 0.57 F1: 0.57	P: 0.5 R: 0.97 F1: 0.65	P: 0.43 R: 1.0 F1: 0.6	P: 0.47 R: 0.44 F1: 0.45	P: 0.01 R: 1.0 F1: 0.03	P: 0.86 R: 1.0 F1: 0.93
Transformers (BETO + distilled)	P: 0.22 R: 0.82 F1: 0.35	P: 0.23 R: 0.61 F1: 0.33	P: 0.81 R: 0.64 F1: 0.72	P: 0.34 R: 0.92 F1: 0.5	P: 0.29 R: 0.92 F1: 0.44	P: 0.33 R: 0.27 F1: 0.3	P: 0.1 R: 0.87 F1: 0.19	P: 0.56 R: 0.65 F1: 0.6
Transformers (GPT-3)	P: 0.58 R: 0.96 F1: 0.72	P: 0.93 R: 1.0 F1: 0.96	P: 0.81 R: 0.92 F1: 0.86	P: 0.74 R: 1.0 F1: 0.85	P: 1.0 R: 0.92 F1: 0.96	P: 0.77 R: 0.73 F1: 0.75	P: 1.0 R: 0.63 F1: 0.77	P: 1.0 R: 1.0 F1: 1.0

Cuadro 7. Resultados generales por variable

[7] https://github.com/cientopolis/OI-NLPExtractorDePares

10. Conclusiones y trabajos futuros

El enfoque de *Matching* Basado en Reglas es uno de los más antiguos. Si bien utiliza machine learning porque requiere procesar el texto usando NLP, no requiere el entrenamiento previo de un modelo ni el etiquetado de datos manual. Basta con definir patrones de extracción y aplicarlos sobre el texto. Una desventaja de este enfoque es que está sesgado a los datos. Es decir, la construcción de los patrones se realiza en función de conocer de qué manera suele aparecer escrita una variable en un anuncio inmobiliario, pero es débil frente a la variabilidad que puede tener el lenguaje natural ya que dado otro conjunto de avisos donde se descubra que las variables aparecen escritas de otra manera, sería necesario revisar y ajustar los patrones para que sean competentes para esos textos. También puede haber inconvenientes cuando los datos tienen mucho ruido [12], por lo que puede ser necesario el curado o normalización de los datos antes de aplicar esta técnica.

NER es una técnica costosa dado que requiere un gran volumen de datos anotados manualmente para entrenar un modelo. Esto requiere la disponibilidad de equipos de expertos en el dominio para realizar la tarea de anotación y la disponibilidad de archivos con información relevante para el etiquetado. Al entrenar un modelo la idea es proporcionarle la mayor cantidad de ejemplos posibles, esto implica mostrarle variabilidad de apariciones en los ejemplos para mejorar su capacidad de predicción. Si bien este trabajo obtuvo buenos resultados en esta técnica, se espera que re-entrenando el modelo con un mayor volumen de datos los resultados sean aún mejores.

`Transformers` está en el auge de las arquitecturas para redes neuronales. Dado que los modelos basados en transformadores están entrenados sobre un inmenso conjunto de datos, no es necesario anotar datos manualmente. Además, la disponibilidad de librerías como `pipeline` mejora la experiencia simplificando el uso estos modelos. Para obtener la respuesta óptima, es necesario evaluar distintos formatos de input para realizar las preguntas con las que se extraen las características. GPT-3 tiene una ventaja frente a los otros enfoques por la capacidad de interpretación y generación de la respuesta, es decir, tendría la capacidad de responder afirmativamente que un lote es irregular si detecta que las medidas del terreno son irregulares. En los otros enfoques, esto podría deducirse estableciendo reglas. Un inconveniente a la hora de procesar resultados es que GPT-3 puede generar respuestas en un formato distinto al que aparece en el *ground truth*, lo que dificulta la tarea de comparación.

Este trabajo deja abiertas varias aristas. Una vez extraídos los pares atributo-valor, será necesario alinear esas extracciones acorde a la ontología que formaliza el OI. La comparación de la información extraída de las descripciones contra los datos estructurados de un anuncio puede revelar inconsistencia en la información. Esta inconsistencia puede ser parcial (por ejemplo, el campo `address` del OI contiene Av. Montevideo al 500 y de la descripción se extrae Montevideo) o total (el campo `irregular` del OI es `false`, y la extracción de descripción da `true`) y esto requerirá acciones para la verificación y corrección de datos. Además, podrían definirse reglas para hallar valores para variables que no pudieron

ser extraídos, en función de los valores de otras variables existentes. Sobre las técnicas analizadas, es posible realizar mejoras en todos los enfoques planteados. Dado que se trabajó sobre anuncios sin ocurrencias de oraciones escritas en negativo, es posible mejorar NER y *Rule-based matching* incorporando ejemplos de este tipo. Por ejemplo, dada la oración "El lote no es irregular", ambos enfoques arrojarían una respuesta positiva a la condición de irregularidad. Pero esto es incorrecto, en su lugar debería analizarse la carga positiva o negativa de la ocurrencia para poder determinar el valor de verdad. Una mejora posible sobre los modelos de QA basados en BERT es aplicar *fine-tuning* [23] utilizando un conjunto de datos específico del dominio inmobiliario anotado con las variables requeridas. De esta manera se evitaría entrenar desde cero un modelo, lo cual es una tarea excesivamente costosa. Finalmente, la evaluación podría recrearse midiendo la performance por cada sub-variable. Por ejemplo, en el caso de dirección podría evaluarse el desempeño que tienen los modelos para identificar cada uno de los formatos posibles. Así mismo, en el caso del FOT podría discriminarse el caso donde haya un único valor, o múltiples valores.

