DS-GA 1012: Text-style Transformation from 19th-Century Literature to Modern Language

Chutang Luo
New York University
cl5293@nyu.edu

Tianshu Chu
New York University
tc2992@nyu.edu

Yuwei Wang
New York University
yw1854@nyu.edu

Jiarui Tang
New York University
jt3869@nyu.edu

Abstract

Transforming text styles of classical literature to plain English helps reader approach contents more easily and can serve for various educational purposes. However, most style transfer tasks focus on transforming attributes of sentiment and formality; only a few work has been done for translating literature pieces. In this paper, we explore existing literature style transfer models on limited parellel data. We introduce a new dataset from *A Tale of Two Cities* by Charles Dickens (scraped from the website Sparknotes [1] with parallel sentence pairs from original text and daily English. Then we trained neural models to translate literary pieces into plain English using the Encoder-Decoder Sequence to Sequence framework. We conduct experiments to compare Seq2Seq model with Pointer Sentinel framework, and test performance on different settings of embedding layer. Our best model with pre-trained embeddings reached BLEU score 31.32 and greatly outperformed the baseline model.

1 Introduction

Text style transfer is an important but challenging task in the field of natural language processing and understanding. It requires a model to automatically transfer the input text of one style to another while keeping the content unchanged. There are many subtasks within text style transfer task (e.g. transformation of text from negative sentiment to positive (Wu et al., 2019) or from informal to formal (Xu et al., 2019)). However, there are only few work on the task of transferring classical literacy style to plain style. Literature style transfer can produce pieces for readers to take as reference when approaching classical pieces, and can benefit the popularization of elusive literature and serve for educational purposes. In recent years, the most robust study on literature style transfer is the work between modern English and ancient Shakespearean English (Jhamtani et al., 2017). Due to the particularity of Shakespearean English, it remains in doubt whether the method of transformation between Shakespearean English and modern English would be applicable to other literacy genres. Therefore, we would like to experiment the method by (Jhamtani et al., 2017) on a new dataset in new literacy genre.

Generally, the text data for style transfer model can be put into two categories: parallel training data (e.g. different versions of Bible) and nonparallel training data (e.g. reviews on Yelp). Due to the large amount of nonparallel data and its application on business operation, most advanced text style transfer models were raised for nonparallel data. The work on parallel style transfer mostly suffers from low-resource of parallel sets. To solve the problem, Jhamtani et al. (2017) pre-trained embeddings with the help of external dictionaries translating Shakespearean words to modern English words. However, for most parallel text beyond Shakespearean corpora, it is not feasible to find such external dictionaries.

Instead of using the external dictionaries, Shang et al. (2019) added large amount of nonparallel text in addition to the original small-scale parallel training data. Following this idea, we experimented some pre-trained embeddings that use external data of Penn Treebank (PTB) (Marcus et al., 1993) and GloVe (Pennington et al., 2014). We also introduce a new dataset A-Tale-of-Two-Cities collected from the literature *A Tale of Two Cities* for literacy style transformation other than Shakespearean English. The new dataset contains 4,325 parallel sentences as main training dataset, which are split and filtered from the original text.

In this paper, the baseline model is a simple Seq2Seq model and reaches only 5.88 in test BLEU

[1] https://www.sparknotes.com/lit/

score. Pointer sentinel is further added on the Seq2Seq model to improve the performance and improve test BLEU score to 23.14. To overcome the issue of limited parallel data size, we use different external word embedding method to pre-train our model, before training encoder-decoder sequence to sequence model on the main training dataset. The most effective way is adding external text of PTB (Marcus et al., 1993) for pre-training embeddings, with BLEU score of 31.32 compared with using GloVe (Pennington et al., 2014) representations directly as embeddings, with BLEU score of 28.67.

In summary, our contributions in this paper are as follows:

1. We introduce a new dataset collected from the literature which contains high quality 4,325 parallel sentences for style transformation tasks.

