bigrama – bigram

Authored by

memjavad

November 7, 2025

Bigrama

Primary Disciplinary Field(s): Lingüística Computacional, Procesamiento del Lenguaje Natural (PLN), Estadística, Criptografía

1. Definición Central

El **bigrama** (del latín *bi-*, dos, y del griego *grámma*, letra o escritura) es un concepto fundamental dentro de la disciplina de la lingüística computacional y el procesamiento del lenguaje natural (PLN). Se define rigurosamente como una secuencia contigua de dos elementos dentro de una muestra de texto o voz. Estos elementos pueden ser unidades discretas como caracteres, sílabas, fonemas o, más comúnmente, palabras. El bigrama constituye el caso más sencillo de un modelo estadístico de lenguaje conocido como **N-grama**, donde N es igual a dos. Su propósito esencial es capturar la dependencia estadística local entre dos unidades adyacentes, permitiendo a los sistemas computacionales predecir la probabilidad de que un elemento siga a otro basándose en la frecuencia observada en un corpus de entrenamiento masivo. Esta capacidad de modelar la coocurrencia inmediata es crucial para tareas que requieren una comprensión probabilística de la estructura superficial del lenguaje.

La utilidad del bigrama reside en su eficiencia para construir modelos de lenguaje que equilibren la complejidad y la capacidad predictiva. A diferencia de los modelos basados en reglas gramaticales explícitas, los modelos de bigramas operan exclusivamente con la estadística observada, asumiendo que la probabilidad de que una palabra específica (w_i) aparezca en una secuencia depende únicamente de la palabra inmediatamente anterior (w_{i-1}). Esta simplificación, conocida como la **suposición de Markov de primer orden**, reduce significativamente la complejidad computacional y la cantidad de datos necesarios en comparación con modelos que intentan capturar dependencias de largo alcance. Aunque esta suposición simplifica inherentemente la riqueza sintáctica y semántica del lenguaje humano, ofrece una base sorprendentemente robusta para muchas aplicaciones prácticas, especialmente cuando se trabaja con grandes volúmenes de texto.

Para generar un modelo de bigramas, se requiere un corpus de texto extenso y representativo del dominio lingüístico de interés. El proceso implica la tokenización del texto (dividirlo en unidades, generalmente palabras) y luego contar la frecuencia con la que aparece cada par ordenado de palabras consecutivas. Por ejemplo, en la frase "El perro corre rápido", se identificarían los bigramas ("El", "perro"), ("perro", "corre"), y ("corre", "rápido"). La acumulación de estas frecuencias en una matriz de coocurrencia permite calcular la probabilidad condicional de cualquier palabra dada su predecesora. Este enfoque estadístico ha sido históricamente la columna vertebral de los primeros sistemas de PLN y sigue siendo relevante como una métrica de referencia y un componente en arquitecturas más avanzadas.

2. Fundamentos Matemáticos: Modelos de Markov y Cadenas

Matemáticamente, el bigrama se fundamenta en la teoría de los Modelos de Markov, específicamente en la cadena de Markov de primer orden. En el contexto del modelado del lenguaje, una cadena de Markov es un proceso estocástico que describe una secuencia de posibles eventos, donde la probabilidad de cada evento depende únicamente del estado alcanzado en el evento anterior. Aplicado a una secuencia de palabras $W = (w_1, w_2, dots, w_m)$, el objetivo es calcular la probabilidad de toda la secuencia, $P(W)$. Utilizando la regla de la cadena para probabilidades conjuntas, $P(W)$ se descompone en el producto de probabilidades condicionales: $P(w_1, w_2, dots, w_m) = P(w_1) cdot P(w_2|w_1) cdot P(w_3|w_1, w_2) cdot dots cdot P(w_m|w_1, dots, w_{m-1})$.

Aquí es donde entra en juego la suposición de Markov de primer orden. En lugar de condicionar la probabilidad de w_i a toda la historia previa de la secuencia, el modelo de bigrama simplifica esta dependencia, asumiendo que $P(w_i | w_1, dots, w_{i-1}) approx P(w_i | w_{i-1})$. Por lo tanto, la probabilidad de la secuencia completa se aproxima mediante el producto de las probabilidades de bigramas individuales: $P(W) approx P(w_1) cdot prod_{i=2}^{m} P(w_i | w_{i-1})$. Esta simplificación drástica permite la construcción de modelos estadísticos robustos que son entrenables eficientemente a partir de datos empíricos. La probabilidad condicional de un bigrama, $P(w_i | w_{i-1})$, se estima utilizando la técnica de máxima verosimilitud (MLE) a partir de las frecuencias observadas en el corpus.

