HW2: Language Modeling

Emily Tseng
et397@cornell.edu

February 17, 2020

1 Introduction

In a sense, a language is just a set of community-driven rules for what words should appear in what order. Language modeling is a generative problem that attempts to formalize this ordering by computing a probability distribution over a given sequence. As an example, consider the incomplete sentence "*Tickets were sold out for the 2 p.m.* _____" A language model properly trained over the English language would compute a greater probability for the sentence ending in "*movie*" than, say, the sentence ending in "*banana*".

The key is to generate a distribution as close as possible to the true distribution of your source language; or, in empirical terms, to maximize the probability (or minimize the perplexity) of the sentences in your test set. In this homework, we use data from the Penn Treebank (Marcus et al., 1993) to implement, train and assess several varieties of language models:

1. A count-based trigram model with linear interpolation
2. A neural network language model (Bengio et al., 2003)
3. A Long Short-Term Memory (LSTM) language model (Zaremba et al., 2014)

For each of the above models, we report perplexity on the validation and testing sets. We additionally consider the following two extensions:

1. Consider 3-4 train and test contexts and do an information-theoretic analysis of the output of the model. What is the entropy of the distribution? What is the KL between different models, and what does this tell us about the predictions and the contexts?
2. Formulate an EM algorithm for estimating α for the count-based model. That is, imagine there is a latent Z categorical variable over three choices, where $p(Z)$ is a prior prob of using a unigram, bigram, or trigram model. This formulation is identical to the one above, but allows us to learn the α values using EM. The approach has two steps: Start with random alpha, 1) Learn a model with that alpha, 2) *On validation* compute $p(z_t|x_t, x_{t-1}, x_{t-2})$ for each position to use as new alphas.

2 Problem Description

Consider a sentence as a sequence of tokens $x_{1:T}$ for a sentence of length T where all tokens are drawn from the same unigram distribution V. In a given language, the probability of the sentence can be described as:

$$p(x_{1:T}) = \prod_{i=1}^{T} p(x_T | x_1 ... x_{T-1})$$

The goal of a language model is to estimate a distribution that maximizes the probabilities of the sentences in a provided test set. From information theory, we have the following ways to quantify a distribution: *entropy* (H), a measure of deviance from the uniform distribution, and *perplexity* (PP), the exponentiated entropy of a distribution and the measure commonly used in NLP. Formally defined, for a distribution p with observations $x \in X$:

$$\begin{aligned} H(p) &= -\sum_i p(x = i) \log[p(x = i)] \\ &= -\mathop{\mathbb{E}}_{x \in p} \log[p(x)] \\ PP(p) &= \exp H(p) \end{aligned}$$

The goal of this assignment will be to train language models that mimic the perplexities reported in the source papers as closely as possible.

3 Model and Algorithms

3.1 Count-based Trigram Model with Linear Interpolation

$$p(y_t | y_{1:t-1}) = \alpha_1 p(y_t | y_{t-2}, y_{t-1}) + \alpha_2 p(y_t | y_{t-1}) + (1 - \alpha_1 - \alpha_2) p(y_t)$$

3.2 EM Algorithm for a Estimation

3.3 Neural Network Language Model

3.4 LSTM Language Model

4 Experiments

5 Conclusion

End the write-up with a very short recap of the main experiments and the main results. Describe any challenges you may have faced, and what could have been improved in the model.

References

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.