Homework 1: n-gram Language Models

CS 388: Natural Language Processing • Fall 2026

Instructions. This assignment can be done entirely by hand – no coding required; however, please transcribe your answers/work into the provided LaTeXfile and submit the PDF, rather than a handwritten response. **Please use the `\B{}` command to wrap your solutions, making them blue (this will help us when grading).** Show your work. Report probabilities as exact fractions (e.g. $\frac{3}{14}$) rather than decimals, and simplify where convenient. Submit a single PDF to Gradescope.

1. Setup and Notation

Throughout, we use the following corpus of 10 sentences as our training data. Assume the text has already been tokenized and lowercased, so each whitespace-separated string is one token.

Training corpus $\mathcal{D}$			
1.	`the cat chased the dog`	6.	`a dog saw the cat`
2.	`the dog chased the cat`	7.	`the mouse saw the cat`
3.	`the cat sat on the mat`	8.	`the mouse slept`
4.	`the dog sat on the mat`	9.	`the cat slept`
5.	`a cat saw the dog`	10.	`a dog barked`

Padding. Every sentence is padded with sentence-boundary symbols. For the *bigram* model, pad each sentence with one $\langle$s$\rangle$ on the left and one $\langle$/s$\rangle$ on the right; for the *trigram* model, pad with two $\langle$s$\rangle$ on the left and one $\langle$/s$\rangle$ on the right. So sentence 8 is treated as

$$\langle \text{s} \rangle \ \langle \text{s} \rangle \ \texttt{the mouse slept} \ \langle \text{/s} \rangle.$$

The symbol $\langle$/s$\rangle$ is predicted like any other token (this is what makes the model a proper distribution over sentences), but $\langle$s$\rangle$ is never predicted – it only ever appears in a conditioning context.

Vocabulary. Let V denote the vocabulary of *possible next tokens*: the 12 word types that occur in $\mathcal{D}$, plus $\langle$/s$\rangle$, and *not* $\langle$s$\rangle$. Hence $|V| = 13$. Use this value wherever $|V|$ appears.

Convention. We consider any probability of the form $\frac{n}{0}$ to be undefined.

2. Counting (30 points)

In class we saw the maximum likelihood estimate (MLE) for n-gram models, based on counting occurrences. Your first task is to build the count tables you will need for this estimator. Include the padding symbols in your counts.

(a) **(25 pts)** Report the full bigram and trigram count tables. Tuples with count zero may be omitted.

Example (bigram)

Bigram	Count
`the cat`	6
...	...

(b) **(5 pts)** Why do we need the padding tokens ⟨s⟩ and ⟨/s⟩? What happens if we do not include them?

3. Sentence Probabilities (40 points)

Consider the following three test sentences.

S_1	`the dog chased the dog`
S_2	`the cat chased the mouse`
S_3	`the mouse chased the dog`

For each part below, write out the full product of conditional probabilities, and their corresponding values, as well as the final product.

(a) **(12 pts)** Compute $P_{\text{bi}}(S_1)$ and $P_{\text{tri}}(S_1)$.

(b) **(14 pts)** Compute $P_{\text{bi}}(S_2)$ and $P_{\text{tri}}(S_2)$.

(c) **(14 pts)** Compute $P_{\text{bi}}(S_3)$ and $P_{\text{tri}}(S_3)$.

4. Laplace Smoothing and Backoff (30 points)

In class we saw both Laplace smoothing and Backoff. As a reminder:

Add-k (Laplace) smoothing, $k = 1$. Add one pseudo-count to every possible continuation:

$$P_{\text{bi}}^{+1}(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i) + 1}{C(w_{i-1}) + V}, \qquad P_{\text{tri}}^{+1}(w_i \mid w_{i-2}w_{i-1}) = \frac{C(w_{i-2}w_{i-1}w_i) + 1}{C(w_{i-2}w_{i-1}) + V}.$$

Backoff. Adding 1 to the numerator does not help address problems stemming from contexts we have never observed. This is especially problematic as we increase n: for a trigram model, there will be many contexts we have never seen. In this case, we can back-off an n-gram to an $n-1$-gram, if we have not seen the particular n-gram in question. We can repeat this until we get to unigrams.

For this assignment, to compute backoff , first use the Laplace-smoothed trigram estimate whenever the trigram conditioning context has been observed. If the bigram context $w_{i-2}w_{i-1}$ has never been observed, i.e. $C(w_{i-2}w_{i-1}) = 0$, we back off to the Laplace-smoothed bigram estimate. **Smoothing is therefore applied in all cases**:

$$P_{\text{bo}}(w_i \mid w_{i-2}w_{i-1}) = \begin{cases} \dfrac{C(w_{i-2}w_{i-1}w_i) + 1}{C(w_{i-2}w_{i-1}) + V}, & \text{if } C(w_{i-2}w_{i-1}) > 0, \\ \dfrac{C(w_{i-1}w_i) + 1}{C(w_{i-1}) + V}, & \text{if } C(w_{i-2}w_{i-1}) = 0. \end{cases}$$

(a) **(11 pts)** Recompute $P_{\text{bi}}^{+1}(S_2)$ and $P_{\text{bo}}(S_2)$, again showing every factor.

(b) **(11 pts)** Recompute $P_{\text{bi}}^{+1}(S_3)$ and $P_{\text{bo}}(S_3)$.

(c) **(8 pts)** Recompute $P_{\text{bi}}^{+1}(S_1)$ and compare it to $P_{\text{bi}}(S_1)$ from Part 2. What difference do you observe, and what is the explanation for the direction of that difference?

What to turn in. A single PDF containing your count tables (Part 1), the full factor-by-factor derivations for all six probability computations in Part 2 and all five in Part 3, and your written answers to the discussion questions.