# Homework 1: $n$-gram Language Models — Solutions

CS 388: Natural Language Processing • Fall 2026

---

Corpus $\mathcal{D}$ (10 sentences), $|V| = 13$ (12 word types + $\langle$/s$\rangle$). Context counts used below: $C(\langle s\rangle) = C(\langle s\rangle\,\langle s\rangle) = 10$, and $C(\texttt{the}) = 14$, $C(\texttt{cat}) = 7$, $C(\texttt{dog}) = 6$, $C(\texttt{a}) = C(\texttt{saw}) = 3$, $C(\texttt{mouse}) = C(\texttt{chased}) = C(\texttt{sat}) = C(\texttt{on}) = C(\texttt{mat}) = C(\texttt{slept}) = 2$, $C(\texttt{barked}) = 1$.

## 1. Counting

**(a)** Full bigram and trigram count tables.

**Solution.** All bigrams with nonzero count (27 types; each column is $w_{i-1}\,w_i$ : count):

- $\langle s\rangle$ `the` : 7
- $\langle s\rangle$ `a` : 3
- `the cat` : 6
- `the dog` : 4
- `the mat` : 2
- `the mouse` : 2
- `a dog` : 2
- `a cat` : 1
- `cat` $\langle/s\rangle$ : 3
- `cat chased` : 1
- `cat sat` : 1
- `cat saw` : 1
- `cat slept` : 1
- `dog` $\langle/s\rangle$ : 2
- `dog barked` : 1
- `dog chased` : 1
- `dog sat` : 1
- `dog saw` : 1
- `chased the` : 2
- `saw the` : 3
- `sat on` : 2
- `on the` : 2
- `mouse saw` : 1
- `mouse slept` : 1
- `mat` $\langle/s\rangle$ : 2
- `slept` $\langle/s\rangle$ : 2
- `barked` $\langle/s\rangle$ : 1

All trigrams with nonzero count (36 types):

- $\langle s\rangle\,\langle s\rangle$ `the` : 7
- $\langle s\rangle\,\langle s\rangle$ `a` : 3
- $\langle s\rangle$ `the cat` : 3
- $\langle s\rangle$ `the dog` : 2
- $\langle s\rangle$ `the mouse` : 2
- $\langle s\rangle$ `a dog` : 2
- $\langle s\rangle$ `a cat` : 1
- `the cat` $\langle/s\rangle$ : 3
- `the cat chased` : 1
- `the cat sat` : 1
- `the cat slept` : 1
- `the dog` $\langle/s\rangle$ : 2
- `the dog chased` : 1
- `the dog sat` : 1
- `the mat` $\langle/s\rangle$ : 2
- `the mouse saw` : 1
- `the mouse slept` : 1
- `a cat saw` : 1
- `a dog barked` : 1
- `a dog saw` : 1
- `cat chased the` : 1
- `cat sat on` : 1
- `cat saw the` : 1
- `cat slept` $\langle/s\rangle$ : 1
- `dog chased the` : 1
- `dog sat on` : 1
- `dog saw the` : 1
- `dog barked` $\langle/s\rangle$ : 1
- `chased the cat` : 1
- `chased the dog` : 1
- `saw the cat` : 2
- `saw the dog` : 1
- `sat on the` : 2
- `on the mat` : 2
- `mouse saw the` : 1
- `mouse slept` $\langle/s\rangle$ : 1

**(b)** Why do we need the padding tokens?

**Solution.** $\langle s\rangle$ lets the model assign a sentence-initial distribution: without it $w_1$ has no context, so the model cannot say which words are likely to *start* a sentence (and the trigram model has no context for $w_2$ either). $\langle/s\rangle$ gives the model a stopping probability; without it the model never terminates a sentence, the scores over strings of all lengths do not sum to 1, and sentences of different lengths are not comparable (shorter strings are always favored, since every extra factor is $\leq 1$).

## 2. Sentence Probabilities

Each table lists one factor per predicted token, as (raw ratio) = (value).

**(a)** $S_1$ = `the dog chased the dog.`

**Solution.**

| predicted $w_i$ | $P_{\text{bi}}(w_i \mid w_{i-1})$ | $P_{\text{tri}}(w_i \mid w_{i-2}w_{i-1})$ |
|---|---|---|
| `the` | $7/10$ | $7/10$ |
| `dog` | $4/14 = 2/7$ | $2/7$ |
| `chased` | $1/6$ | $1/4$ |
| `the` | $2/2 = 1$ | $1/1 = 1$ |
| `dog` | $4/14 = 2/7$ | $1/2$ |
| $\langle/\text{s}\rangle$ | $2/6 = 1/3$ | $2/4 = 1/2$ |
| product | $\mathbf{1/315 \approx 3.17 \times 10^{-3}}$ | $\mathbf{1/80 = 1.25 \times 10^{-2}}$ |

Both are well defined and nonzero: every bigram and trigram of $S_1$ occurs in $\mathcal{D}$, even though $S_1$ itself does not. The trigram value is larger because its contexts are more specific and therefore have far fewer observed continuations (e.g. `dog chased` is always followed by `the`).

