Modelos de Linguagem com N-Gramas

1. Introdução aos Modelos de Linguagem com N-Gramas

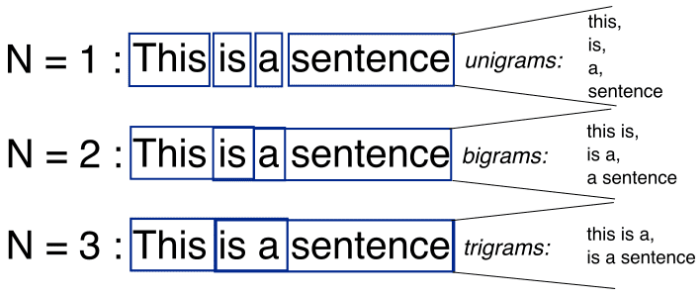
Modelos de linguagem são ferramentas que estimam a probabilidade de uma sequência de palavras. Por exemplo, eles podem prever qual palavra vem a seguir em uma frase, como em sistemas de correção automática, assistentes virtuais ou reconhecimento de fala.

Exemplo: Dada a frase “A água do lago Walden é tão lindamente...”, palavras como azul, clara ou verde são mais prováveis do que geladeira.

2. O Que São N-Gramas

N-gramas são sequências de *n* palavras. Um modelo de n-grama estima a próxima palavra com base nas últimas *n-1* palavras da frase.

- **Bigramas (n = 2)**: “A água”, “do lago”.
- **Trigramas (n = 3)**: “A água do”, “água do lago”.



3. Como os N-Gramas Funcionam

Problema com frases longas: Frases exatas raramente se repetem. Por isso, contar todas as combinações possíveis é inviável. A solução é usar apenas as últimas palavras. Essa é a hipótese de Markov: prever a próxima palavra com base em poucas palavras anteriores.

Exemplo com Bigrama (N = 2): P(verde | lindamente) ≈ probabilidade de “verde” vir após “lindamente”.

Fórmula geral para n-gramas:
P(palavra atual | n-1 palavras anteriores)

4. Estimando as Probabilidades com Frequências (MLE)

Usa-se um corpus (conjunto de textos) para contar quantas vezes as palavras ocorrem juntas.

- Bigramas: P(am | I) = contagem("I am") / contagem("I")
- Trigramas: P(gosto | eu não) = contagem("eu não gosto") / contagem("eu não")

Exemplo com Corpus Simples

<s> Eu sou Sam </s>

<s> Sam eu sou </s>

<s> Eu não gosto de ovos verdes e presunto </s>

Algumas probabilidades:

- P(Eu |) = 2/3
- P(sou | Eu) = 2/3
- P(Sam | sou) = 1/2

5. Calculando a Probabilidade de uma Frase

Para a frase: “Eu quero comida brasileira”

Calculamos:
P(Eu quero comida brasileira) =
P(Eu |) × P(quero | Eu) × P(comida | quero) × P(brasileira | comida) × P(| brasileira)

6. O Que os N-Gramas Capturam

- Sintaxe: Verbos seguidos de substantivos.
- Estilo de linguagem: Frases como “Eu quero” são comuns em aplicativos de voz.
- Cultura: “Comida chinesa” pode ser mais comum do que “comida brasileira”, dependendo do corpus.

7. Dificuldades Técnicas e Soluções

- Multiplicar várias probabilidades pequenas pode causar problemas numéricos (underflow). A solução seria usar logaritmos.
- Para frases longas é comum usar 4-gramas ou 5-gramas, se houver dados suficientes.
- Para otimizar recursos:
    - Reduzir tamanho dos dados,
    - Usar estruturas eficientes,
    - Remover n-gramas com pouca utilidade.

8. Avaliação dos Modelos de Linguagem

- **Avaliação Extrínseca**: mede o desempenho do modelo em tarefas reais, como tradução ou reconhecimento de fala.
- **Avaliação Intrínseca**: avalia diretamente o modelo com métricas como perplexidade, que mede quão bem o modelo prevê a próxima palavra. Quanto menor a perplexidade, melhor. Exemplo: Se um modelo prevê com alta confiança as palavras corretas, ele tem baixa perplexidade.

9. Sampling: Gerar Frases com o Modelo

O modelo pode ser usado para gerar frases automaticamente, sorteando palavras com base em suas probabilidades. Quanto maior o n (ex: trigramas, 4-gramas), mais natural a frase.

Exemplo:

- Unigrama: frases desconexas.
- Bigramas: alguma coerência.
- Trigramas: frases parecidas com o estilo do autor original (como Shakespeare).
- 4-gramas: podem copiar frases do corpus (overfitting).

10. Entropia e Linguagem

Entropia mede a incerteza de uma variável aleatória. Em linguagem indica quantos bits seriam necessários (em média) para representar uma palavra.

Exemplo: Se todos dizem sempre a mesma frase, entropia é baixa. Se o vocabulário é variado, entropia é alta.

11. Conclusão

Modelos de linguagem baseados em n-gramas representam uma abordagem fundamental e historicamente importante no processamento de linguagem natural (PLN). Apesar de sua simplicidade em comparação com modelos neurais modernos, como transformers, eles oferecem uma base sólida para o entendimento da modelagem de sequências linguísticas, fornecendo intuições valiosas sobre probabilidade, entropia, e previsão de palavras.

O estudo desse modelo revela os desafios práticos da modelagem probabilística de texto, como a escassez de dados, o overfitting, e a necessidade de técnicas de suavização, ao mesmo tempo que introduz métricas rigorosas como perplexidade e entropia cruzada, fundamentais para avaliar o desempenho de qualquer modelo de linguagem.