2. We implement the encoder-decoder Seq2Seq Model with the Pointer Sentinel to transfer the literature style written by Charles Dickens to plain English.

3. We show that using pre-trained embeddings from GloVe and PTB text both efectively improve the model performance to overcome the low-resource of parallel data set.

2 Background

Text-style transformation is an extensively discussed NLU topic that could involve sentiment, formality, literacy style or any other sentence attributes. To successfully automate this procedure, a number of researchers have come up with different models and algorithms for a wide range of data sets.

2.1 Non-parallel Data

Considering various difficulties in the process of generating large corpora of paralleled data, most of work are focused on text-style transformation for unparalleled data sets. Shen et al. (2017) managed to solve the problem through adversarial training. With the assumption of shared latent content distribution between original and target sentences, they disentangled a styled-independent content representations for each input sentence and combined it with the target style to generate the desired output. Zhao et al. (2018) extended Shen et al. (2017)'s research to text-style transfer between multiple original styles to one specific target style. They also introduced a style discrepancy loss and a cycle consistency loss to ensure that the style content representations are correctly disentangled and the content information are preserved during transformation respectively. Our model adapts the similar idea of the encoder-decoder framework as in Shen et al. (2017) and Zhao et al. (2018). However, considering the parallel nature of our data, we do not disentangle the style from content during the encoding process.

Li et al. (2018) proposed a totally different Delete-Retrieve-Generate framework for unparalleled data. To generate a fluent sentence in target domain, they first disentangled and deleted the original attribute-dependent phrases from the sentence, then they retrieved new target attribute-dependent phrases to combine with the remaining words using a neural model. This approach performs better at preserving content and structural information than the encoder-decoder framework, especially for longer input sentences. However, We failed to adapt this Delete-Retrieve-Generate framework to our data sets, since it would be difficult to completely separate the attribute related phrases from the remaining content words in the case of transformation between 19th-century literature and plain English. Other researchers have also proposed different models, such as back-translation (Prabhumoye et al., 2018; Subramanian et al., 2018) and seq2seq (Xu et al., 2019), to conduct style transformation for unparalleled data.

2.2 Parallel Data

On the other hand, Carlson et al. (2018) worked on generating a large corpora of highly paralleled data with different Bible versions to improve performance of existing models. Wang et al. (2019) achieved a new state-of-the-art on paralleled data sets for formality transformation by combining GPT-2 model (Radford et al., 2019) with a rule-based system.

Furthermore, Jhamtani et al. (2017)'s model was built upon a sentence level sequence to sequence neural model to perform parallel transformation between contemporary English and Early Modern English used in Shakespeare. To mitigate the greatly increasing number of parameters while learning token embeddings, they implemented a retrofitting method (Faruqui et al., 2015) with a mapping dictionary between original and modern word pairs to pre-train the embeddings on all training sentences along with external data source (PTB). They even-

tually achieved a BLEU score of 31+ by adding a pointer network component (Merity et al., 2016) to directly copy input words. Our model is largely based on Jhamtani et al. (2017)'s work, but we extend it to different paralleled data sets with several different word embeddings. The pointer sentinel mixture architecture first introduced by Merity et al. (2016) largely improves our model performance by maintaining the contextual and structural coherence after the transformation as it did in Jhamtani et al. (2017)'s research. Merity et al. (2016) incorporated the standard softmax classifier with existing pointer networks (Vinyals et al., 2015) to predict words based on immediate context information. Their model has shown high performance on both short sentences and larger corpora with long range dependencies.

3 Dataset

Our dataset is a collection of sentences of the literature *A Tale of Two Cities* from the educational site Sparknotes [2]. The source provides two versions of the text: the original literature written by Charles Dickens, and a plain translation with modern, easy-to-understand English. The two versions are matched almost perfectly on a paragraph level, so it provides a good source for parallel style transformation.

