Generación de Descripciones de Pinturas a Partir de Entidades y Relaciones Usando Llama3.2 3B

Felipe Cruz *fcruzv@unal.edu.co,* David Casallas, *dcasallasm@unal.edu.co,*

***Abstract*—Este artículo presenta un enfoque para la generación automática de descripciones de pinturas a partir de entidades y relaciones identificadas previamente. El trabajo es una continuación de un proyecto anterior titulado *Reconocimiento de Entidades y Relaciones en Descripciones de Pinturas Usando Procesamiento de Lenguaje Natural*. En esta ocasión, se utilizan las entidades y relaciones almacenadas en un formato JSON para generar descripciones coherentes y bien estructuradas. Se exploran dos aproximaciones: una basada en conectores simples y una segunda que emplea un modelo generativo de inteligencia artificial (Llama3.2 3B) para mejorar la redacción de las descripciones. Los resultados demuestran la viabilidad de ambas aproximaciones y sugieren posibles mejoras futuras.**

***Index Terms*—Procesamiento de Lenguaje Natural, Generación de Descripciones, Spacy, Llama3.2, Ollama, Modelo Generativo, Redacción de texto, Formato JSON, Entidades Relacionadas.**

I. Introducción

En el ámbito del procesamiento de lenguaje natural (PLN), la generación automática de texto a partir de datos estructurados es un desafío significativo. Este trabajo se enfoca en la generación de descripciones de pinturas a partir de entidades y relaciones previamente identificadas. El proyecto anterior, titulado *Reconocimiento de Entidades y Relaciones en Descripciones de Pinturas Usando Procesamiento de Lenguaje Natural*, sentó las bases para este trabajo al identificar y almacenar entidades y relaciones en un formato JSON. En este artículo, se presentan dos enfoques para generar descripciones a partir de estos datos: uno basado en conectores simples y otro que utiliza un modelo generativo de inteligencia artificial.

II. Trabajo Relacionado

La generación de descripciones textuales a partir de datos estructurados ha sido un área de interés en el procesamiento de lenguaje natural (PLN). En el contexto de descripciones de pinturas, varios trabajos han abordado problemas similares, desde la identificación de entidades y relaciones hasta la generación de texto coherente y contextualizado.

Uno de los trabajos más relevantes en este ámbito es el de Karpathy y Fei-Fei [1], quienes propusieron un modelo basado en redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN) para generar descripciones de imágenes. Aunque su enfoque se centra en imágenes fotográficas, su metodología inspiró muchos trabajos posteriores en la generación de texto a partir de datos visuales, incluyendo pinturas.

Otro trabajo destacado es el de Vinyals et al. [2], quienes introdujeron el modelo "Show and Tell", que utiliza una arquitectura encoder-decoder para generar descripciones de imágenes. Este enfoque ha sido adaptado en diversos dominios, incluyendo el arte, donde se ha utilizado para describir escenas y objetos en pinturas.

Estos trabajos justificaron la relevancia del uso de nuestro enfoque, que combina la identificación de entidades y relaciones con modelos generativos para producir descripciones de pinturas. A diferencia de los trabajos anteriores, nuestro enfoque se centra en el dominio específico de las pinturas y utiliza un modelo generativo (Llama3.2 3B) para mejorar la redacción de las descripciones.

III. Metodología

En esta sección se describe en detalle el proceso seguido para generar descripciones de pinturas a partir de entidades y relaciones, así como el uso de Llama3.2 3B para mejorar dichas descripciones. El enfoque se divide en cuatro etapas principales: (1) generación de descripciones simples usando conectores, (2) mejora de descripciones mediante un modelo generativo, (3) preprocesamiento de texto (incluyendo lematización), y (4) análisis comparativo de las descripciones generadas.

A. Datos y Estructura

Los datos utilizados en este trabajo consisten en entidades y relaciones identificadas en descripciones de pinturas. Las entidades y relaciones se almacenan en un formato JSON con la siguiente estructura:

```
scene={
  'objects' : [
     <object>
  ]
 'relations' : {
   <relation>
 }
}
```

En dónde Object y Relation están definidos de la siguiente manera:

```
object = {
  'id' : <integer>,
  'type' : <type>,
}
relation = {
  type:   <type>
  'obj1' : <integer>,
  'obj2' : <integer>
}
```

A partir de estos datos, se generan descripciones simples utilizando conectores básicos.

