UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE CIENCIAS

Análisis del sobreajuste en modelos de lenguaje N-grama y técnicas de regularización

Curso: Procesamiento del lenguaje Natural

Alumnos:

- Paredes Lopez Maxwel
- Olivares Ventura Ricardo Leonardo

índice

Introducción

Los modelos de lenguaje N-grama han sido una herramienta fundamental para el procesamiento del lenguaje natural debido a su capacidad para capturar las relaciones locales entre palabras. Sin embargo, como muchos modelos, estos también presentan desafíos importantes, como el sobreajuste y escasez de datos, lo cual limita su capacidad para generalizar a nuevos contextos. En este proyecto, se explora el problema del sobreajuste en modelos de N-grama y se aplican técnicas de regularización, específicamente el suavizado de interpolación y backoff, con el objetivo de mejorar el rendimiento del modelo

El análisis lo realizaremos utilizando dos corpus distintos: El Europarl, un corpus de discursos del Parlamento Europeo que tiene un vocabulario formal y técnico, y el corpus Brown, un corpus diverso que contiene textos de múltiples géneros. La comparación entre estos corpus nos permitirá evaluar el impacto de las técnicas de regularización en diferentes tipos de datos textuales y la capacidad del modelo para generalizar a dominios variados

Marco Teórico

1. Modelos N-grama:

Los modelos N-grama son una clase de modelos probabilísticos utilizados ampliamente en NLP para modelar secuencias de palabra, el principal objetivo es calcular la probabilidad de que una palabra aparezca dada la secuencia de palabras que la precede.

Estos modelos se construyen considerando secuencias de N palabras consecutivas llamadas “N-gramas”, y se asume que la probabilidad de cada palabra en una secuencia depende únicamente de las N-1 palabras anteriores

Los N-gramas se clasifican según el valor de N, que representa la longitud de las secuencias que se consideran:

- Unigramas (N = 1): Cada palabra es considerada de manera independiente, es decir, ignoran la información contextual, lo cual los hace inadecuados para tareas que requieren una comprensión básica del contexto, como la generación de texto fluido. La probabilidad de cada palabra se calcula en base únicamente a su frecuencia en el corpus de entrenamiento
- Bigrama (N = 2): En los modelos de bigramas, la probabilidad depende de la palabra que la precede. Esto brinda una mayor capacidad para capturar dependencias locales en una secuencia. Por ejemplo, las bigramas podrían ayudar a capturar combinaciones comunes de palabras, como “buen día”, hasta luego”, lo cual hace que el modelo sea más preciso al predecir la palabra siguiente
- Trigrama (N = 3): Estos modelos consideran las dos palabras anteriores para predecir la siguiente. Esto permite generar mejores combinaciones, pero también se requiere de una mayor cantidad de datos para poder calcular adecuadamente las probabilidades de las secuencias.

2. Sobreajuste en Modelos de Lenguaje:

Una de las grandes limitaciones de los modelos N-grama es la incapacidad de generalizar adecuadamente a combinaciones de palabras que no han sido vistas en los datos de entrenamiento, lo cual lleva a la necesidad de tener una gran cantidad de datos para estimar con precisión las probabilidades, justamente a esto lo conocemos como Sobreajuste, el cual ocurre cuando el modelo se adapta demasiado a los datos de entrenamiento, capturando patrones significativos como el ruido.

2.1. Características del Sobreajuste en Modelos N-grama:

- **Escasez de datos**: A medida que el valor de N aumenta, la cantidad de posibles N-gramas crece exponencialmente, lo que significa que muchas de las posibles combinaciones de palabras probablemente no aparecerán en el corpus de entrenamiento. Esto causa que los modelos de N-grama de orden alto asignen probabilidades de cero a secuencias no vistas, o que sobrevaloren las pocas secuencias que han visto, lo cual los hace propensos al sobreajuste
- **Memorización de datos específicos del Corpus:** En lugar de aprender patrones generales del lenguaje, los modelos N-grama pueden memorizar secuencias específicas de palabras encontradas en el corpus de entrenamiento. Esto significa que el modelo se ajusta demasiado bien a los ejemplos del corpus, y no puede predecir correctamente en situaciones donde los datos no coinciden con lo que ha “memorizado”

2.2. Impacto del Sobreajuste en Modelos de Lenguaje:

Por todo lo descrito líneas arriba, el sobreajuste tiene un impacto significativo en la calidad de los modelos N-grama:

- **Baja Capacidad de Generalización:** Los modelos sobre ajustados tienen dificultades para generalizar más allá de los datos de

entrenamiento. En el caso de los modelos N-grama, esto se traduce en la incapacidad de generar oraciones coherentes o de predecir correctamente la próxima palabra si la secuencia observada difiere de los ejemplos vistos durante el entrenamiento

- **Alta Varianza:** Un modelo sobre ajustado tiende a tener alta varianza, lo que significa que su rendimiento varía considerablemente entre diferentes conjuntos de datos. Puede tener un rendimiento excepcional en el conjunto de datos de entrenamiento, pero fallar significamente en otros conjuntos, especialmente si esos otros conjuntos contiene patrones lingüísticos que no se observaron durante el entrenamiento

