# Neural Paraphrase Generation using Transfer Learning

**Florin Brad**
Bitdefender, Romania
fbrad@bitdefender.com

**Traian Rebedea**
University Politehnica of Bucharest
traian.rebedea@cs.pub.ro

## Abstract

Progress in statistical paraphrase generation has been hindered for a long time by the lack of large monolingual parallel corpora. In this paper, we adapt the neural machine translation approach to paraphrase generation and perform transfer learning from the closely related task of entailment generation. We evaluate the model on the Microsoft Research Paraphrase (MSRP) corpus and show that the model is able to generate sentences that capture part of the original meaning, but fails to pick up on important words or to show large lexical variation.

## 1 Introduction

Paraphrase generation is the problem of restating a given sentence such that its overall meaning is preserved. This can be seen as a task useful in and of itself or it can serve in proxy applications such as sentence summarization, sentence simplification, question expansion in question answering or rephrasing utterances generated by a conversational agent.

Paraphrase generation has been previously treated as a monolingual machine translation (MT) problem (Quirk et al., 2004; Finch et al., 2004). Lately, Neural Machine Translation (NMT) has revived interest in statistical machine translation through the use of sequence-to-sequence (SEQ2SEQ) models that learn to maximize the probability of a sentence in a target language, given a sentence in a source language (Cho et al., 2014; Sutskever et al., 2014). The SEQ2SEQ model is composed of an encoder that recurrently consumes the words in the source sentence and a decoder that sequentially predicts words in the target sentence, conditioned on the encoder's last hidden state and the previously translated words. This model was later improved by using an attention mechanism (Bahdanau et al., 2014) that allowed the decoder to focus on the relevant words from the source sentence.

NMT can then be used for paraphrase generation by maximizing the probability $P(Y|Y')$, where $(Y, Y')$ is a pair of paraphrases. While parallel corpora are abundantly available for machine translation, paraphrase corpora featuring pairs of complex sentences are prohibitively small for training large models. We propose to overcome this aspect by performing transfer learning from a similar task - entailment generation, which is facilitated by the large number of entailment pairs featured in the Stanford Natural Language Inference (Bowman et al., 2015, SNLI) corpus.

## 2 Related Work

Paraphrase generation has been recently explored as a statistical machine translation problem in a neural setting. Prakash et al. (2016) used a stacked-LSTM (Long Short-Term Memory) SEQ2SEQ network with residual connections and demonstrated strong performance over the simple and attention-enhanced SEQ2SEQ models. They report superior scores on several datasets: the Paraphrase Database corpus (Ganitkevitch et al., 2013, PPDB), captions from Common Objects in Context (Lin et al., 2014, MSCOCO), and question pairs from WikiAnswers (Fader et al., 2013). Mallinson et al. (2017) adapt the NMT architecture to incorporate bilingual pivoting and report improvements over the baseline in simi-

larity prediction, paraphrase identification as well as paraphrase generation.

Our work is different in that we focus on transfer learning to improve performance, using state of the art neural models employed mainly for machine translation.

Transfer learning has been recently investigated by Mou et al. (2016), who distinguish two settings: semantically equivalent transfer (where both source and target tasks are natural language inference) and semantically different transfer (where the source task is natural language inference and the target task is paraphrase detection). They report increased performance only in the former setting. Zoph et al. (2016) train a parent model on a high-resource language pair (such as English-French) in order to improve low-resource language pairs. They manage to improve the baseline with an average 5.6 BLEU points.

## 3 Experiments

Paraphrases can be seen as mostly bidirectional textual entailments (Androutsopoulos and Malakasiotis, 2010). Sentential paraphrase corpora are prohibitively small for training large neural networks, but textual entailment corpora are quite large thanks to the SNLI dataset. Our aim is to exploit this situation by performing transfer learning from the entailment generation (EG) task (given sentence $S$, generate sentence $T$ that can be inferred from $S$) to the paraphrase generation (PG) task. We also fine-tune the weights on the larger PPDB corpus before transferring to the Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005, MSRP), used for the paraphrase detection task. In addition, we also test the multiple transfer in reverse order.

### 3.1 Model

We use the state of the art SEQ2SEQ models with attention[1], described in Luong et al. (2015). We train a 2-layer LSTM with 2000 hidden units and word embeddings of size 1000.