B. Generación de Descripciones Simples

La primera etapa consiste en generar descripciones simples a partir de las entidades y relaciones identificadas. Para cada relación, se construye una oración del tipo

objeto1<type> relation<type> objeto2<type>.

Si un objeto no tiene relaciones, se lista al final de la descripción.
Por ejemplo:

> Objeto1 está a la izquierda de objeto2. Objeto3 está detrás de objeto1. Hay objeto4, objeto5 y objeto6.

Estas descripciones simples sirven como base para la siguiente etapa.

C. Mejora de Descripciones Usando un Modelo Generativo

En la segunda etapa, se utiliza un modelo generativo de inteligencia artificial para mejorar las descripciones simples. Para ello, se emplea la librería `Ollama` en Python, que permite interactuar con modelos de lenguaje avanzados, como el modelo Llama3.2 3B. El proceso es el siguiente:

1) Se toma la descripción simple generada en la etapa anterior.
2) Se utiliza un prompt específico para solicitar al modelo que mejore la redacción de la descripción. El prompt utilizado es:

   ```
   {Description}
   Quiero que me des la descripción
   que te acabo de dar mejor
   redactada pero sin que se
   alargue. Edítala lo mínimo posible
   intentando conectar las diferentes
   oraciones, no agreges nada extra.
   Tu respuesta debe ser únicamente
   la descripción, no digas nada
   extra.
   ```

3) El modelo generativo devuelve una versión mejorada de la descripción, con una redacción más fluida y coherente, respetando las instrucciones del prompt.

D. Generación de un PDF Comparativo

Una vez obtenidas las descripciones simples y las mejoradas, se genera un PDF que incluye:

- La imagen de la pintura.
- La descripción original.
- La descripción simple generada con conectores.
- La descripción mejorada por el modelo generativo.

Este PDF permite una comparación visual y textual de las diferentes versiones de las descripciones.

E. Preprocesamiento de Texto: Lematización

Antes de realizar el análisis de similitud, se aplica un preprocesamiento de texto a las descripciones (tanto las originales como las generadas). Este preprocesamiento incluye la lematización, que consiste en reducir las palabras a su forma base o raíz (por ejemplo, "corriendo" se convierte en "correr"). Para ello, se utiliza el modelo de lenguaje en español `es_core_news_sm` de la librería `spaCy`. El proceso es el siguiente:

- Se carga el modelo `es_core_news_md` de `spaCy`.
- Se aplica la lematización a cada descripción.

Este paso es crucial para garantizar que las palabras se representen de manera consistente antes de aplicar técnicas como TF-IDF.

F. Análisis de Similitud entre Descripciones

Para cuantificar las diferencias entre las descripciones generadas y las originales, se realiza un análisis de similitud utilizando la técnica TF-IDF (Term Frequency-Inverse Document Frequency) y la similitud del coseno. El proceso es el siguiente:

1) **Vectorización con TF-IDF**: Se utiliza la librería `sklearn` para convertir las descripciones lematizadas en vectores numéricos basados en la frecuencia de las palabras. Esto permite representar las descripciones en un espacio vectorial.
2) **Cálculo de la Similitud del Coseno**: Se calcula la similitud del coseno entre los vectores de las descripciones originales, las descripciones simples y las descripciones mejoradas. La similitud del coseno es una medida que varía entre 0 (sin similitud) y 1 (idénticas), lo que permite evaluar cuánto se parecen las descripciones entre sí.

Este análisis proporciona una métrica cuantitativa para comparar la calidad y la fidelidad de las descripciones generadas.

IV. Resultados

En esta sección, se presentan los resultados obtenidos con ambas aproximaciones. Se comparan las descripciones generadas por los conectores simples con las mejoradas por el modelo generativo. Se discuten las ventajas y desventajas de cada enfoque, así como la calidad de las descripciones finales.

Se realizó el proceso con `1412` descripciones.

Descripción Original	Descripción Generada	Descripción Mejorada
La imagen muestra a una mujer sentada en un muelle, mirando el agua.Está rodeada de una cerca de madera, y en el fondo hay árboles y un cielo con nubes.	Imagen a una mujer. Mujer en un muelle. Agua Está rodeada de madera. Madera en el árboles. Cielo con nubes.	Una mujer está en un muelle rodeada de agua y madera que llega hasta los árboles que hay en el cielo con nubes.

Fig. 1. PDF Pintura de Mujer en el Muelle.