Referencias

1. Automatic Extraction of Product Information, https://dida.do/projects/numeric-attribute-extraction-from-product-descriptions, accedido: 2024-2-15
2. Se presentó el "observatorio de valores del suelo e instrumentos de financiamiento del desarrollo urbano". https://www.gba.gob.ar/habitat/noticias/se_present%C3%B3_el_%E2%80%9Cobservatorio_de_valores_del_suelo_e_instrumentos_de_financiamiento, accedido: 2024-2-15
3. ¿Qué es? | Observatorio de valores de suelo, https://observatoriosuelo.gba.gob.ar/institucional/que-es, accedido: 2024-2-15
4. Anantharangachar, R., Ramani, S., S, R.: Ontology Guided Information Extraction from Unstructured Text. International journal of Web & Semantic Technology **4**(1), 19–36 (Jan 2013). https://doi.org/10.5121/ijwest.2013.4102, http://www.airccse.org/journal/ijwest/papers/4113ijwest02.pdf
5. Baur, K., Rosenfelder, M., Lutz, B.: Automated real estate valuation with machine learning models using property descriptions. Expert Systems with Applications **213**, 119147 (Mar 2023). https://doi.org/10.1016/j.eswa.2022.119147, https://www.sciencedirect.com/science/article/pii/S0957417422021650
6. Blandón Andrade, J.C., Zapata Jaramillo, C.M.: Gate-Based Rules for Extracting Attribute Values. Computación y Sistemas **25**(4) (Feb 2021). https://doi.org/10.13053/cys-25-4-3493, https://cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3493, number: 4
7. Brinkmann, A., Shraga, R., Der, R.C., Bizer, C.: Product information extraction using chatgpt (2023), https://api.semanticscholar.org/CorpusID:259262489
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin,

H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
9. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
10. Del Río, J.P., Dioguardi, F., May, M., Torres, D.: Normalización y análisis exploratorio de datos inmobiliarios web. In: XI Jornadas de Sociología de la UNLP 5-7 de diciembre de 2022 Ensenada, Argentina. Sociologías de las emergencias en un mundo incierto. Departamento de Sociología. Facultad de Humanidades y Ciencias de la ... (2022)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019), https://api.semanticscholar.org/CorpusID:52967399
12. Dey Chowdhury, R., Sarkar, A., Banik, M., Bobbili, P.: Product attribute extraction and product listing analysis from e-commerce websites (06 2023). https://doi.org/10.13140/RG.2.2.11045.47842
13. Dioguardi, F., Torres, D., Antonelli, R.L., Río, J.P.d.: Construcción de un grafo de conocimiento para un observatorio inmobiliario. In: XXVIII Congreso Argentino de Ciencias de la Computación (CACIC)(La Rioja, 3 al 6 de octubre de 2022) (2022)
14. Huynh, S., Le, K., Dang, N., Le, B., Huynh, D., Nguyen, B.T., Nguyen, T.T., Ho, N.Y.T.: Named entity recognition for vietnamese real estate advertisements. In: 2021 8th NAFOSTED Conference on Information and Computer Science (NICS). pp. 23–28 (2021). https://doi.org/10.1109/NICS54270.2021.9701519
15. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications **82**(3), 3713–3744 (Jan 2023). https://doi.org/10.1007/s11042-022-13428-4, https://link.springer.com/10.1007/s11042-022-13428-4
16. Linková, M., Gurskỳ, P.: Attributes extraction from product descriptions on e-shops. In: ITAT. pp. 23–26 (2017)
17. Pham, L.V., Pham, S.B.: Information Extraction for Vietnamese Real Estate Advertisements. In: 2012 Fourth International Conference on Knowledge and Systems Engineering. pp. 181–186. IEEE, Danang, Vietnam (Aug 2012). https://doi.org/10.1109/KSE.2012.27, http://ieeexplore.ieee.org/document/6299417/
18. Probst, K., Ghani, R., Krema, M., Fano, A.E., Liu, Y.: Semi-supervised learning of attribute-value pairs from product descriptions. In: International Joint Conference on Artificial Intelligence (2007), https://api.semanticscholar.org/CorpusID:619505
19. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018), https://api.semanticscholar.org/CorpusID:49313245
20. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 784–789. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-2124, https://aclanthology.org/P18-2124
21. Sabeh, K., Kacimi, M., Gamper, J.: CAVE: Correcting Attribute Values in E-commerce Profiles. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 4965–4969. ACM, Atlanta GA USA

(Oct 2022). https://doi.org/10.1145/3511808.3557161, https://dl.acm.org/doi/10.1145/3511808.3557161
22. Sharma, A., Amrita, Chakraborty, S., Kumar, S.: Named entity recognition in natural language processing: A systematic review. In: Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021. pp. 817–828. Springer (2022)
23. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? In: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18. pp. 194–206. Springer (2019)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, , Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
25. Vijayarajan, V., Dinakaran, M., Lohani, M.: Ontology based object-attribute-value information extraction from web pages in search engine result retrieval. In: Kumar Kundu, M., Mohapatra, D.P., Konar, A., Chakraborty, A. (eds.) Advanced Computing, Networking and Informatics- Volume 1. pp. 611–620. Springer International Publishing, Cham (2014)
26. Wang, Q., Yang, L., Kanagal, B., Sanghai, S., Sivakumar, D., Shu, B., Yu, Z., Elsas, J.: Learning to Extract Attribute Value from Product via Question Answering: A Multi-task Approach. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 47–55. ACM, Virtual Event CA USA (Aug 2020). https://doi.org/10.1145/3394486.3403047, https://dl.acm.org/doi/10.1145/3394486.3403047