Modelos de Semântica Vetorial e Embeddings de Palavras

1. Introdução

A semântica vetorial é uma abordagem computacional que representa palavras como vetores em um espaço multidimensional. Essa técnica permite que relações semânticas entre palavras sejam capturadas de forma quantitativa. Em vez de depender exclusivamente de regras linguísticas, os modelos vetoriais utilizam padrões de uso das palavras em grandes corpora de texto para inferir significado e similaridade.

Nesse relatório, abordarei os principais conceitos de embeddings estáticos, que associam a cada palavra um vetor fixo, e as diferentes formas de construí-los: modelos esparsos e densos.

2. Representação Vetorial de Palavras

No paradigma vetorial, cada palavra é representada como um ponto (vetor) em um espaço de alta dimensão, chamado de espaço de embeddings. A principal ideia é que palavras com significados semelhantes tendem a ocorrer em contextos semelhantes e, por isso, devem ter representações vetoriais próximas.

Esse tipo de representação permite calcular similaridade semântica entre palavras e documentos utilizando operações matemáticas simples, como produto escalar ou cosseno de ângulo entre vetores.

3. Embeddings Estáticos

Embeddings estáticos significam que cada palavra tem um único vetor fixo, independentemente do contexto em que aparece. Isso difere de modelos mais recentes (como BERT ou GPT), que produzem embeddings contextuais.

4. Modelos Esparsos

Os modelos esparsos foram uma das primeiras abordagens para representar palavras como vetores. Neles, cada vetor tem tantas dimensões quanto o tamanho do vocabulário. Ou seja, se o vocabulário tem 10.000 palavras, cada vetor também terá 10.000 posições.

Existem duas estruturas comuns:

- **Matriz termo-documento**: cada linha representa uma palavra (termo), e cada coluna representa um documento. Os valores das células são baseados em quantas vezes a palavra aparece em cada documento.
- **Matriz palavra-contexto (ou termo-termo)**: cada linha representa uma palavra-alvo e cada coluna representa uma palavra de contexto. Os valores refletem a frequência de coocorrência entre essas palavras.

4.1 Técnicas de ponderação

Para melhorar a qualidade semântica dos vetores, duas técnicas de ponderação são amplamente utilizadas:

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: ajusta a frequência da palavra com base em quão comum ela é no corpus. Palavras muito frequentes em todos os documentos (como "de", "o", "é") recebem pesos menores.
- **PPMI (Pointwise Positive Mutual Information)**: mede a associação estatística entre uma palavra-alvo e uma de contexto, destacando relações significativas de coocorrência. É comum em matrizes palavra-contexto.

Apesar de simples e interpretáveis, os modelos esparsos têm vetores muito grandes e com muitos zeros, o que dificulta seu uso em tarefas de aprendizado de máquina em larga escala.

5. Modelos Densos

Para superar as limitações dos modelos esparsos, surgiram os modelos densos, nos quais cada palavra é representada por um vetor de dimensão reduzida, geralmente entre 50 e 1000 dimensões.

Nesses vetores, os valores não têm um significado direto por dimensão, mas são aprendidos automaticamente a partir de grandes corpora de texto.

5.1 Word2Vec e o modelo Skip-gram

O Word2Vec é um dos modelos densos mais influentes. Ele aprende embeddings a partir do contexto de uso das palavras usando redes neurais simples.

O modelo Skip-gram é uma das variantes mais populares e funciona da seguinte forma:

- Para cada palavra-alvo, o modelo tenta prever as palavras de contexto ao seu redor.

- Isso é feito como um problema de classificação binária, usando regressão logística para estimar a probabilidade de uma palavra estar no contexto da outra.
- O treinamento utiliza gradiente descendente estocástico para ajustar os vetores, aumentando o produto escalar (dot product) entre palavras que ocorrem juntas e diminuindo-o para pares aleatórios (ruído).