La fórmula de estimación por máxima verosimilitud para un bigrama es: $P_{MLE}(w_i | w_{i-1}) = frac{text{Count}(w_{i-1} w_i)}{text{Count}(w_{i-1})}$. El numerador cuenta el número de veces que el par de palabras $w_{i-1} w_i$ aparece consecutivamente en el corpus, mientras que el denominador cuenta el número total de ocurrencias de la palabra w_{i-1} (la palabra de contexto). La implementación práctica de esta fórmula implica la construcción de una matriz de transición de bigramas, donde las filas representan las palabras de contexto y las columnas representan las palabras siguientes. Esta matriz codifica todas las probabilidades de transición y es la esencia del modelo estadístico de bigramas.

3. Desarrollo Histórico y Contexto Lingüístico

Aunque el término **N-grama** y su aplicación formal al lenguaje se popularizaron con el auge de la lingüística computacional en la segunda mitad del siglo XX, las ideas subyacentes de modelar secuencias de símbolos mediante estadísticas de pares tienen raíces más profundas. Los trabajos pioneros de Claude Shannon en la Teoría de la Información (décadas de 1940 y 1950) fueron fundamentales. Shannon demostró que las secuencias de texto, incluso aquellas generadas por procesos estocásticos simples (como los modelos de Markov), podían exhibir propiedades sorprendentemente similares al lenguaje humano, especialmente en términos de entropía y

redundancia. Sus experimentos utilizando modelos de orden cero, primer orden (bigramas) y segundo orden (trigramas) para predecir la siguiente letra en un texto inglés establecieron las bases empíricas para el uso de estadísticas de coocurrencia.

Durante las décadas siguientes, la lingüística estuvo dominada por el paradigma racionalista de Noam Chomsky, que priorizaba la gramática formal y las reglas sintácticas sobre la estadística. Sin embargo, a finales de los años 80 y principios de los 90, hubo un resurgimiento de los enfoques empíricos y estadísticos, impulsado por el aumento de la capacidad de procesamiento y la disponibilidad de grandes corpus de texto. Los modelos de N-gramas, incluyendo los bigramas, se convirtieron en la técnica estándar para el modelado de lenguaje en aplicaciones prácticas como el reconocimiento de voz. La simplicidad, robustez y facilidad de entrenamiento de los bigramas les otorgaron una ventaja significativa sobre los complejos sistemas basados en reglas.

El desarrollo del PLN moderno se ha caracterizado por una continua búsqueda de modelos que capturen dependencias más largas (trigramas, 4-gramas) o estructuras profundas (redes neuronales). No obstante, los bigramas han mantenido su relevancia. Sirven como una línea base de rendimiento (*baseline*) contra la cual se miden modelos más sofisticados. Además, los principios de frecuencia de bigramas son esenciales no solo para las palabras, sino también para el análisis de secuencias de caracteres, lo que tiene aplicaciones directas en la detección de errores tipográficos, la segmentación de palabras en idiomas sin espacios explícitos (como el chino) y el análisis de estilos de escritura.

4. Cálculo de Probabilidades y Suavizado

La implementación práctica de los modelos de bigramas enfrenta un desafío crítico conocido como el problema de la **escasez de datos** o *data sparsity*. En cualquier corpus de tamaño finito, inevitablemente habrá secuencias de bigramas que no se observaron durante el entrenamiento, aunque sean gramaticalmente válidas. Cuando el modelo intenta calcular la probabilidad de un bigrama no visto, la fórmula MLE arroja un conteo de cero, resultando en una probabilidad condicional de $P(w_i | w_{i-1}) = 0$. Asignar una probabilidad de cero a una secuencia nunca antes vista es catastrófico, ya que, por la regla del producto, cualquier secuencia de texto que contenga ese bigrama no visto tendrá una probabilidad total de cero, independientemente de la frecuencia de los otros bigramas en la secuencia.

Para mitigar este problema, se emplean técnicas de **suavizado** (*smoothing*). El objetivo del suavizado es ajustar las probabilidades asignadas, robando una pequeña porción de la probabilidad total a los bigramas observados y redistribuyéndola entre los bigramas no observados. Esto asegura que todos los bigramas posibles, incluso aquellos con conteo cero, reciban una probabilidad positiva, aunque muy pequeña. Las técnicas de suavizado varían en complejidad y efectividad.