**(b)** $S_2$ = `the cat chased the mouse.`

**Solution.**

| predicted $w_i$ | $P_{\text{bi}}$ | $P_{\text{tri}}$ |
|---|---|---|
| `the` | $7/10$ | $7/10$ |
| `cat` | $6/14 = 3/7$ | $3/7$ |
| `chased` | $1/7$ | $1/6$ |
| `the` | $2/2 = 1$ | $1/1 = 1$ |
| `mouse` | $2/14 = 1/7$ | $\mathbf{0/2 = 0}$ |
| $\langle/\text{s}\rangle$ | $\mathbf{0/2 = 0}$ | $\mathbf{0/2 = 0}$ |
| product | $\mathbf{0}$ | $\mathbf{0}$ |

Both are 0 but both are *well defined*: every conditioning context is attested, so no factor is $\frac{n}{0}$. The bigram model is killed by `mouse` $\langle/\text{s}\rangle$ (count 0; `mouse` never ends a training sentence), the trigram model additionally by `chased the mouse` (count 0).

**(c)** $S_3$ = `the mouse chased the dog.`

**Solution.**

| predicted $w_i$ | $P_{\text{bi}}$ | $P_{\text{tri}}$ |
|---|---|---|
| `the` | $7/10$ | $7/10$ |
| `mouse` | $2/14 = 1/7$ | $2/7$ |
| `chased` | $\mathbf{0/2 = 0}$ | $\mathbf{0/2 = 0}$ |
| `the` | $2/2 = 1$ | $\mathbf{0/0 = \text{undefined}}$ |
| `dog` | $4/14 = 2/7$ | $1/2$ |
| $\langle/\text{s}\rangle$ | $2/6 = 1/3$ | $2/4 = 1/2$ |
| product | $\mathbf{0}$ | **undefined** |

$P_{\text{bi}}(S_3) = 0$, well defined: the context `mouse` was seen ($C(\texttt{mouse}) = 2$), the continuation `mouse chased` was not. $P_{\text{tri}}(S_3)$ is undefined: the context `mouse chased` has count 0, so $P(\texttt{the} \mid \texttt{mouse chased})$ divides by zero. This is the key contrast with $S_2$ — zero probability is a claim the model makes, while an unseen context means the model has no estimate at all.

## 3. Laplace Smoothing and Backoff

*Convention used.* $P_{\text{bo}}$ uses the add-1 trigram estimate when $C(w_{i-2}w_{i-1}) > 0$, and otherwise backs off to the add-1 bigram estimate $P^{+1}_{\text{bi}}(w_i \mid w_{i-1})$; add-1 denominators are $C(\text{context}) + |V| = C(\text{context}) + 13$.

**(a)** $P^{+1}_{\text{bi}}(S_2)$ and $P_{\text{bo}}(S_2)$.

**Solution.**

| predicted $w_i$ | $P^{+1}_{\text{bi}}$ | $P_{\text{bo}}$ |
|---|---|---|
| `the` | $8/23$ | $8/23$ |
| `cat` | $7/27$ | $4/20 = 1/5$ |
| `chased` | $2/20 = 1/10$ | $2/19$ |
| `the` | $3/15 = 1/5$ | $2/14 = 1/7$ |
| `mouse` | $3/27 = 1/9$ | $1/15$ |
| $\langle/\text{s}\rangle$ | $1/15$ | $1/15$ |
| product | $\mathbf{28/2095875 \approx 1.34 \times 10^{-5}}$ | $\mathbf{16/3441375 \approx 4.65 \times 10^{-6}}$ |

Both are now nonzero. No backoff was triggered: every trigram context in $S_2$ is attested. The two previously-zero factors (1/15 each) now carry the smallest pseudo-count mass.

**(b)** $P^{+1}_{\text{bi}}(S_3)$ and $P_{\text{bo}}(S_3)$.

**Solution.**

| predicted $w_i$ | $P^{+1}_{\text{bi}}$ | $P_{\text{bo}}$ |
|---|---|---|
| `the` | $8/23$ | $8/23$ |
| `mouse` | $3/27 = 1/9$ | $3/20$ |
| `chased` | $1/15$ | $1/15$ |
| `the` | $3/15 = 1/5$ | $3/15 = 1/5$ (**backoff**) |
| `dog` | $5/27$ | $2/15$ |
| $\langle/\text{s}\rangle$ | $3/19$ | $3/17$ |
| product | $\mathbf{8/530955 \approx 1.51 \times 10^{-5}}$ | $\mathbf{4/244375 \approx 1.64 \times 10^{-5}}$ |

Exactly one factor backs off: $P(\texttt{the} \mid \texttt{mouse chased})$, since $C(\texttt{mouse chased}) = 0$. It is replaced by the add-1 *bigram* estimate $P^{+1}_{\text{bi}}(\texttt{the} \mid \texttt{chased}) = (2+1)/(2+13) = 1/5$, read off the bigram table. $S_3$ is now defined *and* nonzero, and (unlike under MLE) ranks above $S_2$.

**(c)** $P^{+1}_{\text{bi}}(S_1)$ vs. $P_{\text{bi}}(S_1)$.

**Solution.** $P^{+1}_{\text{bi}}(S_1) = \frac{8}{23} \cdot \frac{5}{27} \cdot \frac{2}{19} \cdot \frac{3}{15} \cdot \frac{5}{27} \cdot \frac{3}{19} = 80/2017629 \approx 3.97 \times 10^{-5}$, about $80\times$ *smaller* than the MLE value $1/315 \approx 3.17 \times 10^{-3}$. Smoothing moves probability mass off the observed continuations and onto the $|V| = 13$ unseen ones for every context; since the corpus is tiny, the added pseudo-counts dominate the real counts (e.g. $C(\texttt{the}) = 14$ vs. $+13$, nearly halving every estimate conditioned on `the`, and $P(\texttt{the} \mid \texttt{chased})$ drops from 1 to 1/5). Add-1 is therefore far too aggressive here; a smaller $k$ or a backoff/interpolation scheme with discounting is preferable.