5.2 GloVe (Global Vectors for Word Representation)

Outro modelo importante é o GloVe, que também gera embeddings densos, mas utiliza uma abordagem diferente. Ele constrói uma matriz de coocorrência de palavras e aprende embeddings com base na razão de probabilidades de coocorrência entre pares de palavras.

Enquanto o Word2Vec foca nos contextos locais (janelas de palavras), o GloVe captura padrões globais de coocorrência em todo o corpus.

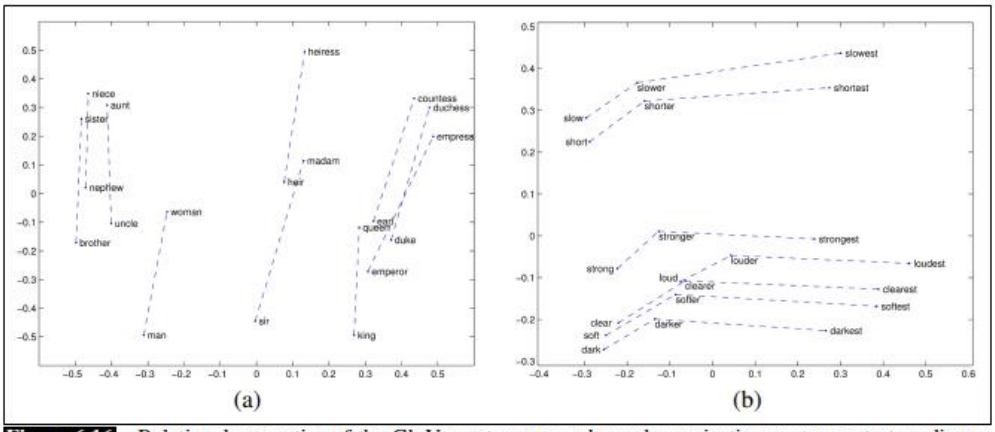

Figure 6.16 Relational properties of the GloVe vector space, shown by projecting vectors onto two dimensions. (a) $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$ is close to $\overrightarrow{queen}$. (b) offsets seem to capture comparative and superlative morphology (Pennington et al., 2014).

6. Medidas de Similaridade Vetorial

Tanto nos modelos esparsos quanto nos densos, a similaridade semântica entre palavras ou documentos pode ser calculada usando funções baseadas no produto escalar entre vetores.

A métrica mais comum é o cosseno de similaridade, que mede o ângulo entre dois vetores, ignorando suas magnitudes. Isso é útil porque muitas vezes queremos comparar a direção dos vetores (isto é, o significado), e não seu comprimento (frequência).

Valores próximos de 1 indicam alta similaridade, enquanto valores próximos de 0 indicam pouca ou nenhuma semelhança.

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{cherry}, \text{information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = .018$$

$$\cos(\text{digital}, \text{information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = .996$$

7. Aplicações Práticas

Modelos de embeddings são a base de muitas aplicações modernas de linguagem natural, como:

- **Sistemas de busca**: encontrar documentos ou respostas relevantes com base na similaridade semântica.
- **Análise de sentimentos**: representar textos como média dos vetores das palavras e aplicar classificadores.
- **Recomendações de produtos ou conteúdo**: por similaridade semântica entre descrições.
- **Tradução automática e question answering**: como parte do pipeline de entendimento semântico.

8. Conclusão

A semântica vetorial transformou a forma como representamos linguagem em sistemas computacionais. A transição de representações simbólicas para vetores numéricos permitiu uma nova era de modelos de aprendizado profundo e aplicações mais eficazes.

Resumo das principais ideias:

- As palavras podem ser representadas como vetores numéricos.
- Existem modelos **esparsos** (como TF-IDF e PPMI) e **densos** (como Word2Vec e GloVe).
- Modelos densos são mais eficientes e precisos, e capturam melhor relações semânticas.
- Similaridades semânticas são medidas principalmente por funções como o cosseno entre vetores.
- Esses modelos são fundamentais para diversas aplicações modernas de NLP.