Una de las técnicas más simples es el **suavizado de Laplace** (o adición de uno), donde se añade una unidad (o un valor $delta$) a todos los conteos, tanto en el numerador como en el denominador, antes de calcular la probabilidad. Otras técnicas más sofisticadas y eficientes incluyen el suavizado de Good-Turing, que reasigna la masa de probabilidad basándose en la frecuencia de las frecuencias, y el suavizado de **Kneser-Ney**, que es generalmente considerado el estándar de oro para los modelos de N-gramas. Kneser-Ney es particularmente efectivo porque utiliza información de órdenes inferiores (unigramas) para estimar la probabilidad de bigramas no vistos, basándose en cuán probable es que la palabra siguiente aparezca en cualquier contexto, no solo después de la palabra de contexto específica.

5. Aplicaciones en Procesamiento del Lenguaje Natural

El modelo de bigramas, debido a su simplicidad y eficacia, ha encontrado aplicaciones extensas y variadas en el campo del PLN, sirviendo a menudo como el núcleo estadístico de sistemas complejos. Una de las aplicaciones más importantes es el **reconocimiento de voz**. En estos sistemas, el bigrama actúa como el modelo de lenguaje que ayuda a desambiguar entre varias posibles transcripciones fonéticas. Si el sistema acústico sugiere que la palabra siguiente podría ser "sol" o "sal", el modelo de bigramas consulta su matriz para determinar cuál de los pares ("palabra anterior", "sol") o ("palabra anterior", "sal") tiene una probabilidad de ocurrencia significativamente mayor en el lenguaje natural.

Otra aplicación crucial es la **corrección de errores ortográficos y gramaticales**. Los correctores de texto utilizan bigramas para identificar secuencias de palabras que son estadísticamente improbables. Por ejemplo, si un usuario escribe "de casa", el sistema puede evaluar la probabilidad de este bigrama y compararla con bigramas similares como "en casa" o "la casa". Al trabajar a nivel de caracteres, los bigramas también son utilizados en la detección de errores tipográficos comunes (como la transposición de letras). Además, en la **traducción automática estadística** (antes de la era de la traducción neuronal), los modelos de bigramas se utilizaban para garantizar la fluidez gramatical de la frase traducida en el idioma de destino.

Finalmente, los bigramas son fundamentales en el desarrollo de herramientas de **segmentación de texto** y **etiquetado de partes del discurso** (Part-of-Speech Tagging, POS). En el etiquetado POS, los modelos estadísticos, como los modelos ocultos de Markov (HMM), utilizan las probabilidades de transición de bigramas de etiquetas (por ejemplo, la probabilidad de que un sustantivo sea seguido por un verbo) para determinar la secuencia de etiquetas más probable para una oración dada. Esta capacidad para modelar transiciones locales hace que los bigramas sean herramientas versátiles y esenciales para el análisis de la estructura superficial del lenguaje.

6. Uso en Criptografía y Análisis de Frecuencia

Históricamente, el análisis de bigramas no se limitó a la lingüística computacional, sino que también desempeñó un papel vital en la **criptografía** y, más específicamente, en el criptoanálisis. Los sistemas de cifrado clásicos, como los cifrados de sustitución simple o los cifrados polialfabéticos (como el Vigenère), se basan en la sustitución de letras individuales. El método más básico para romper estos cifrados es el análisis de frecuencia de letras individuales. Sin embargo, los bigramas proporcionan una capa de información estadística mucho más rica y potente.

En cualquier idioma natural, no todas las combinaciones de letras son igualmente probables. Por ejemplo, en español, la secuencia "qu" es extremadamente común, mientras que "qz" es virtualmente inexistente. Los criptoanalistas, especialmente durante la Primera y Segunda Guerra Mundial, recopilaban y utilizaban tablas de frecuencias de bigramas (y trigramas) para el idioma objetivo. Cuando se aplicaba un cifrado de sustitución, las secuencias de dos letras en el texto cifrado que correspondían a los bigramas más frecuentes del idioma natural proporcionaban pistas cruciales para descifrar la clave de sustitución, incluso cuando el análisis de letras individuales no era concluyente.