2.3. Ejemplo de Sobreajuste en Modelos de N-grama:

Consideremos un modelo de trigramas entrenado en un corpus específico de noticias deportivas. Si el corpus tiene repetidas menciones de “la liga española de fútbol”, el modelo de trigramas podría aprender a predecir con una probabilidad alta que, dado “la liga”, la siguiente palabra debería ser “española”. Sin embargo, en un contexto diferente, como “la liga de baloncesto”, el modelo podría asignar una probabilidad incorrectamente baja a “baloncesto”, debido a que no ha visto esta combinación con la misma frecuencia

3. Técnicas de regularización en Modelos N-Grama

La regularización se refiere a un conjunto de técnicas utilizadas en el aprendizaje automático para prevenir el sobreajuste y mejorar la capacidad del modelo para generalizar. En el contexto de los modelos de lenguaje, la regularización ayuda a evitar que el modelo dependa demasiado de secuencias de palabras específicas que ha visto durante el entrenamiento, permitiéndole ser más flexible y asignar probabilidades razonables a secuencias de palabras no vistas.

En los modelos N-grama, la regularización se logra principalmente mediante técnicas de suavizado. Estas técnicas asignan una pequeña

probabilidad no nula a los N-gramas no observados durante el entrenamiento, lo cual permite al modelo manejar mejor los datos nuevos y generalizar más efectivamente.

3.1 Suavización de interpolación

La interpolación es una técnica de suavizado más sofisticado que combina probabilidades de diferentes modelos N-grama. En lugar de usar únicamente un modelo trigramas, por ejemplo, la interpolación combina los modelos unigramas, bigramas y trigramas, asignando un peso a cada uno. De esta forma, la probabilidad de una palabra se calcula como una combinación de diferentes órdenes de N-gramas:

$$P(wi|wi-2,wi-1) = \lambda 1 P(wi) + \lambda 2 P(wi|wi-1) + \lambda 3 P(wi|wi-2,wi-1)$$

Donde λ1, λ2, λ3 son pesos que suman 1. Esta técnica ayuda a que el modelo sea más robusto frente a secuencias no vistas, ya que siempre puede recurrir a probabilidades de menor orden, mejorando así la capacidad de generalización y mitigando el sobreajuste.

3.2 Suavizado Backoff

El suavizado backoff es otra técnica utilizada para lidiar con secuencias no observadas. En lugar de calcular la probabilidad de una secuencia únicamente a partir de los N-gramas de orden más alto, el backoff “retrocede” a un modelo de menor orden si no se encuentra una secuencia específica en el corpus de entrenamiento. Esto significa que si un trigramas no se encuentra en el corpus, el modelo usará el correspondiente bigrama, y si este tampoco está disponible, usará el unigrama.

Por ejemplo, la probabilidad de una palabra en un modelo backoff se define como

$$P(w_i|w_{i-2},w_{i-1}) = \begin{cases} \text{probabilidad estimada} & \text{si } (w_{i-2},w_{i-1},w_i) \text{ existe} \\ \alpha P(w_i|w_{i-1}) & \text{si } (w_{i-2},w_{i-1},w_i) \text{ no existe} \end{cases}$$

Donde α es un factor de normalización que asegura que las probabilidades sumen 1. Esta técnica ayuda a reducir el

sobreajuste ya que distribuye las probabilidades de manera más uniforme, evitando asignar valores excesivamente altos a secuencias que fueron vistas pocas veces.

4. Descripción de Corpus asignado

4.1 Corpus Europarl

El corpus Europarl es una colección de transcripciones de los debates del Parlamento Europeo. Se creó como parte de un proyecto para entrenar y evaluar sistemas de traducción automática y modelos de lenguaje multilingües. Dado que los debates se llevan a cabo en varios idiomas oficiales de la Unión Europea (UE), el corpus incluye textos paralelos, lo que significa que el mismo contenido está disponible en varios idiomas.

Características:

- Idiomas: El corpus contiene textos en más de 20 idiomas europeos, incluyendo inglés, francés, alemán, español, italiano, entre otros.
- Volumen de datos: Es un corpus extenso, con millones de palabras en cada idioma.
- Estructura: Se compone de transcripciones reales de discursos parlamentarios, lo que lo hace valioso para tareas de procesamiento de lenguaje natural como la traducción automática, el alineamiento de oraciones entre idiomas y la creación de diccionarios bilingües.
- Disponibilidad: Fue creado y es mantenido por la comunidad académica, y está disponible de manera gratuita para fines de investigación y desarrollo de modelos de NLP.

4.2 Corpus Brown

El corpus Brown es un conjunto de textos en inglés compilado en 1961 por investigadores de la Universidad de Brown en los Estados Unidos. Es uno de los primeros corpus ampliamente utilizados en la investigación del procesamiento del lenguaje natural y la lingüística computacional. A

diferencia del Europarl, el corpus Brown incluye una variedad de géneros y estilos de escritura en inglés estadounidense.

