Contrastive Perplexity: A new evaluation metric for sentence level language models

Kushal Arora
kshel22@ufl.edu

arXiv:1601.00248v1 [cs.CL] 3 Jan 2016

Abstract

Perplexity(per word) is the most widely used metric for evaluating language models. This is mostly due to a its ease of computation, lack of dependence on external tools like speech recognition pipeline and a good theoretical justification for why it should work. Despite this, there has been no dearth of criticism for this metric. Most of this criticism center around lack of correlation with extrinsic metrics like word error rate(WER), dependence upon shared vocabulary for model comparison and unsuitability for un-normalized language model evaluation. In this paper we address the last problem of inability to evaluate un-normalized models by introducing a new discriminative evaluation metric that predicts model's performance based on its ability to discriminate between test sentences and their deformed version. Due to its discriminative formulation, this approach can work with un-normalized probabilities while retaining perplexity's ease of computation. We show a strong correlation between our new metric and perplexity across a range of models on WSJ datasets. We also hypothesize a stronger correlation between WER and our new metric vis-a-vis perplexity due to similar discriminative objective.

1 Introduction

There are two standard evaluation metrics for language models: perplexity or word error rate(WER). The simpler of these measures, WER, is simply the percentage of erroneously recognized words E (deletions, insertions, substitutions) to total number of words N, in a speech recognition task i.e.

$$WER = \frac{E}{N} \times 100\%$$

The second metric, perplexity(per word), is an information theoretic measure that evaluates the similarity of proposed model m to the original distribution p. It can be computed as a inverse of (geometric)average probability of test set T

$$\begin{aligned} PPL(D) &= \sqrt[N]{\frac{1}{m(T)}} \\ &= 2^{-\frac{1}{N}lg(m(T))} \end{aligned} \quad (1)$$

where N is the number of words in test set T.

Equation 1 can be seen as exponentiated cross entropy, where cross entropy $H(p, m)$ is approximated as

$$H(p, m) = -\frac{1}{N}lg(m(T))$$

In many ways, WER is an better metric as any improvement on language modeling benchmarks is meaningful only if it translates in to improvements in Automatic Speech Recognition(ASR) or Machine Translation. The problem with WER is that it needs a complete ASR pipeline to evaluate. Also, almost all benchmarking datasets are behind pay-wall, hence not readily available for evaluation.

Perplexity, on the other hand, is a theoretically elegant and easy to compute metric which correlates well with WER for simpler n-gram models. This makes PPL a good substitute for WER when evaluating n-grams model. For more complex language models, the correlation is not so strong[]. In addition to this, perplexity is an unsuitable metric to evaluate un-normalized models like sentence level models for which partition function computation is intractable. Also, two compare two models using perplexity, they must share the same vocabulary.

Most of the previous work done to improve upon perplexity has been focused on achieving better correlation with WER. Iyer et al. (Iyer et al., 1997) proposed a decision tree based metric that uses additional features like word length, POS tags and phonetic length of words to improve the WER correlation. Chen et al. (Chen et al., 1998) propose a new metric *M-ref* in which they attempt to learn likelihood curve between WER and perplexity. Clarkson et al.(Clarkson et al., 1999) attempt to use entropy in conjugation with perplexity, empirically learning mixing coefficient.

In this paper we focus on a different problem of extending perplexity to enable it to be used for un-normalized language models. We do so by introducing a discriminative approach to language model evaluation. Our approach is very much inspired by Contrastive Estimation by (Smith and Eisner, 2005) and works on the philosophy that a superior language model would be able better to distinguish between the sentence from the test set and its slightly deformed version.

In next Section, we derive our new metric, Contrastive perplexity, from scratch and give an intuitive understanding of why it should work. In Section 4, we analyze this new metric across various models on most widely used datasets, namely, Pen-TreeBank section of WSJ dataset and Brown Corpus. We report a very strong correlation between our new metric and the perplexity. We conclude this paper by hypothesizing a better correlation between WER and contrastive perplexity due to similar objective of minimizing the errors in prediction.

2 Contrastive Entropy

Let T be the test set. We pass this test set through a noise channel and let the distorted version of test set be $\hat{T}$. Now, we define Contrastive Entropy can as:

$$\begin{aligned} H_C(T) &= H(\hat{T}) - H(T) \\ &= -\frac{1}{N} lg \left(\frac{p(\hat{T})}{p(T)} \right) \\ &= -\frac{1}{N} lg \left(\frac{\tilde{p}(\hat{T})/Z}{\tilde{p}(T)/Z} \right) \\ &= -\frac{1}{N} lg \left(\frac{\tilde{p}(\hat{T})}{\tilde{p}(T)} \right) \end{aligned}$$

N here is size of test set.