El conocimiento de los patrones de bigramas, tales como los bigramas iniciales (secuencias de inicio de palabra) o los bigramas finales (secuencias de fin de palabra), permitía a los analistas reducir drásticamente el espacio de búsqueda para la clave. Esta dependencia de la distribución estadística de los pares de caracteres subraya la debilidad inherente de los cifrados que no ocultan completamente la estructura probabilística del idioma subyacente. Los bigramas, por lo tanto, son un testimonio de que la estadística local del lenguaje es una huella digital poderosa, útil tanto para la generación de texto como para su decodificación.

7. Limitaciones Teóricas y Comparación con N-gramas Superiores

La principal limitación teórica del modelo de bigramas se deriva directamente de su suposición fundacional: la **suposición de Markov de primer orden**. Esta suposición ignora las dependencias de largo alcance que son omnipresentes en el lenguaje humano. La elección de una palabra a menudo está influenciada no solo por la palabra inmediatamente anterior, sino también por el sujeto de la oración, el verbo principal o el contexto temático establecido varias palabras atrás. Por ejemplo, en la frase "El científico que inventó la máquina... y que fue galardonado... su estudio", la concordancia del pronombre o del verbo puede depender del sujeto original ("científico"), no de la palabra inmediatamente precedente ("galardonado").

Para intentar capturar un contexto más amplio, se utilizan **N-gramas superiores**, como los trigramas (N=3) o los 4-gramas (N=4). Un trigrama, por ejemplo, modela $P(w_i | w_{i-1}, w_{i-2})$, condicionando la palabra actual a las dos palabras precedentes. Si bien los N-gramas superiores son intrínsecamente más precisos porque utilizan más contexto, introducen un

problema de complejidad exponencial: a medida que N aumenta, el número de posibles secuencias N-grama crece exponencialmente, exacerbando el problema de la escasez de datos. Un corpus que es adecuado para estimar bigramas puede ser insuficiente para estimar trigramas o 4-gramas de manera confiable.

El uso de bigramas representa un compromiso práctico. Proporcionan suficiente contexto para capturar gran parte de la sintaxis local y las idiosincrasias del idioma (como "a pesar de" o "sin embargo") sin inflar excesivamente el tamaño del modelo ni sucumbir completamente a la escasez de datos. En la práctica moderna, los sistemas suelen emplear un enfoque de **modelado de respaldo** (*backoff model*) o **interpolación**, donde se combinan modelos de N-gramas de diferentes órdenes (por ejemplo, 4-gramas, trigramas, bigramas y unigramas). Si un 4-grama no se encuentra en el corpus, el sistema "retrocede" al trigrama, y si este tampoco está, recurre al bigrama, asegurando siempre una estimación de probabilidad positiva y aprovechando el contexto más amplio disponible.

8. Impacto en la Modelización del Lenguaje

El concepto de bigrama ha dejado un legado duradero como el andamiaje sobre el cual se construyó la modelización estadística del lenguaje. Antes del surgimiento de las arquitecturas de aprendizaje profundo (como las Redes Neuronales Recurrentes o los Transformers), los modelos de N-gramas, y por extensión, los bigramas, eran la tecnología dominante. Su impacto no solo fue técnico, sino también metodológico, al cimentar la idea de que el lenguaje puede ser tratado como un fenómeno estadístico medible, alejándose del dogma puramente formalista.

Incluso con la hegemonía actual de los modelos neuronales, los bigramas mantienen su importancia. Sirven como una referencia de rendimiento indispensable para evaluar la complejidad y la efectividad de los modelos neuronales. Si un modelo complejo de aprendizaje profundo no supera significativamente el rendimiento de un simple modelo de bigramas suavizado, esto sugiere que el modelo avanzado está mal entrenado o es excesivamente complejo para la tarea. Además, la métrica de **perplejidad**, utilizada para evaluar la calidad de un modelo de lenguaje, se basa directamente en la probabilidad de la secuencia de palabras, un cálculo que en sus formas más básicas se deriva de los principios de los bigramas.

En resumen, el bigrama es más que una simple secuencia de dos elementos; es la manifestación más simple y eficiente del principio de la dependencia local en el lenguaje. Ha demostrado ser una herramienta estadísticamente potente, fundamental para la evolución del PLN, y que continúa ofreciendo soluciones prácticas y eficientes para un vasto número de problemas computacionales relacionados con la lengua.

9. Lecturas Adicionales

N-grama (Wikipedia en español)
Cadena de Márkov (Wikipedia en español)
Smoothing (statistics) (Wikipedia en inglés)
Language model (Wikipedia en inglés)