Descripción Original	Descripción Generada	Descripción Mejorada
Un río con un bote sobre él, rodeado de árboles y un cielo azul claro.El agua está tranquila y quieta, reflejando los árboles y el cielo de arriba.	Río con un bote. Bote sobre árboles. Hay cielo, agua y cielo.	Un bote flota sobre un río, rodeado de árboles que se reflejan en el agua cristalina y en el cielo azul claro.

Fig. 2. PDF Pintura de Bote en el Agua.

A. Descripciones Beneficiadas por el uso del Modelo

En el caso de la pintura de la mujer en el muelle, la descripción mejorada presenta una redacción más fluida y coherente en comparación con la descripción generada. Sin embargo, se observa que el modelo generativo omite algunos detalles presentes en la descripción original, como la acción de la mujer (“sentada, mirando el agua”) y la especificación de la “cerca de madera”. A pesar de esto, la descripción mejorada es más comprensible que la generada. De manera similar la pintura del bote sobre el río aunque tiene un descripción simple, se beneficia del modelo pues complementa la descripción y le da una estructura más natural.

B. Descripción Perjudicada por el uso del Modelo

En el caso de la pintura del chico con arco, la descripción mejorada introduce información adicional que no está presente en la descripción original, como la descripción del cuerpo del niño (“en un estado de flujo como agua”) y la interpretación subjetiva de las plantas (“se extienden hacia el cielo”). Esto demuestra una limitación del modelo generativo, ya que tiende a agregar detalles inventados que no corresponden a la descripción original. Aunque la redacción es más elaborada, la fidelidad al contenido original se ve comprometida.

Descripción Original	Descripción Generada	Descripción Mejorada
Un niño de pie en el suelo, sosteniendo un arco y una flecha en sus manos.En el fondo, hay un cuerpo de agua, árboles, colinas y un cielo.En la parte inferior.Hay plantas.	Niño de pie. Pie en el suelo. Flecha en sus manos. Manos En el cuerpo. Cuerpo de agua. Cielo En la plantas. Hay arco, árboles y colinas.	Un niño de pie, con una pie en el suelo y una flecha apuntada hacia adelante entre sus manos, que también se encuentran en su cuerpo formando parte integral del mismo. Su cuerpo se encuentra en un estado de flujo como agua, mientras que las plantas se extienden hacia el cielo. En este entorno natural, se encuentran un arco, árboles y colinas.

Fig. 3. PDF Pintura del Chico con Arco.

C. Análisis General

Los resultados muestran que el modelo generativo es efectivo para mejorar la fluidez y coherencia de las descripciones generadas con conectores simples. Sin embargo, en algunos casos, como en el caso de la pintura del chico con arco, el modelo tiende a agregar información no presente en la descripción original, lo que puede llevar a interpretaciones incorrectas. Esto sugiere que, aunque el modelo es útil para mejorar la redacción, es necesario implementar mecanismos de control para garantizar que las descripciones mejoradas mantengan la fidelidad al contenido original.

D. Análisis de la similitud

La figura 4, presenta las medidas de similitud obtenidas. El proceso para realizar la gráfica fue el siguiente:

1) Calcular la similitud entre cada descripción generada y la descripción original.
2) Ordenar los valores de similitud de menor a mayor.
3) Graficar un punto por cada valor.
4) Repetir los pasos anteriores para cada descripción mejorada.

La gráfica, en el eje horizontal, muestra el porcentaje de datos graficados hasta ese punto viéndola de izquierda a derecha. La línea roja punteada muestra la ubicación del 50% de similitud. A partir de la gráfica, se puede deducir lo siguiente:

- La descripción generada es más similar a la descripción original que la descripción mejorada.
- Más del `80%` de las descripciones generadas y mejoradas tienen una similitud mayor al `50%` con la descripción original.

