

Week 1: Foundations & Statistical Language Models

Building the Groundwork for NLP

Natural Language Processing Course

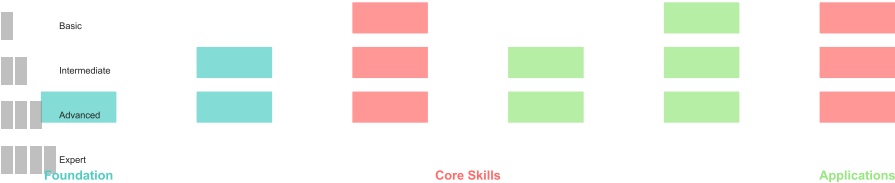
BSc Computer Science

What We'll Learn This Week

Week 1 Learning Journey



Difficulty:



Key Concept: N-gram Language Models

N-gram Extraction

Unigrams: The quick brown fox jumps

Bigrams: The quick quick brown brown fox fox jumps

Trigrams: The quick brown quick brown fox brown fox jumps

Probability Calculation

$P(\text{jumps} \mid \text{brown fox})$

brown fox 10

... jumps 3

... runs 2

... walks 5

$$P(\text{jumps} \mid \text{brown fox}) = 3/10 = 0.3$$

The Markov Assumption

- Future depends only on recent past
- $P(w_i \mid w_1 \dots w_{i-1}) \approx P(w_i \mid$

Probability Estimation

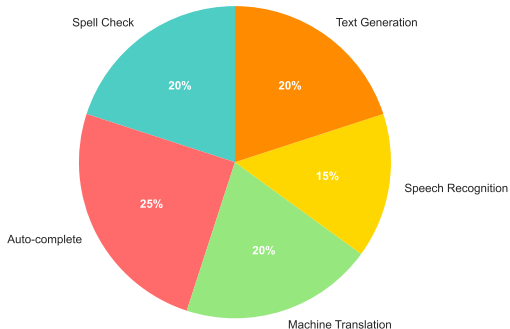
- Count n-grams in corpus
- Maximum likelihood:

Challenges

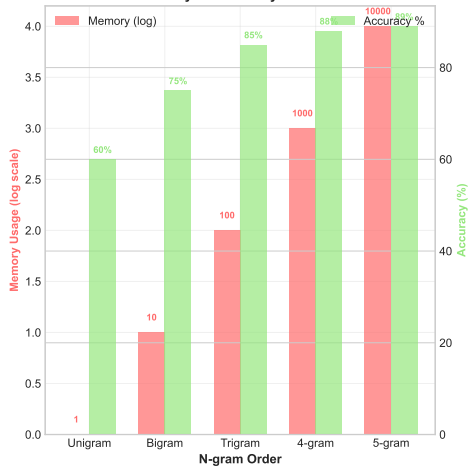
- Data sparsity
- Storage requirements

Applications & Real-World Impact

N-gram Applications



Memory vs Accuracy Trade-off



Where N-grams Excel:

- **Spell Checking:** Find likely corrections
- **Auto-complete:** Predict next words

Historical Significance:

- Dominated NLP for decades (1980s-2000s)
- Still used in hybrid systems