Sampling Informative Training Data for RNN Language Models

Jared Fernandez, Doug Downey
jared.fern@u.northwestern.edu, ddowney@eecs.northwestern.edu
Department of Electrical Engineering and Computer Science, Northwestern University

Introduction

We seek to determine if it is possible to select a training set substantially more informative than a set randomly drawn sentences. By selectively training on high information sentences, language models can learn the language distribution faster and more accurately than models trained on randomly sampled training sets.

Our approach preferentially samples high perplexity sentences, as determined by an easily queryable n-gram language model. RNNLMs are then trained with corrective importance weights to remove sampling bias.

Methodology

1. Train an n-gram model on randomly sampled sentences from the corpus.
2. Determine n-gram perplexities for each of the remaining sequences
3. Sample training sequences using a distribution determined as a function of the calculated n-gram perplexities
4. Train on each sequence s with weight $w_s = \frac{1}{Pr_{keep}(s)}$

Importance Sampling Distributions

We propose multiple sampling distributions for selecting training sequences according to their n-gram perplexity.

- **Z_{Full} Sampling:**

$$Pr_{Z_{Full}}(s) \propto \left|\alpha \frac{ppl(s) - \mu_{ppl}}{\sigma_{ppl}} + 1\right|$$

- **Z_α Sampling:**

$$Pr_{Z_\alpha}(s) \propto \begin{cases} \alpha^{\frac{ppl(s)-\mu_{ppl}}{\sigma_{ppl}}} + 1, & \text{if } ppl(s) > \mu_{ppl} \\ 1, & \text{else} \end{cases}$$

- **Z^2 Sampling:**

$$Pr_{Z^2}(s) \propto \begin{cases} \alpha \left(\frac{ppl(s)-\mu_{ppl}}{\sigma_{ppl}}\right)^2 + 1, & \text{if } ppl(s) > \mu_{ppl} \\ 1, & \text{else} \end{cases}$$

μ_{ppl}, σ_{ppl} : Sentence ngram ppl mean and standard deviation

Sequences with perplexities in the 100th percentile were generally esoteric, and were assigned boosted selection probability.

Results

Model	Tokens	μ_{ngram}	σ_{ngram}	RNN Ppl
n-gram	500k	—	—	492.3
Random	500k	449.0	346.4	749.1
$Z_{0.5}$	500k	497.1	398.8	643.9
$Z_{1.0}$	500k	544.1	440.1	645.2
$Z_{2.0}$	500k	615.7	481.3	593.2
$Z_{4.0}$	500k	729.0	523.6	**571.4**
Z^2	500k	576.5	499.7	720.0
Z_{full}	500k	627.1	451.9	663.7
n-gram	1M	—	—	502.7
Random	1M	448.9	380.2	550.6
$Z_{0.5}$	1M	495.7	431.8	545.7
$Z_{1.0}$	1M	540.4	475.4	435.4
$Z_{2.0}$	1M	615.6	528.4	426.9
$Z_{4.0}$	1M	732.9	584.4	420.1
Z^2	1M	571.5	535.7	435.7
Z_{Full}	1M	608.6	489.9	**416.3**
n-gram	2M	—	—	502.6
Random	2M	430.45	392.1	341.3
$Z_{0.5}$	2M	471.8	445.2	292.7
$Z_{1.0}$	2M	514.6	493.9	289.8
$Z_{2.0}$	2M	582.8	544.6	346.9
$Z_{4.0}$	2M	684.6	604.7	294.6
Z^2	2M	518.4	522.9	**287.9**
Z_{Full}	2M	568.4	506.5	312.5

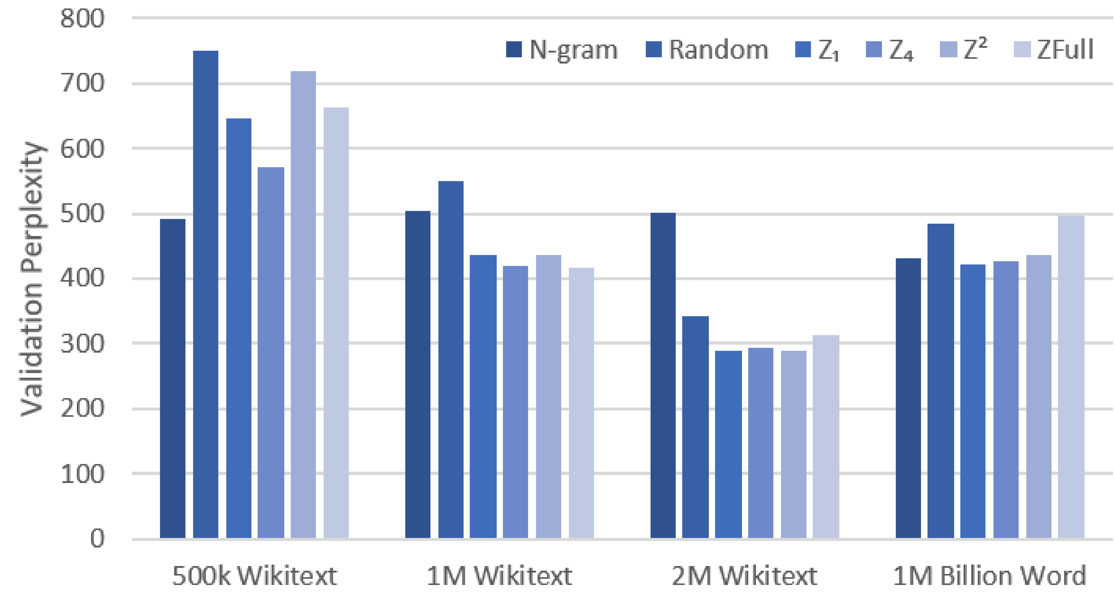
Table 1: Perplexities for Wikitext models.

Model	Tokens	μ_{ngram}	σ_{ngram}	RNN Ppl
n-gram	1M	—	—	432.5
Random	1M	433.2	515.4	484.0
$Z_{0.5}$	1M	476.8	410.9	436.6
$Z_{1.0}$	1M	543.8	529.0	**421.5**
$Z_{4.0}$	1M	726.4	517.3	427.3
Z_{full}	1M	635.19	458.69	495.75
Z^2	1M	639.2	593.7	435.3

Table 2: Perplexities for Billion Word models.

Experimental Details

- **Datasets & Vocab Size:**
  - Wikitext-103 (267k), One Billion Word (250k)
- **Models:**
  - *RNNLMs:*
    - 2-layer LSTM network with 200D hidden and embedding layers
    - Trained for 10 epochs with batch size of 12.
  - *N-grams:* Order 5 with Kneyser-Ney discounting.



Conclusions

- Selecting training sequences with higher average n-gram perplexity reduces the perplexity of the resulting RNNLM
- Low perplexity sequences should be selected with relatively high probability
  - Likely because low perplexity sequences contain subsequences shared with rare sequences

Future Work

- Alternative sampling distributions based on:
  - Sentence perplexity calculated by a pilot RNNLM
  - Sentence's unique n-gram content
- Sampling in a streaming setting and periodically updating the sampling distribution

Acknowledgements

Research supported in part by NSF grants IIS-1351029 & the Allen Institute for Artificial Intelligence. Support for travel provided in part by the ACL Student Travel Grant (NSF Grant IIS-1827830).