Comparing methods for preventing Posterior Collapse in Deep Generative Models for Text

Stefan F. Schouten
University of Amsterdam
stefan.schouten@student.uva.nl

Guido Ansem
University of Amsterdam
guidoansem@gmail.com

Abstract

This paper examines a phenomenon called Posterior Collapse in sentence variational auto encoders (S-VAE). Posterior Collapse, also known as the strong decoder problem, occurs when the KL divergence in the ELBO of the S-VAE reaches zero. This causes the Encoder to recreate the prior, preventing it from effectively conditioning the Decoder. In this paper three different methods of preventing posterior collapse are tested and compared. These methods include, Word Dropout, freebits and μ-forcing. We conclude that combining Word Dropout and μ-Forcing is an effective way to obtain a conditionable Decoder network.

1 Introduction

In recent years there has been a big increase in neural models for text generation. Different neural models, for example GPT2 (Radford et al., 2019) are able to generate text which is almost indistinguishable from text written by humans. One of the limitations of these models is that they are not able to include external variables in the generation of the text, for example a certain style or subject. Deep generative models (DGM) can provide a solution for this limitation. DGM's such as variational auto-encoders (VAE) (Kingma and Welling, 2013) have been widely used for different NLP tasks including language modeling (Bowman et al., 2015), machine translation (Zhang et al., 2016) and controllable text generation (Fang et al., 2019). These models are able to capture holistic properties of input such as style and subject, which guide the generation of diverse relevant sentences.

The quality and the number of latent factors learned is influenced by a phenomenon called posterior collapse. Posterior collapse occurs when the variational distribution starts to closely match the prior for a subset of latent variables and the model learns to ignore these latent variable (Lucas et al., 2019) (Bowman et al., 2015) (Chen et al., 2016). This posterior collapse is caused by the KL-Divergence term in the ELBO objective. We contribute to this problem by providing a review and comparison of different methods that have been proposed to counter this problem. The goal of the paper is to give more insight in the effect of the different methods. The research question for this paper can be formulated as *Which of the methods tested in this paper or a combination there off performs best with respect to the prevention of posterior collapse in Deep Generative Models.*

2 Related Work

Multiple methods have been developed over the years in an attempt to reduce posterior collapse. In this paper three different methods will be reviewed and compared.

The first of these methods is Word Dropout (Bowman et al., 2015). This method tries to reduce the posterior collapse by weakening the decoder. During learning the decoder tries to predict each word based on the previous words, by removing some of this prior information during learning the decoder can easily be weakened. Bowman et al., (2015) did this by randomly replacing some of the previous words with a token that represents unknown words (UNK). This reduction in prior information should force the model to rely on the latent variable for its predictions and therefore reducing posterior collapse.

The second method examined in this paper is usually referred to as *freebits* (Kingma et al., 2016). This method uses a constant objective function that is constant throughout training. The latent dimension is divided into K groups in which the parameters are shared. The objective (shown in Equa-

tion 1) is then used to ensure that no less than λ nats of information per mini-batch are used. By keeping some space for the KL divergence on every dimension of the latent variables, this places a lower bound on the value of the divergence thereby making it impossible to collapse.

$$\mathcal{L}_\lambda = \mathbb{E}_{x\sim M}[\mathbb{E}_{q(z|x)}[\log p(x|z)]] - \sum_{j=1}^{K} max(\lambda, \mathbb{E}_{x\sim M}[D_{KL}(q(\mathbf{z_j}|\mathbf{x})||p(\mathbf{z_j}))]) \quad (1)$$

The last method was proposed by (Liu et al., 2019) and is called *μ-forcing*. The idea behind this method is that when the model collapses the approximation of every datapoint collapses to the same simple distribution $\mathcal{N}(0, 1)$ (the prior). This method punishes the lack of variance (in each batch) that would be caused by the collapse. It does so by adding an additional loss term that is inversely proportional to the sample variance, up to a threshold β (see Equation 2). The idea is that if the model is encouraged to have variance within the batches in this way, that together with the pressure from the Negative Log-Likelihood, it will not just add arbitrary variation. Instead it is assumed and hoped that the model will find a way to add variation such that the latent variables contain useful information.

