Contrastive Search: How It Works

Step 1: Get top-k candidates

- city: 0.45
- town: 0.18
- area: 0.15
- place: 0.12

Step 2: Compute similarity to recent context

Context: "...the"

- city: 0.92 (cosine similarity)
- town: 0.75 (cosine similarity)
- area: 0.68 (cosine similarity)
- place: 0.60 (cosine similarity)

Step 3: Apply diversity penalty (α=0.6)

$$score = (1-\alpha) \times P(token) - \alpha \times similarity$$

- city: $0.4 \times 0.45 - 0.6 \times 0.92 = -0.372$
- **town: $0.4 \times 0.18 - 0.6 \times 0.75 = -0.378$**
- area: $0.4 \times 0.15 - 0.6 \times 0.68 = -0.348$
- place: $0.4 \times 0.12 - 0.6 \times 0.60 = -0.312$

Winner: "town" (high prob, lower similarity)

Key Insight: Balance probability (coherence) with diversity (novelty)

α=0: Pure greedy | α=0.6: Balanced | α=1.0: Maximum diversity