Now, using the definition of Contrastive Entropy Rate, we calculate Contrastive Perplexity as:

$$PPL_C(T) = 2^{-H_c(T)}$$

For a sentence level language model, the probability distribution $m(D)$ can be modeled as product of sentence probabilities, i.e.

$$m(D) = \prod_{w_d \in D} m(w_d)$$

Now, contrastive entropy can be modeled as

$$\begin{aligned} H_c(D) &= -\frac{1}{N} \sum_{w_d \in D} lg \left(\frac{m(w_d)}{m(w_d)} \right) \\ &= -\frac{1}{N} \sum_{w_d \in D} lg \left(\frac{\tilde{m}(w_d)}{\tilde{m}(w_d)} \right) \end{aligned}$$

where $\tilde{m}$ is the un-normalized probability of sentence w_d.

The intuition behind our evaluation technique is that the distorted sentence $\hat{W}$, should be seen as a out of domain text and that a better probability model should be able to distinguish between an in-domain sentence from the language versus a malformed sentence that is less likely to be generated by the language.

Now we look at distortion generation mechanism. We allow only two type of distortions: substitution and transpositions. For substitutions, we randomly select a word from the vocabulary to substitute. For transposition, we randomly select a word from the same sentence to swap. For each word in a sentence there are three possible outcomes: no distortion with probability $x_{\mathcal{N}}$, substitution with probability x_S and transposition with probability x_T such that $x_{\mathcal{N}} + x_S + x_T = 1$.

3 Results

Pen TreeBank corpus is one of the most widely used dataset in statistical modeling community to report perplexity results. We use Pen TreeBank dataset with following split and preprocessing: Sections 0-20 were used as training data, sections 21-22 for validation and 23-24 for testing. The training, validation and testing token sizes are 930k, 74k and 82k respectively. Vocabulary is limited to 10k words with all words outside this set mapped to a special token $< unk >$.

We start by looking at the distortion process. Table 1 shows some example sentences for this

dataset and corresponding output produced by the noisy channel. As we can see at 20% distortion the sentence is still coherent and meaning is still being conveyed. At 40% distortion, it is difficult even for human beings to discern what original sentence meant to say. This observation clearly indicates that a better language model should have considerably high contrastive perplexity for 40 % distortion as compared to 20%.

Table 1: Example sentence with 20% and 40% distortion

Original Sentence	Sentence with 20% distortion	Sentence with 40% distortion
no it was n't black monday	no it deeply black n't monday/	no it generating proceeds black monday
at the end of the day N million shares were traded	nights the meantime of the day N million shares fourth traded	centrust fundamentals away too encourage to secretary government for stop them now
things have gone too far for the government to stop them now	things have gone too far charles goldberg government to stop them openly	concentrate did a n't get even chance to do we slight the wanted to bear
but stocks kept falling	but ride kept falling	falling stocks kept but
they never considered themselves to be anything else	they never considered themselves to anything be else	promises never be themselves considered to anything else
businesses were borrowing at interest rates higher than their own earnings	businesses were borrowing their interest rates higher advise at own investigated	rates were borrowing intense interest businesses own than their equivalents ibm

Now, we look at the contrastive perplexities of various well known models at different distortion rates. The objective here is to verify the

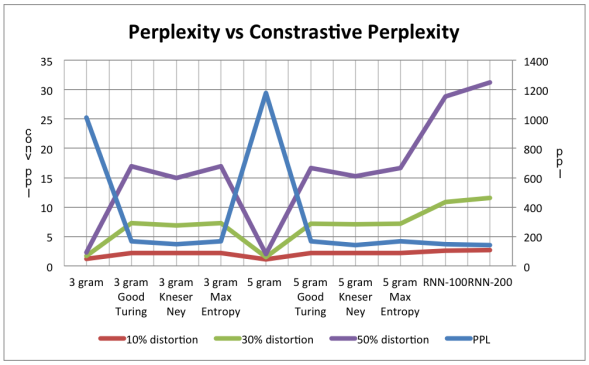

Figure 1: Contrastive Perplexity and Perplexity for all models, all distortion level

following hypothesis about contrastive perplexity. For a good language model contrastive perplexity should rise faster with the distortion level. This is akin to saying that a good language model should do a lot better job at differentiating the language generated by the model and distorted language as the distortion level increases. At the same time contrastive perplexity should be somehow correlated to perplexity across the models. This means that for same distortion level, range of models should be ranked similarly on two metrics, perplexity and contrastive perplexity. Table 2 shows the results for our experiments. The results were generated using open source language modeling SRILM toolkit(Stolcke and others, 2002) for n-gram models and RNNLM toolkit(Mikolov et al.,) for RNN based models. The results shown in Table 2 were averaged for 10 runs.