### 3.2 Datasets

We use the MSRP and PPDB datasets featuring paraphrase pairs and the SNLI dataset featuring textual entailments. We discard the negative examples from the MSRP dataset. We discard the neutral and contradiction examples and only keep entailment pairs from the SNLI corpus. We also use the small (XS) phrasal subset of the PPDB dataset, due to its higher-scoring pairs as compared to the other variants of PPDB. We also augmented all datasets with the inverse pair $(Y, X)$ for each pair of sentences $(X, Y)$ - this approach is completely justified for paraphrases, but it also makes sense for SNLI if we treat an entailment pair just as a paraphrase pair.

| Dataset | Train | Validation | Test |
|---|---|---|---|
| MSRP | 3,854 | 1,652 | 2,294 |
| PPDB(XS) | 457,000 | 114,888 | - |
| SNLI | 183,416 | 3,329 | 3,368 |

**Table 1:** Datasets statistics (number of pairs)

The MSRP dataset is small, but it features long sentences with lots of numbers and proper nouns, which is rather problematic when predicting words from fixed-size vocabularies. The PPDB dataset contains a large number of short, but high-quality paraphrase pairs. We hypothesize that the SNLI entailments could prove useful in paraphrase generation, due to the large lexical overlap between the premise and the hypothesis.

An overview of the datasets and their train/validation/test sizes is shown in Table 1.

### 3.3 Transfer learning

In order to perform transfer learning in scenarios of type $X \longrightarrow Y$, where $X$ and $Y$ are two datasets for the same or different tasks, we follow the next steps. We train the SEQ2SEQ models on dataset X, keeping the configuration with the lowest perplexity on the validation set of X. We then transfer the parameters to a new model that are retrained on dataset Y.

Transfer learning in scenarios of type $X \longrightarrow Y \longrightarrow Z$ is similar to the process described above, but with the additional transfer from task/dataset Y to task/dataset Z.

All models are compared with the MSRP baseline, where a SEQ2SEQ model is trained on the MSRP training set alone.

[1] https://github.com/harvardnlp/seq2seq-attn

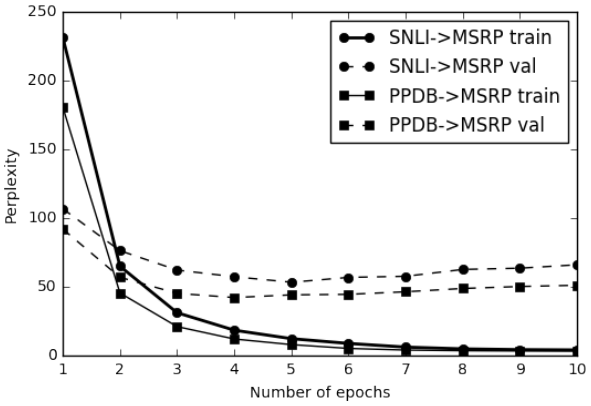


**Figure 1:** Perplexities with direct transfer

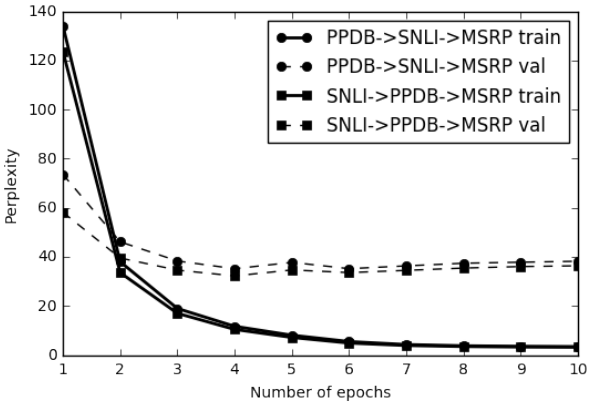


**Figure 2:** Perplexities with one-hop transfer

| Experiment name | test perplexity per word | BLEU score |
|---|---|---|
| MSRP | 14.37 | 0.09 |
| SNLI $\longrightarrow$ MSRP | 3.97 | 7.17 |
| PPDB $\longrightarrow$ MSRP | 3.73 | 10.29 |
| SNLI $\longrightarrow$ PPDB $\longrightarrow$ MSRP | 3.08 | 15.76 |
| PPDB $\longrightarrow$ SNLI $\longrightarrow$ MSRP | 3.78 | 12.91 |

**Table 2:** BLEU score and perplexity on the MSRP test set over different transfer scenarios

### 3.4 Training

All models are trained using stochastic gradient descent (SGD), with a learning rate decay of 0.5 if the validation perplexity does not decrease on consecutive epochs. The models are trained for 20 epochs each. We perform early stopping by keeping the configurations with the lowest perplexity on the validation set.