$$\mathcal{L}_\mu = \max\left\{0, \beta - \frac{1}{2}q_\mu\right\} \quad (2)$$
$$with \quad q_\mu = \frac{1}{N}\sum_{n=1}^{N}(\mu^{(n)} - \bar{\mu})^\top(\mu^{(n)} - \bar{\mu})$$

3 Method

3.1 Dataset

The dataset for this work is based on the Penn Treebank (Marcus et al., 1994), a large annotated corpus of English, which is commonly used for language-modeling tasks. The corpus is annotated with syntax trees and POS-tags, neither of which were used in this work. The words in the dataset come already tokenized, we convert all character to lowercase. For our dictionary we only take words that occurred at least twice in the corpus. Finally this results in a dictionary of a little over 22.000 tokens. We use 300-dimensional fastText pretrained word embeddings (Mikolov et al., 2018).

3.2 Baseline Model (RNNLM)

The baseline model we use is a standard RNN language model (RNNLM). This model used a autoregressive definition of the likelihood $P(X_i|x_{<i}; \theta)$. This model takes the words from the previous time steps as input to predict the following word while keeping track of the hidden states that incorporate useful information from all time steps. Specifically we used a (uni-directional) gated recurrent unit (GRU) with one hidden layer and 500-dimensional hidden states.

3.3 Sentence VAE

The VAE we use to study posterior collapse is based on previous research by (Bowman et al., 2016), the core structure of this model is shown in figure 1. In this model the the decoder is a language model identical to the baseline RNNLM, but is now conditioned on a latent variable. The latent variable is drawn from a Gaussian distribution parameterized by the output of the Encoder network. The Encoder is a bi-directional GRU with again one hidden layer and 500-dimensional hidden states.

3.4 Learning

We optimize our networks with Adam, we use a learning rate of 0.001 and a learning rate decay of 0.95 applied each epoch. Each pass of the network encodes one sentence from the dataset, regardless of its length, which the decoder then attempts to reproduce. We do not fine-tune our word embeddings.

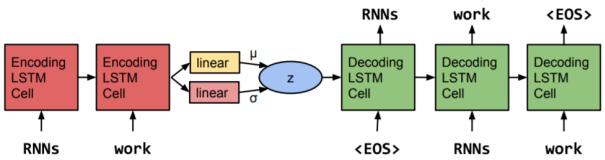

Figure 1: Basic structure of the Sentence-VAE

4 Experiments and Results

4.1 Quantitative Evaluation

We compare the methods for preventing posterior collapse in the Sentence-VAE against our RNNLM baseline by reporting their Perplexity and Negative Log-Likelihood. We also report the KL-divergence between the Sentence-VAE's variational posterior and prior. These metrics were calculated on the test set, using estimations based on 10 samples for each data point; and are reported in Table 1 for multiple combinations of hyperparameters.