Table 2: Comparing n-gram models and RNNLM model perplexity for different level of distortion levels.

Model	Original	10% dist	30% dist	50% dist
3-gram	1009.90	1.18	1.68	2.34
3-gram GT	166.57	2.185	7.29	16.91
3-gram KN	148.28	2.183	6.91	14.93
5-gram	1177.85	1.112	1.49	2.02
5-gram GT	169.33	2.17	7.20	16.70
5-gram KN	141.46	2.22	7.08	15.26
RNN-100	148.78	2.57	10.87	28.84
RNN-200	141.31	2.649	11.52	31.21

Figure 1 shows the results discussed in Table 2. First thing to observe is the perplexity is inversely correlated to contrastive perplexity. This

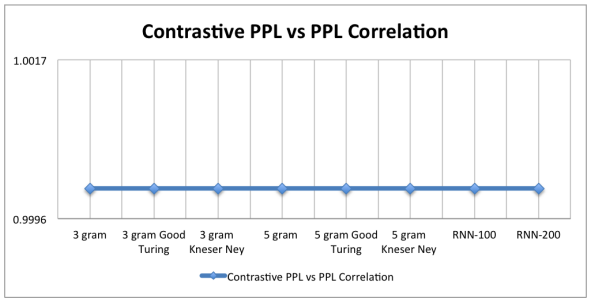

Figure 2: Correlation between Contrastive Perplexity vs Perplexity over models

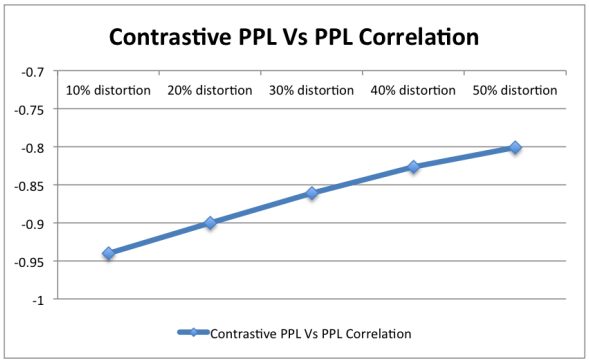

Figure 3: Correlation between Contrastive Perplexity and Perplexity over percentage distortion.

is in line with what we expect. A better language model would lead to a lower perplexity score as it would to a better job of lowering the cross entropy of model and original distribution but contrastive perplexity should increase as it should be able to do a better job at differentiating between original and distorted language. Another interesting observation here is about increase in contrastive perplexity with distortion across models. Contrastive perplexity increase rapidly for models with lower perplexity like RNN-100, RNN-200 and 5-gram KN. This is in line with the hypothesis that state of the art models should do a lot better job at discriminating between good and bad language as compared to the bad models for example 5-grams without any smoothing. We can see here that increase in contrastive perplexity is minimal for it.

Now, let's consider correlations between perplexity and contrastive perplexity across models and distortions. Figure 2 and Figure 3 plots the correlation between perplexity and contrastive perplexity at various models and distortion levels respectively. Figure 2 shows a very strong correlation between contrastive perplexity and perplexity across various models. From Figure 3, we have two observations. Firstly, as expected the correlation is negative which indicates the inverse relation between the two quantities across distortion level. Second, and more interesting observation is the the increasing slope of correlation across the distortion levels. This can be explained by the same rationale that at higher distortion levels contrastive perplexity rises considerably faster as compared to decrease of perplexity, hence the slope. This indicates in many way contrastive perplexity might be a better metric to evaluation language models than perplexity.

4 Conclusion

In this paper we proposed a new evaluation criteria which can be used to evaluate un-normalized language models. We showed that this new criteria has a very strong correlation with perplexity. The correlation across distortions indicate it might be a better metric than perplexity. Contrastive perplexity ranks models with better differentiating ability higher and WER ranks the model with lesser number of errors higher, we hypothesize there might be a higher correlation between these two as compared to WER's correlation with perplexity.

References

[Chen et al.1998] Stanley F Chen, Douglas Beeferman, and Roni Rosenfield. 1998. Evaluation metrics for language models.

[Clarkson et al.1999] Philip Clarkson, Tony Robinson, et al. 1999. Towards improved language model evaluation measures. In *EUROSPEECH*. Citeseer.

[Iyer et al.1997] Rukmini Iyer, Mari Ostendorf, and Marie Meteer. 1997. Analyzing and predicting language model improvements. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 254–261. IEEE.

[Mikolov et al.] Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukar Burget, and Jan Cernocky. Rnnlm-recurrent neural network language modeling toolkit.

[Smith and Eisner2005] Noah A Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 354–362. Association for Computational Linguistics.

[Stolcke and others2002] Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.