### 3.5 Evaluation

To generate paraphrases, we use beam-search with a beam size of 5. We report the BLEU score (Papineni et al., 2002) and the perplexity of the reconstructed sentences for the MSRP test corpus. Although no standard metric has proved conclusive for evaluating paraphrase generation, BLEU score has been shown to correlate fairly well with human judgements (Chen and Dolan, 2011), especially when more references are being used. We also plot the perplexity on the training and validations sets of different transfer scenarios.

## 4 Results

### 4.1 Quantitative results

**Transfer learning improves perplexity and BLEU score**

In Figure 1 we notice that transferring from the PPDB paraphrases yields lower perplexities than from the SNLI entailments. Not surprisingly, the additional transfer further lowers perplexity. However, Figure 2 shows that perplexity is slightly lower when transferring from entailments through additional paraphrases than the other way around.

The higher BLEU scores in Table 2 also seem to correlate well with the lower validation perplexities, with SNLI $\longrightarrow$ PPDB $\longrightarrow$ MSRP proving to be the best transfer setting. One possible explanation is that entailment pairs have a stronger lexical, but a weaker semantic overlap. Then phrasal (multi-word) paraphrases from PPDB are used to improve the semantic equivalence needed for paraphrasing. We turn to qualitative analysis, where we generate sentences using the SNLI $\longrightarrow$ PPDB $\longrightarrow$ MSRP model.

### 4.2 Qualitative results

Results in Table 3 show that the model is able to restate parts of the input sentences, but fails to retain the whole meaning.

In the first two examples, the models drops the proper names and the description following the dialogue. The second example shows little variation in the input sentence. The third example reflects a more diverse vocabulary, but again suffers from dropping parts of the input. The fourth example re-

| Source sentence | Decoded sentence | Target sentence |
|---|---|---|
| current chief operating officer mike butcher and group chief financial officer alex arena will report to so | the chief executive officer and chief financial officer to report to so . | pccw ' s chief operating officer , mike butcher , and alex arena , the chief financial officer , will report directly to mr so . |
| " there ' s no reason for you to keep your skills up , " the judge told the convicted crack cocaine kingpin . | " there is no reason for you to keep your skills up . " | " there ' s no reason for you to keep your skills up , " u . s . district judge j . frederick motz told mcgriff after he was sentenced . |
| those reports were denied by the interior minister , prince nayef . | such reports were refused by internal affairs . | however , the saudi interior minister , prince nayef , denied the reports . |
| the letter bomb sent to prodi exploded in his hands but he was unhurt . | the letter was sent to prodi in his hand but he was surrounded . | it exploded in his hands , but the former italian prime minister was unhurt . |

**Table 3:** Four examples of source sentences from the MSRP test set, along with the decoded and the target sentences

tains part of the original meaning, but doesn't contain important words such as 'bomb' and 'exploded'.

The truncation effect may be due to training on the entailment pairs, because most of the hypotheses featured in the SNLI dataset are shorter than the premises.

Also, without a copying mechanism, it is challenging for SEQ2SEQ models to predict proper names, especially if they are rare or out of training vocabulary.

## 5 Conclusions and future work

In this paper, we investigated the use of SEQ2SEQ neural models for paraphrase generation. The major limitation in training such models is the shortage of corpora with (complex) sentential paraphrases, which we overcame by performing transfer learning, first using textual entailment and then phrasal paraphrase pairs.

We showed that transfer learning improves the BLEU score of the generated paraphrases in all transfer settings and that transfer works best when transferring entailments to short paraphrases and then to the longer paraphrases from the MSRP corpus.

Qualitative results showed promising results, with the model being able to restate parts of the input sentence fairly well. Further areas of research should address the lexical variety and should look into incorporating copying mechanism into the network so that rare or unknown words are picked up during paraphrasing.

## References


Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

D L Chen and W B Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 190–200.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk,

and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pages 9–16.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-Driven Learning for Open Question Answering. *Acl*, pages 1608–1618.

Andrew Finch, Taro Watanabe, Yasuhiro Akiba, and Eiichiro Sumita. 2004. Paraphrasing as machine translation. *Journal of Natural Language Processing*, 11(5):87–111.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: the paraphrase database. pages 758–764.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir D Bourdev, Ross B Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft {COCO:} Common Objects in Context. {*arXiv*}*:1405.0312*, pages 740–755.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain, April. Association for Computational Linguistics.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? pages 479–489.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural Paraphrase Generation with Stacked Residual LSTM Networks. *Coling*.

Chris Quirk, Chris Brockett, and William B. Dolan. 2004. Monolingual machine translation for paraphrase generation. pages 142–149.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. *Emnlp*, page 8.