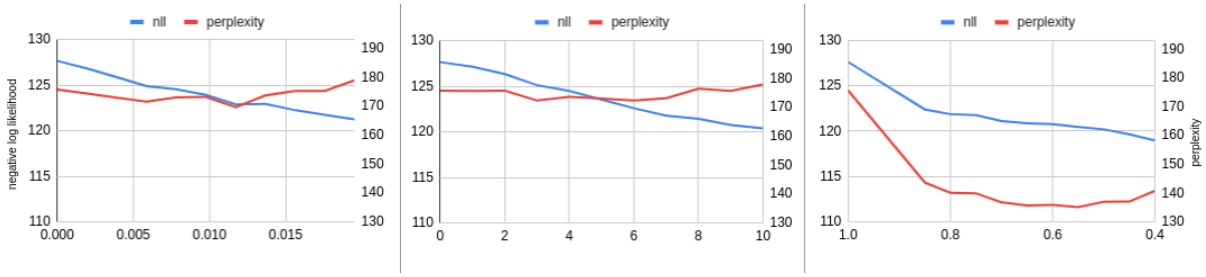

Figure 2: From left to right we see the effect of: *freebits*, μ-Forcing, and Word Dropout; on the negative log-likelihood and perplexity. The maximum of the axis is chosen equal to the performance of the RNNLM baseline (nll=129.8, perplexity=193.3). The leftmost datapoint in each plot is the S-VAE without any of the methods for preventing posterior collapse.

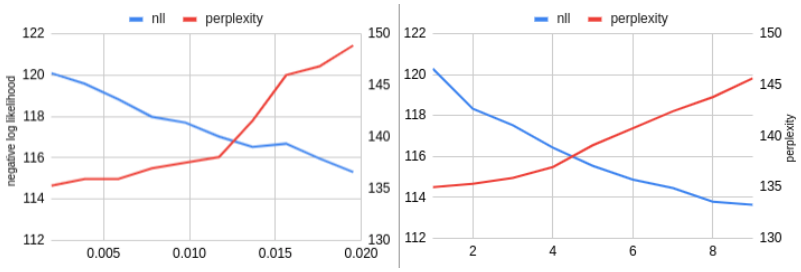

Figure 3: The effect of *freebits* (left) and μ-Forcing (right) on the negative log-likelihood and perplexity, when used with Word Dropout (p=0.55).

The combinations of hyperparameters that are reported, were selected based on a hyperparameter search. We first searched for the best performing singular method for preventing posterior collapse. The results of this first step can be seen in Figure 2. The following hyperparameters were investigated: $\lambda \in \{\frac{1}{512}, \frac{2}{512}, \ldots, \frac{10}{512}\}$, $\beta \in \{1, 2, \ldots, 10\}$, $p \in \{0.4, 0.45, \ldots, 0.85\}$. Where λ is the threshold for *freebits*, β is the threshold for μ-Forcing, and p is the keep probability of Word Dropout.

In Figure 2 we can see that Word Dropout is most capable of reducing Perplexity. By removing a large part of the input, Word Dropout pushes the model to encode information about the input in the latent vector. By the much lower Perplexity, we know that this makes a considerable difference for the model's ability to model the data. All methods seem roughly equally capable of reducing the Negative Log-Likelihood.

Given Word Dropout's ability to effectively reduce the perplexity, we only consider combining it with one of the other methods, and not the other methods with each other. We think this way we have the best chance of producing a model with equal or better Perplexity and lower Negative Log-Likelihood. To investigate how the addition of *freebits* or μ-Forcing to Word Dropout affects both metrics, we again search through the same hyperparameters for λ and β, but now while keeping p fixed at 0.55. The results for this can be seen in Figure 3.

We can see that there is not a large difference between the two methods, but also that μ-Forcing is able to reduce the Negative Log-Likelihood slightly more while keeping the Perplexity slightly higher. It seems that μ-Forcing's penalty on low-variance is somewhat better at making the model use the latent vector while also still effectively modelling the data. This is in line with our expectations; since μ-Forcing is essentially introducing an inductive bias that encourages the model to produce what we, as designers, consider a useful property. *freebits* on the other hand leaves the model free to decide in what way to 'spend' it's penalty-free use of the latent space. In this case it seems that introducing this bias works in our advantage.

Because we know we can reduce the Negative Log-Likelihood and Perplexity using Word Dropout and reduce the Negative Log-Likelihood further most effectively with μ-Forcing, we decide to test and compare variants of the Sentence-VAE with: no methods for preventing posterior collapse (as an additional baseline) that we call 'vae-vanilla'; the combination of methods that achieved lowest Perplexity ($p = 0.55, \mu = 1$) which we call 'vae-min-

Name	Model	Word Dropout	μ-Forcing	Test NLL (KL)	Test PPL
baseline	RNNLM	-	-	131.2 (-)	206.2
vae-vanilla	S-VAE	1	0	129.7 (0.5)	190.0
vae-min-ppl	S-VAE	0.55	1	120.8 (2.8)	145.1
vae-min-nll	S-VAE	0.55	25	107.8 (28.5)	197.4

Table 1: Our main results, the Negative Log-Likelihood, Perplexity and KL-Divergence of the models on the test set when trained with the given relevant hyperparameters.

ppl'; and finally, a combination of methods that obtains the lowest Negative Log-Likelihood while achieving Perplexity comparable to the baselines ($p = 0.55, \mu = 25$) that we call 'vae-min-nll'.

4.2 Qualitative Evaluation

For a qualitative evaluation we perform interpolation between sentences. We do so by selecting two sentences by hand (see Appendix A), mostly arbitrarily but making sure to select a pair of medium length sentences, and a pair with one short and one long sentence. We use the mean from the Encoder's output as the corresponding latent variables $(\vec{z}_a, \vec{z}_b)$. These latent variables are interpolated between, producing a batch $\mathbb{Z}$ where $\mathbb{Z}_i = \vec{z}_a + \frac{i}{n} \cdot (\vec{z}_b - \vec{z}_a)$. We condition the Decoder once for each of these and have it produce a sentence through greedy sampling. The results of this interpolation are shown in Appendices B.1, C.1, and D.1. We also chose three sentences for which we compare the output of the Decoder when conditioned on the mean, and on one of three samples from the posterior. These can be found in Appendices B.2, C.2, D.2 and B.3, C.3, D.3 respectively.

They first thing these experiments show is that the vae-vanilla does not use latent space, indicating that the model has collapsed, as expected. This becomes apparent from both the interpolations, which always produce the same sentence, and the reconstructions, which have no apparent similarity to the encoded sentence. Things only improve slightly when we move on the vae-min-ppl: the interpolation still consists mostly of the same sentences over and over again; but the reconstructions start to show some signs indicating that the model is using the latent space. Each of the reconstructions of sentence **(4)** starts with "mr.". But otherwise there is not much to connect the generated sentences back to what was encoded. So even though this model had the highest perplexity, and a non-zero KL-Divergence, this was not yet enough. When we look at the vae-min-nll however we *do* start to see the sentences change during the interpolation under the influence of the latent variable. For example, we see evidence of the length of the sentence being encoded, sentences are gradually increased in length during interpolation. We also see that the subject and main verb are often the same, and negations are often preserved.

5 Conclusion

This paper examines three different methods of preventing posterior collapse in S-VEA's, namely Word Dropout, *freebits* and μ-forcing. These different methods were examined and compared to each other as well as to a baseline. From the quantitative evaluation we conclude that Word Dropout is efficient in reducing the Perplexity. We also conclude that μ-Forcing is more efficient than Free Bits at reducing the Negative Log-Likelihood. And that when Word Dropout and μ-forcing are combined, the perplexity won by the Word Dropout can be effectively 'spent' by μ-forcing to reduce the Negative Log-Likelihood. We qualitatively compared different combinations of Word Dropout and μ-forcing against a Sentence-VAE with a collapsed posterior. We confirm that a low Negative Log-Likelihood is key in determining how well conditioning of the decoder works. By combining Word Dropout and μ-forcing we obtain a model that has a Perplexity similar to the baseline, but with much reduced Negative Log-Likelihood.

6 Future Work

Other methods for preventing posterior collapse than examined in this paper have also been proposed, for example (Razavi et al., 2019) and (Pelsmaeker and Aziz, 2019). These methods or combinations of these methods with those tested in this paper might even further decrease the Negative Log-Likelihood and lead to even better results of the model overall. Future research to these methods is needed to test this hypothesis.

References

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*.

Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. 2019. Implicit deep latent variable models for text generation. *arXiv preprint arXiv:1908.11527*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751.

Dayiheng Liu, Yang Xue, Feng He, Yuanyuan Chen, and Jiancheng Lv. 2019. μ-forcing: Training variational recurrent autoencoders for text generation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–17.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. 2019. Understanding posterior collapse in generative latent variable models.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, page 114–119, USA. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tom Pelsmaeker and Wilker Aziz. 2019. Effective estimation of deep generative language models. *arXiv preprint arXiv:1904.08194*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. 2019. Preventing posterior collapse with delta-vaes. *arXiv preprint arXiv:1901.03416*.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. *arXiv preprint arXiv:1605.07869*.

A Sentences used for the Qualitative Analysis

The sentences that were used for the evaluation of the model are displayed below. The first for sentences show were used for the interpolation and the last three sentences for the mean and sample.

- Interpolation pair 1:
  - **(1a)** *but so far the company has n't complied with that request , the spokesman said .*
  - **(1b)** *others wonder how many more of these shocks the small investor can stand .*
- Interpolation pair 2:
  - **(2a)** *iowa is making a comeback .*
  - **(2b)** *food and drug administration spokesman jeff nesbit said the agency has turned over evidence in a criminal investigation concerning vitarine pharmaceuticals inc. to the u.s. attorney 's office in baltimore .*
- Sentences we attempt to reconstruct using the mean, and samples:
  - **(3)** *the company asked for a 15-day extension sept. 30 , when the financial reports were due .*
  - **(4)** *mr. spielvogel said he would n't launch a hostile bid .*
  - **(5)** *officials at drexel said they had n't seen the suit and thus could n't comment .*

The Interpolations, and Reconstructions from the mean and samples, are all given in the same order as listed here. They are given separately for each of the model-variants based on the Sentence-VAE given in Table 1. The sentences were all retrieved from the test set, and have the following indices respectively: 1981 **(1a)**, 2119 **(1b)**, 1417 **(2a)**, 1977 **(2b)**, 672 **(3)**, 625 **(4)**, 555 **(5)**.

B Qualitative Analysis Outputs: vae-vanilla

B.1 Interpolation (1) & (2)

```
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
------
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
the company said it expects to report a loss for the third quarter . <bos>
```

B.2 Reconstructed from mean

```
the company said it expects to report a loss for the third quarter . <bos>
----
the company said it expects to report a loss for the third quarter . <bos>
----
the company said it expects to post a third-quarter loss of about $ 600,000 million
    , or about 8,000 cents a share , in the quarter . <bos>
```

B.3 Reconstructed from samples

```
the government 's assertions that the government is n't the only reason to deny the
    existence of the drug . <bos>

`` i think it 's a little disingenuous , '' says mr. <unk> , an analyst with raymond
     james & associates . <bos>
```

on the other hand , the agriculture department said that the government 's <unk> refusal to face the opposition of the government 's demands to keep the government 's so-called and <unk> . <bos>

----

the company said it expects to report a loss for the fourth quarter . <bos>

the company said it expects to report a loss for the fourth quarter . <bos>

the company said it expects to report a loss for the third quarter . <bos>

----

the company said it expects to report a loss for the third quarter . <bos>

`` we 're not going to be stampeded in the fourth quarter , '' said <unk> <unk> , an analyst with painewebber inc . <bos>

mr. the company said the sale of its <unk> % stake in the company , will be able to reduce debt . <bos>

C Qualitative Analysis Outputs: vae-min-ppl

C.1 Interpolation (1) & (2)

but the company said it will take a $ 50.9 million charge of the company 's stock , which will be sold by the company 's <bos>
but the company said it will take a $ 50.9 million charge of the company 's stock , which will be sold by the company 's <bos>
but the company said it will take a $ 50.9 million charge of the company 's stock , which will be sold by the company 's <bos>
but the company said it will take a $ 50.9 million charge of the company 's stock , which will be sold by the company 's <bos>
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>
------
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>
the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>