In general, the original version is more verbose, containing longer and complex sentences. It also includes larger vocabulary and unique expressions in mid-19th century of England. Also, we observed that one original sentence with many clauses may have been broken into several shorter sentences in modern version. These attributes are critical for us to take into consideration how to reconstruct parellel sentence pairs for training and test data.

Type	Sentence
Original	There were a king with a large jaw and a queen with a plain face , on the throne of England .
Modern	A stern looking king and a plain looking queen ruled England.
Original	Never you mind what it is .
Modern	Never mind what it is .
Original	Their conference was very short , but very decided .
Modern	Their conversation was short but to the point .

Table 1: Sentence pair sample.

We crawled both versions of texts and performed initial preprocessing (remove spaces, line breaks, replace special characters with ASCII letters). Then for each paragraph, we compared number of sentences from both versions, and filtered out paragraphs with unequal number of sentences, in order to make sure the remaining text are parellel on the sentence level. Finally, we filtered sentences by length and kept only those no more than fitty words; this is because our model would suffer with too large variance of input sequence length. The original text consists of 3269 paragraphs; after splitting and filtering, the real training set contains 4325 parallel sentences. We splited them in random orders into train, valid, test sets of sizes 2395, 865, 865 pairs, respectively.

[2] https://www.sparknotes.com/lit/

4 Model Overview

Most of the model architecture follows the framework of Jhamtani et al. (2017)'s research. In general, we use Encoder-Decoder Sequence to Sequence Model to complete the text-style transformation. The highlight of previous research was to adopt the Pointer Sentinel framework, to firstly preserve rare words such as character names from the input, prevent them from going into neural transformation, and directly map onto the output text (Merity et al., 2016). In our work, we also adopt Pointer framework to remember character names of the novel, and have proven it successful in improving the transformation.

Following the framework of Jhamtani et al. (2017), our model uses a bidirectional LSTM to encode the sentence from the original version of the literature. The decoder model is combination of RNN and pointer network.

To overcome the problem of limited training data size, we used two different pre-trained embeddings from larger external corpora. Out first attempt was to pre-train word embeddings from PTB with the internal training set, which followed the work of Jhamtani et al. (2017). The second was to directly using GloVe as the pre-trained embedding layer.

We used BLEU as our evaluation metrics to compare model performance.

5 Experiments

Following the work done by Jhamtani et al. (2017), we used a minibatch-size of 32 and the ADAM optimizer with learning rate 0.001. The original implementations are written in Python using Tensorflow 1.1.0 framework, we have adapted minor changes for codes to work for Tensorflow 1.15.0.

We share our codes and work to public [3].

For every model, we experimented LSTM sizes of 128-128 and 192-192; and max sequence length 32 and 48. We are not performing dropout for now when training models, although dropout function is enabled in the implementation. At test time, we used greedy decoding to generate sentences with highest probability.

We have found the prediction with LSTM 128-128 with max sequence input size 32 produces the best score. All model performance presented used this combination of hyperparameters. We trained each experiment with 15 epoches, and used the epoch with highest validation score to predict test results.

5.1 Baseline model - Simple Seq2Seq

We used simple Seq2Seq model as our baseline. The baseline model performs poorly on transformation, reaches only 5.88 in BLEU score and can hardly predict any reasonable English sentences.

5.2 Seq2Seq with Pointer Sentinel

Our first improvement is to add Pointer on top of Seq2Seq as comparison. It significantly improved prediction accuracy, reaches test BLEU score 23.14. We found out the model can successfully predict some short sentences, but still struggles with putting longer sentences into readable English.

5.3 Pre-trained embeddings from PTB text

From this step on, we want to explore whether adopting pre-trained embeddings from external text will improve our model performance. We pre-trained embeddings from Penn Treebank (PTB) text (Marcus et al., 1993) in combination with our training set. The test BLEU score reaches 31.32, significantly higher than embeddings from training text alone.