Características:

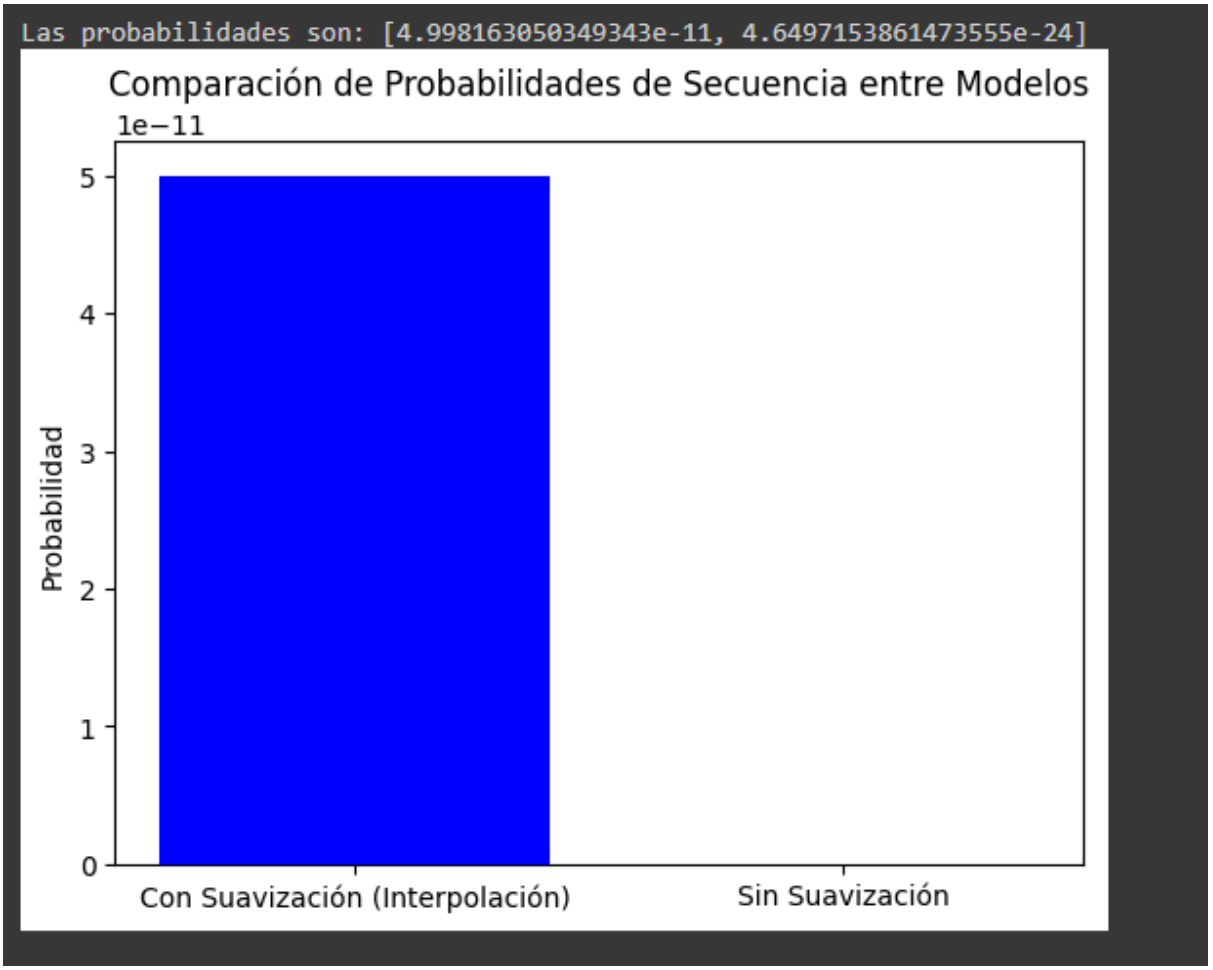
- Tamaño: El corpus contiene aproximadamente 1 millón de palabras distribuidas en 500 textos.
- Diversidad: Se recopilaron en textos de 15 géneros diferentes, incluyendo ficción, no ficción, ensayos, artículos periodísticos, textos académicos y más. Esto lo convierte en un corpus muy diverso en términos de temas y estilos de lenguaje.
- Estructura: Cada texto está etiquetado con su género y categoría, lo que permite realizar estudios de comparación de estilos lingüísticos entre diferentes tipos de escritos.
- Influencia histórica: El Corpus Brown fue el primer corpus de texto creado con el objetivo de analizar el uso del lenguaje de manera sistemática y ha sido utilizado como base para otras corpus.

Código REPOSITORIO

https://github.com/rlov/pc1-NLP

Conclusiones:

Los modelos con regularización tienen mayor probabilidad que los sin regularización, mostrando sus eficiencias.



```
Las probabilidades son: [4.9981630503493415e-08, 4.6497153861473555e-24]
```