Erratum to "Learning Semantic Sentence Embeddings using Pair-wise Discriminator"

Badri N. Patro* **Vinod K. Kurmi*** **Sandeep Kumar*** **Vinay P. Namboodiri**
Indian Institute of Technology, Kanpur
{badri,vinodkk,sandepkr,vinaypn}@iitk.ac.in

Abstract

An erratum is being provided in this paper to address an error in the comparison in table 4 for our paper (Patro et al., 2018). We also update table 2 of the paper as we are not sure about the baseline score from a previous work and we provide a clarification in this regard.

1 Introduction

We had made the code for our work publicly available on Github [1]. It was pointed out that the evaluation for table 4 had a mistake as we were comparing Root results with all phrase results for the sentiment analysis task. It was also raised that the BLEU score referred to in the paraphrase question generation task may not be comparing the right BLEU score. Both these points are being addressed in this erratum through updated tables.

2 Paraphrase Question Generation Task

Dataset	Model	BLEU1	METEOR	TER
50K	Unsupervised VAE (Gupta et al., 2017)	–	12.2	83.7
	VAE-S (Gupta et al., 2017)	–	17.4	69.4
	VAE-SVG (Gupta et al., 2017)	–	21.3	63.1
	VAE-SVG-eq (Gupta et al., 2017)	–	**21.4**	61.9
	EDD-G (**Ours**)	40.7	19.7	51.2
	EDD-LG(**Ours**)	40.9	19.8	51.0
	EDD-LG(shared)(**Ours**)	**41.1**	20.1	**50.8**
100K	Unsupervised (Gupta et al., 2017)	–	14.3	79.9
	VAE-S (Gupta et al., 2017)	–	21.6	67.1
	VAE-SVG (Gupta et al., 2017)	–	24.6	55.7
	VAE-SVG-eq (Gupta et al., 2017)	–	**24.7**	55.0
	EDD-G (**Ours**)	42.1	20.4	49.9
	EDD-LG(**Ours**)	44.2	22.1	48.3
	EDD-LG(shared)(**Ours**)	**45.7**	23.1	**47.5**

Table 1: The table provides an analysis of baselines and state-of-the-Art methods for paraphrase generation on Quora dataset. This table provides a modified comparison for the table 2 of our previous work (Patro et al., 2018). In that table we had indicated a BLEU1 score for the work by (Gupta et al., 2017). However, we are not sure which BLEU score is indicated in the paper. Hence, the revised table omits the BLEU score for their work. Note that all BLEU scores and other measures are available in table 1 of our paper (Patro et al., 2018).

[1] Source Code: https://github.com/badripatro/PQG

Model	Root (Fine-Grained)	All (Fine-Grained)
Naive Bayes (Socher et al., 2013)	59.0	32.8
SVMs (Socher et al., 2013)	59.3	35.7
Bigram Naive Bayes (Socher et al., 2013)	58.1	29.0
Word Vector Averaging (Socher et al., 2013)	67.3	26.7
Recursive Neural Network (Socher et al., 2013)	56.8	21.0
Matrix Vector-RNN (Socher et al., 2013)	55.6	21.3
Recursive Neural Tensor Network (Socher et al., 2013)	54.3	19.3
Paragraph Vector (Le and Mikolov, 2014)	51.3	–
EDD-LG(Random) (Ours)	61.3	40.0
EDD-LG(Shared) (Ours)	58.7	37.5

Table 2: Performance of our method compared to other approaches on the Stanford Sentiment Treebank Dataset. The error rates of other methods are reported in (Le and Mikolov, 2014)

3 Sentiment Analysis with Stanford Sentiment Treebank (SST) Dataset

3.1 Tasks and Baselines

In this erratum, we consider both root level and all phrase-based comparison with respect to (Socher et al., 2013). The earlier comparison was based on work by (Le and Mikolov, 2014) and the distinction between root and all-phrase was not clear from that work. However, based on the comments received on our Github repository, we are able to provide an updated comparison in this erratum. We have also updated our repository with the updated comparison code.

3.2 Results

We report the error rates of different methods in table 2. We compare our method with various other methods and a relative baseline with both root and all-phrase-based comparisons. The encoder-LSTM with a random initialisation for the "encoder-LSTM" model is treated as a comparative baseline model (EDD-LG(Random)). When we initialise the "encoder-LSTM" module with pre-trained "EDD-LG(shared) encoder-LSTM" weights, we treat it as our proposed module. We have also uploaded our models to the online competition on Rotten Tomatoes dataset [2] and obtained an accuracy of 62.606% on their test-set of 66K phrases. The table shows that with respect to a baseline random initialization there is an improvement of 2.5%. However, it does not improve over other models, particularly the model by Le and Mikolov has the best performanace of 51.3% and is considerably better than our work.

We thank the PCs for providing the chance to add an erratum for our work. Lastly, it was due to our belief in open research that this error could be spotted and we thank Aykut Firat who took the trouble for identifying this error and helping us in improving our work.

References

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2017. A deep generative framework for paraphrase generation. *arXiv preprint arXiv:1709.05074*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Badri Narayana Patro, Vinod Kumar Kurmi, Sandeep Kumar, and Vinay Namboodiri. 2018. Learning semantic sentence embeddings using sequential pair-wise discriminator. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2715–2729.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

[2]website: `www.kaggle.com/c/sentiment-analysis-on-movie-reviews`