5.4 Pre-trained embeddings using GloVe

We also conducted an experiment to use Global Vectors for Word Representations (Pennington et al., 2014) directly as embedding layer. We used GloVe.6B 200d vectors, which was trained on Wikipedia 2014 + Gigaword 5 text sets, and arbitrarily chunked out the last 8 dimensions in order to fit our 192-192 LSTM size. The test BLEU score reaches 28.67, lower than pre-trained PTB, but still significantly higher than no pre-trained embedding.

5.5 Performance summary

Table 2 shows the complete validation and test results from all experiments.

Table 2: Tuned hyperparameters on Ranking Metrics

Model	(Best) Validation BLEU	Test BLEU
Seq2Seq	5.76	5.88
Seq2Seq with Pointer	23.66	23.14
Pre-train with PTB	27.56	**31.32**
GloVe	27.10	28.67

Table 3 shows some sample outputs from our best performance model (pre-trained with PTB).

No	Type	Text
1	Original	He didn't precisely remember where it was.
	Modern	He didn't quite remember.
	Prediction	He didn't remember where it.
2	Original	Have you seen him , to your certain knowledge?
	Modern	Have you ever seen him before?
	Prediction	Have you seen him , your certain?
3	Original	Give me your authority , like a dear good man.
	Modern	Give me your permission , like a good man.
	Prediction	Give me your authority , like a man.

Table 3: Sample outputs with PTB pre-trained embeddings

6 Conclusion

In this paper, we propose a new parallel data set collected from *A Tale of Two Cities*, and prove it can be used for style transfer tasks with seq2seq model. We also show Pointer Sentinel works effectively for improving performance of literature style transfer, in comparison with simple seq2seq model. In order to compensate the limited size of the parallel sentences, we build our model with two different external text embeddings to compare the performance with the baseline model. Both models with external text perform significantly better than embeddings from training set alone. Thus, we show that the pre-training process with larger external corpora could be beneficial for text-style transformation with limited parallel data size.

For the future studies, we may experiment with more state-of-art pre-trained embeddings such as BERT to improve model performance. Also, more explorations can be done to overcome the limitation of low-resource parallel sentences, including effective sentence alignment strategies, auto-encoder for pre-train, etc. to utilize the text which was not included in our reconstructed sentence sets.

[3] https://github.com/yuwei-jacque-wang/Literature-Style-Transfer-DSGA1012

Collaboration Statements

- Research and idea: everyone
- Literature Review: Tianshu Chu, Jiarui Tang
- Data scraping and processing: Chutang Luo, Tianshu Chu, Jiarui Tang
- Code implementation: Yuwei Wang, Chutang Luo
- Experiments: Yuwei Wang
- Paper write-up: everyone

References

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society Open Science* 5(10):171920. https://doi.org/10.1098/rsos.171920.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1606–1615. https://doi.org/10.3115/v1/N15-1184.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*. Association for Computational Linguistics, Copenhagen, Denmark, pages 10–19. https://doi.org/10.18653/v1/W17-4902.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 1865–1874. https://doi.org/10.18653/v1/N18-1169.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.* 19(2):313–330.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR* abs/1609.07843. http://arxiv.org/abs/1609.07843.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pages 866–876. https://doi.org/10.18653/v1/P18-1080.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners .

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pages 4937–4946. https://doi.org/10.18653/v1/D19-1499.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *CoRR* abs/1705.09655. http://arxiv.org/abs/1705.09655.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *CoRR* abs/1811.00552. http://arxiv.org/abs/1811.00552.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *ArXiv* abs/1506.03134.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pages 3573–3578. https://doi.org/10.18653/v1/D19-1365.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and infill: Applying masked language model for sentiment transfer. *ArXiv* abs/1908.08039.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *CoRR*

abs/1903.06353. http://arxiv.org/abs/1903.06353.

Yanpeng Zhao, Wei Bi, Deng Cai, Xiaojiang Liu, Kewei Tu, and Shuming Shi. 2018. Language style transfer from sentences with arbitrary unknown styles. *CoRR* abs/1808.04071. http://arxiv.org/abs/1808.04071.