CSCI 8115 Fall 2023
Hardik Gupta

Authorship Attribution

Language Models

Language models assign a probability to a sequence of words or tokens, effectively predicting the likelihood of the next token in a sequence. While Large Language Models (LLMs) have seen a surge in popularity in recent years due to their impressive capabilities, **n-gram models** remain a robust and interpretable subset of language models that are still valuable in certain contexts.

N-grams

N-grams are probabilistic models that assign a probability to a given word based on the preceding $n-1$ words. Formally, an n-gram model calculates the probability of a sequence of words as follows:

$$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2) \cdots P(w_n \mid w_1, \ldots, w_{n-1}) = \prod_{k=1}^{n} P(w_k \mid w_1, \ldots, w_{k-1})$$

This is known as the **chain rule of probability**. However, as n increases, it becomes impractical to estimate the probability of each sequence due to the **sparsity** of language data—i.e., the chances of encountering a particular sequence of words, even in a large corpus, are low due to the creativity and subjectivity of natural language. Therefore, we often approximate these probabilities using shorter contexts, such as bigrams ($n = 2$) or trigrams ($n = 3$).

For an n-gram model, the probability of the next word is calculated based on the previous $n-1$ words:

$$P(w_n \mid w_{n-1}, \ldots, w_1) \approx \frac{C(w_{n-1}, \ldots, w_1, w_n)}{C(w_{n-1}, \ldots, w_1)}$$

where $C(w_{n-1}, \ldots, w_1, w_n)$ represents the count of occurrences of the sequence $(w_{n-1}, \ldots, w_1, w_n)$ in the training corpus, and $C(w_{n-1}, \ldots, w_1)$ represents the count of the preceding sequence. However, relying purely on counts often leads to issues with **zero probabilities** for unseen sequences, which can be mitigated using **smoothing techniques** such as Laplace smoothing or Kneser-Ney smoothing.

Evaluating Language Models: Perplexity

One common metric used to evaluate language models is **perplexity**. Perplexity measures how well a probability model predicts a sample and is defined as the exponentiated average negative log-likelihood of the sequence:

$$PP(W) = P(w_1, w_2, \ldots, w_n)^{-\frac{1}{n}} = 2^{-\frac{1}{n}\sum_{i=1}^{n} \log_2 P(w_i|w_1,\ldots,w_{i-1})}$$

A lower perplexity indicates a better predictive model, as it means the model assigns higher probabilities to the observed data.

References

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*, 3rd edition, Section 3.5. `https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf`