Analysis and Optimization of Text Generation Models using Brown Corpus and WordNet

Pranav Verma
2015EE30528
Varun Srivastava
2015EE10835

Facilitator:
Prof. Jayadeva
Dept. of Electrical Engineering, IIT Delhi

Department Head:
Prof. S. D. Joshi
Department of Electrical Engineering, IIT Delhi

Objective

- To create a fused dataset by incorporating the semantic information from WordNet with word distribution data from the Brown Corpus.
- To extend the current RNN architectures to incorporate semantic information to improve their performance on language modelling tasks.

WordNet and Brown Corpus

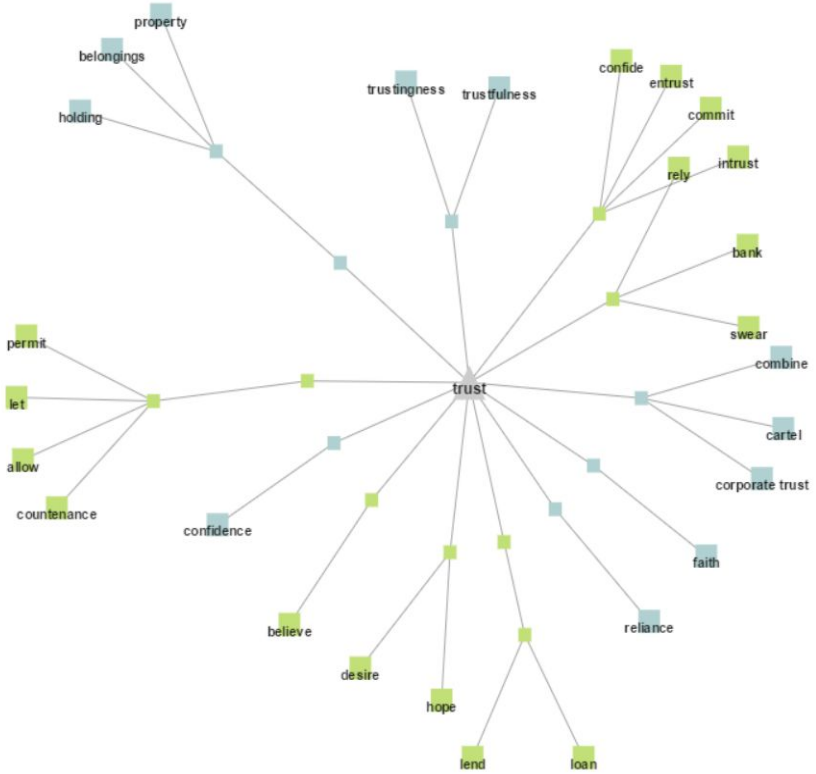
WordNet

- Lexical database
- Groups words together based on their meanings
- Labels the relations between words

Brown Corpus

- Contains about a million words drawn from various English sources.
- Provides syntactical data for words
- Standard Database often used in NLP tasks

Our dataset incorporates the semantic and syntactic data from both the datasets



Language Modeling

A language model is a probability distribution over a sequence of words.

Let $w_1, \ldots, w_m$ be a sequence of words, then a language model assigns a probability P given by

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1})$$

Perplexity

Perplexity is the measure of how likely a given language model will predict a given sequence of words.

The perplexity per word of a discrete probability distribution *p* is defined as

$$2^{-\frac{1}{N}\sum_{i=1}^{N} log(p_i)}$$

LSTM

The previous hidden state : h_{t-1}

The current input : x_t

The previous cell state : C_{t-1}

The forget gate output : $f_t = \sigma(W_f.[h_{t-1}x_t] + b_f)$

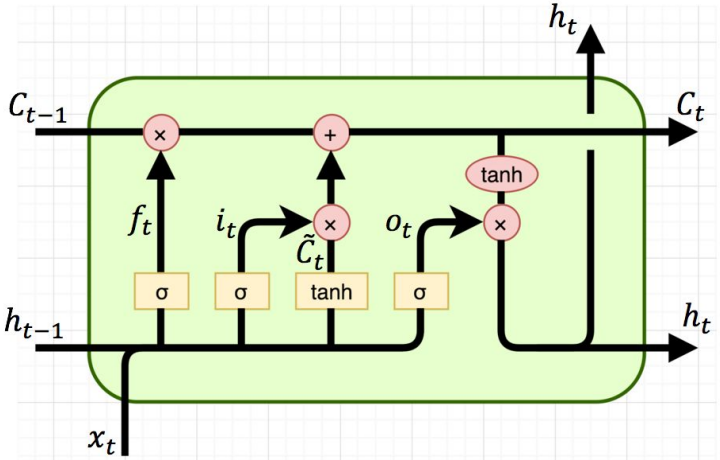
The input gate output : $i_t = \sigma(W_i.[h_{t-1}x_t] + b_i)$

The candidate state : $\tilde{C}_t = tanh(W_C.[h_{t-1}x_t] + b_C)$

The output gate output : $o_t = \sigma(W_o.[h_{t-1}x_t] + b_o)$

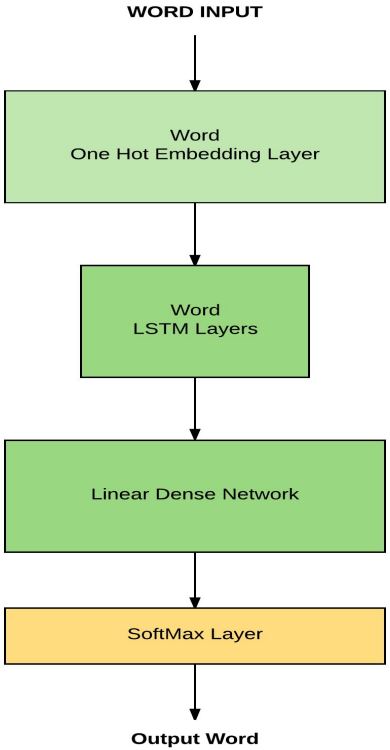
The next output : $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

The next hidden state : $h_t = o_t * tanh(C_t)$

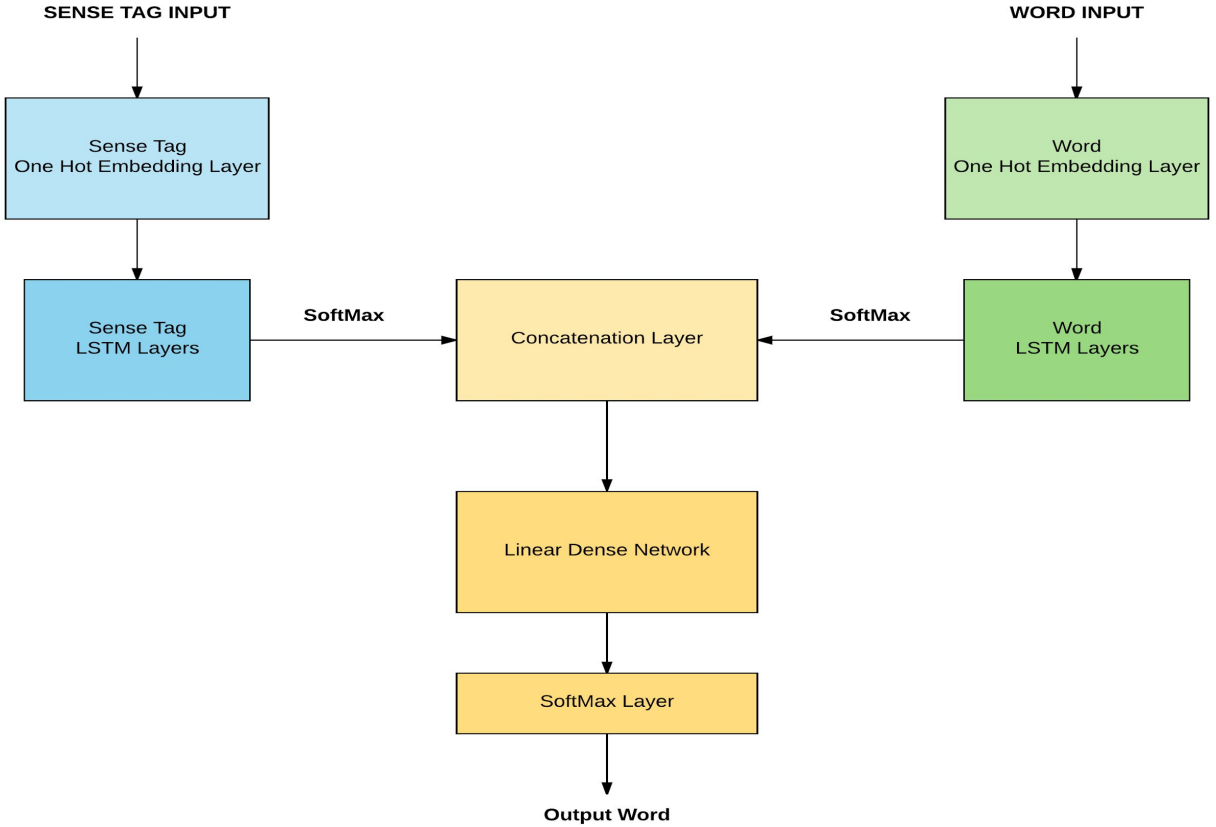


(a) Long Short-Term Memory

Base Model


WORD INPUT
Word
One Hot Embedding Layer
Word
LSTM Layers
Linear Dense Network
SoftMax Layer
Output Word

Our Augmented model



Testing Methodology

The best fit models based on a training, validation and test split of 70:20:10 were evaluated on the test set.

A comparison of these models was performed based on the following parameters:

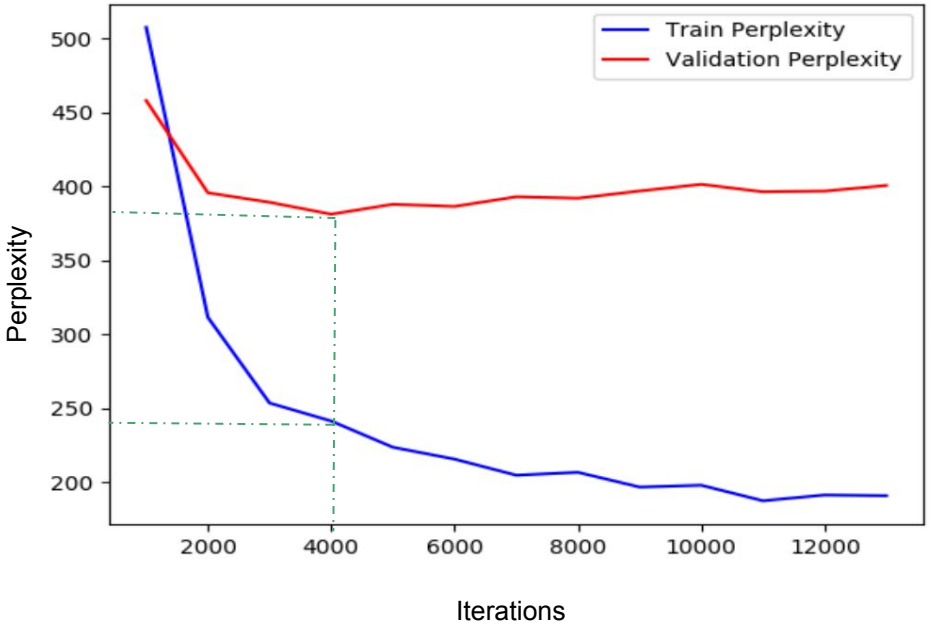
1. Perplexity of the model on the Test Set (Quantitative Analysis)
2. A qualitative analysis of the sentences produced by the model

Model Comparison

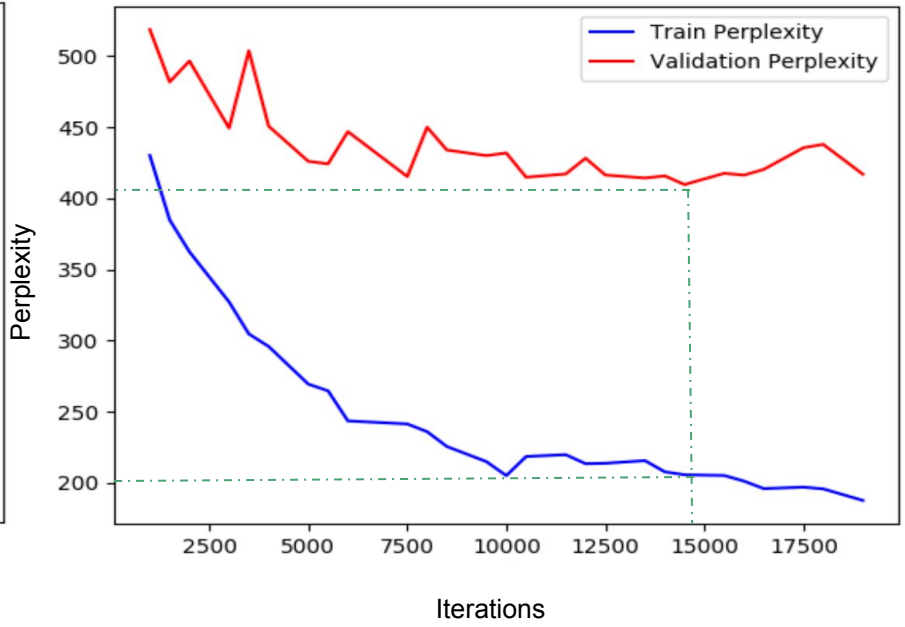
	Train Perplexity	Validation Perplexity	Test Perplexity
Base Model	241.34	***381.08***	737.91
Augmented Model	***205.54***	409.67	***383.17***

Model Comparison

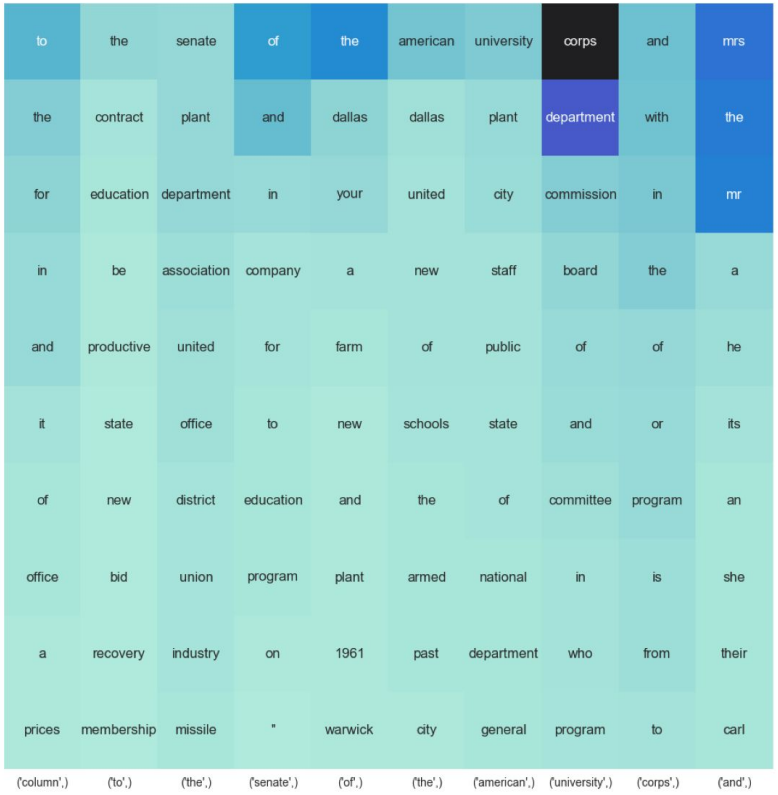
Base Model

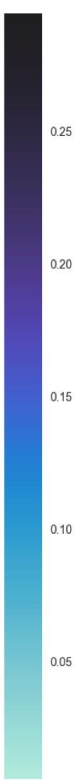


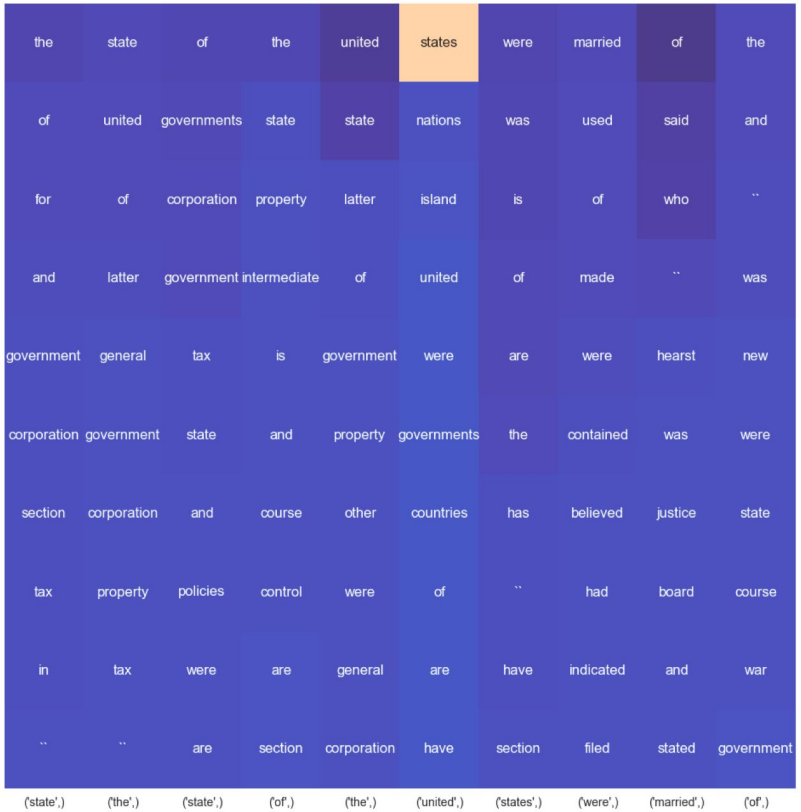
Augmented Model



Heat Map for Text Generated by Non Teacher Forcing



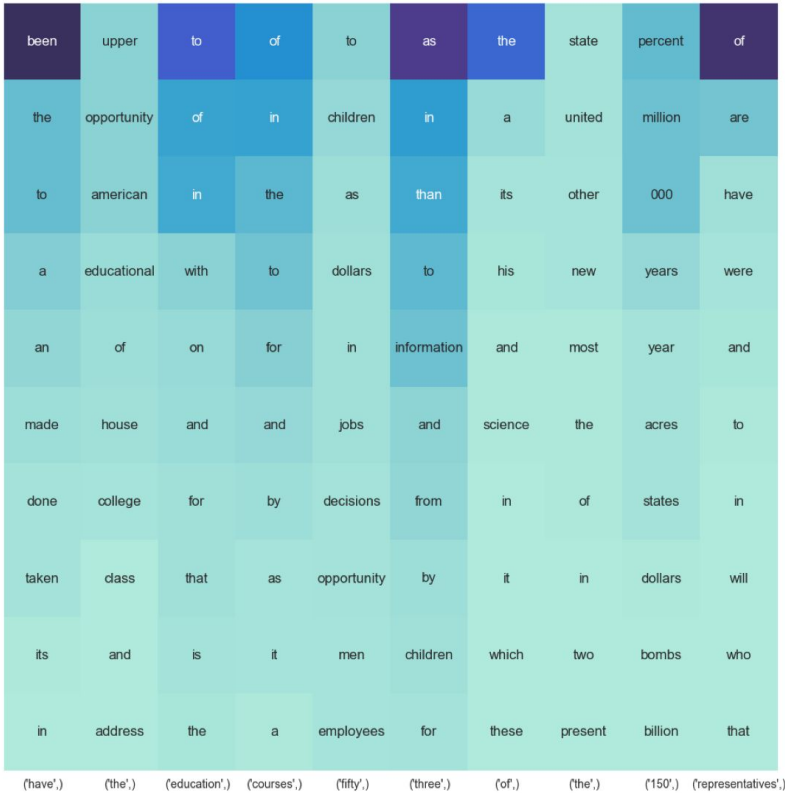


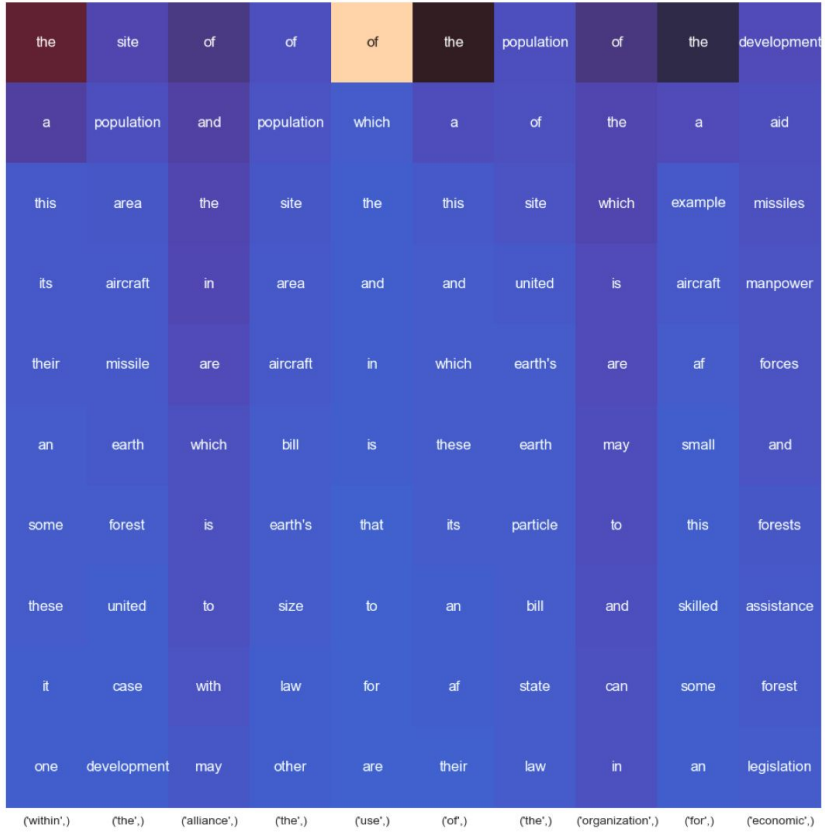


Base Model

Augmented Model

Heat Map of Text Generated by Teacher Forcing





Base Model

Augmented Model

Conclusion

1. A ***novel approach*** for incorporating syntactic and semantic knowledge on a LSTM based language model
2. Demonstrated our hypothesis that using sense tags and semantic information with word corpora improves text generating models.
3. Achieved lower perplexity on the Brown Corpus compared to the current language models.
4. Improvement in the grammatical quality of the text produced.
5. Future scope of the model includes
    a. Improved Sentence Quality in Image Captioning tasks
    b. More accurate autocorrect in applications involving text prediction (*eg: Mobile Keyboards)*

References

1. I. Sutskever, J. Martens and G. E. Hinton., "Generating text with recurrent neural networks," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011.
2. S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," *Advances in neural information processing systems,* 1997.
3. C. E. Shannon, "Prediction and Entropy of Printed English," vol. 30, no. 1, 1951.
4. A. Graves, "Generating sequences with recurrent neural networks," vol. arXiv:1308.0850, 2013.
5. G. Hinton and Tieleman, "Lecture 6.5 - RMSPROP-Neural Networks for Machine Learning - Coursera," 2012.
6. Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research,* pp. 1137-1155.

Thank you