C.2 Reconstructed from mean

the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>

mr. <unk> , who is retiring as president and chief operating officer of this maker of specialty and <unk> products . <bos>

the company said it will take a $ 50.9 million charge of the sale of its leasing unit to paris-based the company 's <bos>

C.3 Reconstructed from samples

```
the new york stock exchange composite trading yesterday , closed at $ <unk> , down
    12.5 cents . <bos>

the company said it will take a $ 50.9 million charge of the sale of its leasing
    unit to paris-based the company 's <bos>

the company said it will take a $ 50.9 million charge of the sale of its leasing
    unit to paris-based the company 's <bos>

----

mr. , a maker of optical devices , said it will sell versions of the machines , and
    the company 's the <unk> of the plant . <bos>

mr. <unk> , who is retiring as president and chief operating officer of this maker
    of specialty and <unk> products . <bos>

the company said it will sell its non-food cartridge for the first time of yamatake-
    honeywell , and the company 's n't be able to make it . <bos>

----

but the company said it will take a $ 50.9 million charge of the company 's stock ,
    which will be sold by the company 's <bos>

the <unk> , which is puttable , and <unk> , will be sold by the end of the year . <
    bos>

the company said it will take a $ 50.9 million charge of the sale of the company 's
    stock and warrants . <bos>
```

D Qualitative Analysis Outputs: vae-min-nll

D.1 Interpolation (1) & (2)

```
but even though the company has n't yet been family-run , and the company said . <
    bos>
but even though the company has n't yet to be determined by the company . <bos>
but even though the company has n't yet to be profitable . <bos>
but even though the company 's n't to be profitable in the <unk> . <bos>
some gripes that the stock market is too much too . <bos>
they are telling the truth about how much of the market . <bos>
you can ignore the risk of the risk of risk . <bos>
------
<unk> is a closed-end . <bos>
<unk> is a <unk> . <bos>
<unk> is a <unk> of the company 's <bos>
<unk> , the u.s. said the project has been working in a <unk> . <bos>
<unk> and the administration has been working with a <unk> of the new york city , n.
    j . <bos>
officials in washington and mr. bush has been diagnosed the possibility of renewing
    military aid in the u.s. of the u.s. defense initiative . <bos>
officials in defense and defense ministry said that the u.s. has been challenged by
    the new york state of new york , and the u.s. of the u.s. intelligence agency .
    <bos>
```

D.2 Reconstructed from mean

```
the company was a <unk> % decline in the company 's but declined to comment on the
    debt-rating . <bos>

mr. nadeau said he would n't comment on the board . <bos>

officials at drexel said they were n't aware that the contract was n't . <bos>
```

D.3 Reconstructed from samples

```
the company reported a $ <unk> million charge , which was largely financed by the
    company 's <bos>
```

```
the company noted that the results were bolstered by the company 's results were n't
     by a provisional . <bos>

the company received a $ 300 million for damages , but the government did n't . <bos
    >

----

mr. he said he would n't a bid for the company . <bos>

mr. guber said he would n't a full-fledged . <bos>

mr. bush would n't have a veto threat . <bos>

----

officials at the company said that they were n't to be reached for comment . <bos>

officials at the company , which had been notified about the request was n't . <bos>

officials at officials said they were n't aware that the company had n't comment on
    the suit . <bos>
```

E Notes on the code

While putting the finishing touches on our paper, we noticed a small mistake in the source code. Instead of saving the perplexity for the best found negative log-likelihood, we saved the best perplexity separately. Theoretically it is possible that some models achieved a minimum perplexity and minimum log-likelihood during a different epochs. Although when we checked the logs of our hyperparameter search, we did not find this to be the case in practice.