Modelado de relaciones semánticas en lenguaje natural mediante Word2Vec y redes neuronales LSTM

Autora: Dora Melizza Amarilla Cardozo

Resumen

Este proyecto tiene como objetivo principal analizar la capacidad de modelos de procesamiento de lenguaje natural para identificar y clasificar relaciones semánticas complejas. Inicialmente se propuso una exploración semántica mediante Word2Vec, evaluando agrupamientos de términos. Sin embargo, durante el proceso, el enfoque se reformuló hacia un modelo predictivo más robusto. El nuevo título del proyecto pasó a ser 'Modelado de relaciones semánticas en lenguaje natural mediante Word2Vec y redes neuronales LSTM'. Este cambio permitió explorar no solo representaciones semánticas estáticas, sino también relaciones secuenciales y clasificación semántica. Se entrenaron dos modelos LSTM sobre datasets sintéticos —uno binario y otro multiclase— y se evaluaron mediante métricas cuantitativas y pruebas interactivas. Los resultados muestran precisión perfecta en ambos modelos, destacando el potencial de la arquitectura Word2Vec + LSTM incluso en entornos sintéticos.

1. Planteamiento del problema

El entendimiento semántico de relaciones entre palabras es fundamental para el procesamiento del lenguaje natural. Word2Vec ha demostrado ser una herramienta poderosa para capturar relaciones de co-ocurrencia, pero su carácter estático limita su capacidad para representar contextos variables. Este proyecto se pregunta si una red LSTM, apoyada por embeddings preentrenados de Word2Vec, puede superar estas limitaciones y clasificar correctamente frases según su contenido semántico, incluso cuando las categorías son cercanas o ambiguas.

2. Descripción del corpus

Se trabajó con dos conjuntos de datos sintéticos diseñados específicamente para evaluar la capacidad de generalización semántica del modelo.

- Dataset binario: 500.000 frases divididas entre dos categorías (nobleza y transporte).
- Dataset multiclase: 20.000 frases distribuidas en cuatro clases (nobleza, transporte, ciudad y clima).

Cada frase fue generada aleatoriamente con vocabulario temático controlado, incluyendo entre 2 y 4 palabras de ruido común para simular un contexto más natural. Las frases tienen una longitud uniforme de 15 palabras, lo cual facilita el padding y la vectorización para los modelos LSTM.

3. Metodología

Se utilizó el modelo Word2Vec preentrenado 'GoogleNews-vectors-negative300.bin' como base de representación semántica. El preprocesamiento incluyó tokenización, eliminación de stopwords y vectorización. Las frases fueron transformadas a secuencias de vectores, y posteriormente normalizadas mediante padding.

Se entrenaron dos modelos LSTM:

- Modelo binario: arquitectura LSTM con 128 unidades, capa densa con activación sigmoide.
- Modelo multiclase: arquitectura similar, pero con salida softmax y codificación one-hot en las etiquetas.

Ambos modelos fueron entrenados en Google Colab con aceleración GPU. Se utilizó el optimizador Adam con tasa de aprendizaje 0.001 y función de pérdida binary_crossentropy o categorical_crossentropy según el caso.

4. Evaluación de resultados

4.1 Métricas cuantitativas

Modelos	Accuracy en test	Loss:
Modelo binario	1.0000	3.37e-11
Modelo multiclase:	1.0000	2.85e-06

Ambos modelos convergieron rápidamente y sin señales de sobreajuste. Las curvas de entrenamiento mostraron estabilidad desde las primeras épocas.

4.2 Pruebas interactivas

Se desarrolló una función de predicción interactiva que permite al usuario ingresar frases y obtener la predicción del modelo correspondiente. Estas pruebas cualitativas permitieron verificar que los modelos clasificaban correctamente frases con ruido y ambigüedad. El modelo multiclase mostró una capacidad de generalización superior al binario.

Ejemplo:

- Entrada: 'bus garage wheel tire event'
- Salida: Clase 1 (Transporte), confianza: >95%

4.3 Complementariedad de enfoques

El proyecto inicialmente exploraba relaciones astronómicas con Word2Vec, pero evolucionó hacia una propuesta predictiva con LSTM. Ambas etapas fueron complementarias: la primera permitió comprender las limitaciones de modelos estáticos, y la segunda demostró cómo redes neuronales pueden superar esas barreras, aun sin depender de un corpus específico. La prueba con dataset multiclase muestra que se puede simular

escenarios complejos con control semántico, y que los modelos aprenden representaciones robustas.

5. Conclusiones y recomendaciones

Este proyecto comenzó con una propuesta inicial titulada “Exploración semántica de relaciones léxicas mediante Word2Vec en lenguaje natural”, centrada en visualizar y analizar relaciones semánticas usando embeddings estáticos. Sin embargo, en el transcurso del desarrollo, se observó que dicho enfoque, aunque útil para análisis exploratorios, era insuficiente para capturar relaciones secuenciales y contextuales más complejas. Esto motivó una modificación sustancial del enfoque, que fue debidamente informada y aprobada, pasando a un nuevo título: “Modelado de relaciones semánticas en lenguaje natural mediante Word2Vec y redes neuronales LSTM”.

La reformulación permitió incorporar redes neuronales LSTM para fortalecer el componente predictivo y evaluar cómo estas pueden interpretar patrones semánticos en textos a partir de embeddings preentrenados. Dos modelos fueron entrenados y evaluados: uno binario y uno multiclase, demostrando excelente desempeño en tareas de clasificación semántica.

Un aspecto importante de este diseño fue el uso de datasets sintéticos. En particular, se optó por no reutilizar el mismo dataset del enfoque exploratorio inicial. El uso de datos sintéticos en esta fase no es una limitación, sino una estrategia válida y eficaz para comprobar si un modelo es capaz de aprender patrones semánticos bajo condiciones controladas. Esto sienta las bases para futuras extensiones con corpus reales, ya que se parte de modelos previamente validados en entornos sintéticos.

Los resultados obtenidos fueron notables: ambos modelos lograron una precisión del 100% en los datos de prueba, sin evidencia de sobreajuste. Las pruebas interactivas al final del proyecto sirvieron para validar la capacidad de generalización, confirmando que los modelos no solo memorizan, sino que pueden interpretar correctamente frases nuevas con combinaciones semánticas similares a las clases entrenadas.

Se recomienda en futuros trabajos:

1. Evaluar estos modelos en datasets reales y no controlados, como corpus de noticias, Wikipedia o foros especializados.
2. Introducir embeddings contextuales (BERT, RoBERTa) para comparar su rendimiento frente a Word2Vec.
3. Realizar análisis de error más exhaustivos y visualización de activaciones neuronales para comprender mejor el proceso de decisión del modelo.
4. Ampliar el enfoque a tareas más complejas como clasificación jerárquica o etiquetado secuencial.

En resumen, el proyecto no solo cumplió sus objetivos reformulados, sino que demuestra que la combinación de Word2Vec y redes LSTM puede capturar estructuras semánticas relevantes de manera efectiva. La decisión metodológica de emplear corpus sintéticos balanceados permitió establecer una base sólida para investigaciones futuras en PLN y aprendizaje profundo aplicado a relaciones semánticas.