Decoding Strategies

Week 9 - From Greedy to Creative

NLP Course 2025

October 27, 2025

Two-Tier BSc Discovery

The Quality-Diversity Tradeoff

../figures/quality_diversity_tradeoff_bsc.pdf

Three Decoding Families

Deterministic	Stochastic	Controlled
Greedy, Beam	Temperature, Top-k	Nucleus (top-p)
Always same output	Random sampling	Balance both
High quality	Creative	Tunable creativity
No diversity	Can be nonsense	Modern standard

Different tasks need different decoding strategies

Beam Search: Explore Multiple Paths

../figures/beam_search_visual_bsc.pdf

Worked Example: Beam Search (width=3)

Task: Generate after "The cat"

Step 1: Top-3 words
sat (0.4), is (0.3), was (0.2)
Keep 3 hypotheses

Step 2: Expand each
"The cat sat" → on (0.5), there (0.3)
"The cat is" → sleeping (0.6), black (0.2)
"The cat was" → happy (0.4), tired (0.3)

Step 3: Score and prune
Total scores: sat+on (0.2), is+sleeping (0.18), sat+there (0.12)
Keep top-3, continue...

Final: "The cat is sleeping" wins!

Beam search finds better sequences than greedy

Temperature: Control Randomness

../figures/temperature_effects_bsc.pdf

Worked Example: Temperature Scaling

Logits: [2.0, 1.0, 0.5, 0.2]

T=0.5 (focused):

$$p_i = \frac{\exp(logit_i/0.5)}{\sum \exp(logit_j/0.5)}$$

Result: [0.61, 0.22, 0.11, 0.06] - peaked!

T=1.0 (normal):

$$p_i = \text{softmax}(logits)$$

Result: [0.42, 0.23, 0.16, 0.13] - balanced

T=2.0 (flat):
Result: [0.32, 0.26, 0.23, 0.19] - uniform!

Lower T = more deterministic. Higher T = more random.

Key Takeaways

1. Beam search: Deterministic, high quality, no diversity
2. Temperature: Simple randomness control
3. Top-k: Filter unlikely words, sample from top
4. Nucleus (top-p): Dynamic cutoff, modern standard
5. Choose based on task: translation=beam, creative=sampling

Decoding strategy matters as much as model quality

Technical Appendix