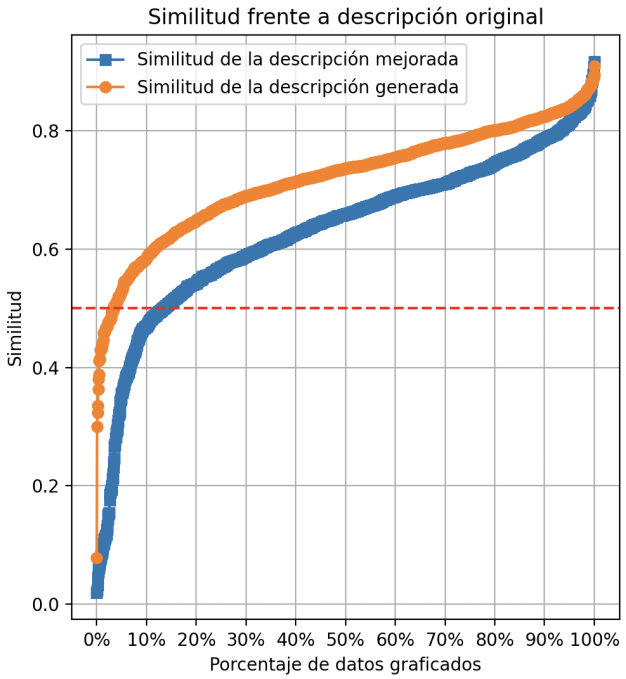

Fig. 4. Gráfica de similitudes entre las descripciones con la descripción original.

Al ordenar los dos conjuntos de datos de manera independiente, se pierde información sobre si para una pintura específica, la descripción mejorada (usando modelo generativo), es mejor o peor que la descripción generada (conectores simples). Para analizar este aspecto, se comparó la similitud entre la descripción generada y la mejorada para cada pintura (uno a uno):

- La descripción mejorada fue más similar que la descripción generada en `376` de las `1412` pinturas, lo que corresponde con el `26.63%`.
- La descripción generada fue más similar que la descripción mejorada en `1036` de las `1412` pinturas, lo que corresponde con el `73.37%`.

Lo anterior es consistente con lo deducido anteriormente desde la gráfica, donde la descripción generada es más similar a la descripción original que la descripción mejorada.

V. Discusión

Los resultados obtenidos en este trabajo muestran que las descripciones generadas con conectores simples son más similares a la descripción original según la medida de similitud del coseno. Esto sugiere que, desde un punto de vista cuantitativo basado en la frecuencia de términos (TF-IDF), las descripciones simples conservan una mayor proximidad léxica con las descripciones originales. Sin embargo, es importante destacar que la similitud del coseno no captura completamente la semántica o el sentido de las oraciones.

Por otro lado, las descripciones mejoradas por el modelo generativo, aunque menos similares léxicamente, presentan una redacción más fluida y coherente, lo que sugiere una mayor cercanía semántica al sentido original de las descripciones. Este hallazgo indica que, aunque las descripciones mejoradas no coinciden exactamente en términos de palabras clave, su estructura y coherencia las hacen más comprensibles y naturales.

Un aspecto importante a considerar es que este trabajo se centró en la similitud léxica (basada en TF-IDF y similitud del coseno) y no en la similitud semántica. Sería valioso en futuros estudios incorporar técnicas de evaluación semántica, como el uso de modelos de embeddings contextuales (por ejemplo, BERT o SBERT) o métricas basadas en la comprensión del significado, como BLEU o ROUGE. Estas técnicas permitirían evaluar no solo la coincidencia de palabras, sino también la similitud en el significado y la coherencia global de las descripciones.

En resumen, aunque las descripciones generadas con conectores simples son más similares a las originales según la métrica de similitud del coseno, las descripciones mejoradas por el modelo generativo ofrecen una redacción más natural y coherente. Futuros trabajos deberían explorar métodos para evaluar la similitud semántica y comparar ambos enfoques desde una perspectiva más integral.

VI. Conclusión

Este trabajo demuestra la viabilidad de generar descripciones de pinturas a partir de entidades y relaciones utilizando tanto conectores simples como modelos generativos de inteligencia artificial. Los resultados sugieren que el uso de modelos generativos puede mejorar significativamente la calidad de las descripciones, aunque se requiere un mayor refinamiento para alcanzar un nivel de detalle y coherencia óptimo.

References

[1] Karpathy, A., & Fei-Fei, L. (2015). *Deep visual-semantic alignments for generating image descriptions*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128–3137).

[2] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). *Show and tell: A neural image caption generator*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156–3164).

[3] Pedregosa, F., et al. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[5] Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Software available at https://spacy.io.

[6] Ollama. (2023). *Ollama: A framework for interacting with generative models*. Software available at https://ollama.com.

[7] Cruz, F., Casallas, D. & Rodríguez, L. (2024). *Reconocimiento de Entidades y Relaciones en Descripciones de Pinturas Usando Procesamiento de Lenguaje Natural*. Artículo primer proyecto para la materia PLN semestre 2024-2 UNAL.