DL Project Interim Report: Shakespeare - English Sequence to Sequence Modeling

Morris Kraicer
Johns Hopkins University
mkraice1@jhu.edu

Riley Scott
Johns Hopkins University
rscott39@jhu.edu

William Watson
Johns Hopkins University
billwatson@jhu.edu

Abstract

We present a interim report on our efforts to contruct a sequence-to-sequence neural translation model with attention. We present the inital work done on data procurement and processing, and provide statitics on our datasets. Additionally, we describe in detail our models that we are experimenting with. Finally, we describe the overall strucutre of our code to support sequence to sequence modeling.

1 Introduction

2 Data Procurement

Most of our data was procured from [6]. This includes all of Shakespeares plays translated line by line into modern English. However, since the data was aligned, not all of the original lines from the plays are included (the sentences could not be aligned properly). [1]

3 Preprocessing

3.1 SOS and EOS

3.2 Propernouns

3.3 Train, Validation, Test Split

We will use spacy.io to tokenize our dataset to replace proper nouns with pronouns.[2] In addition, we will tokenize numbers and punctuation to reduce our vocabulary size. Finally, we will lowercase the data. We will randomly sample a 70/20/10 split for train/dev/test on the full merged dataset.

[1] `https://github.com/cocoxu/Shakespeare`
[2] `https://spacy.io`

4 Architectures

We seek to develop several models to improve our translations, incorporating context, attention, and novel training methods.

4.1 Baseline RNN Sequence to Sequence

Our baseline model will be a simple RNN sequence to sequence model. It will accept a source sentence s and will decode it to an output sequence t. A simple RNN will only incorporate previous states, i.e. source words, during prediction, and might be prone to the vanishing gradient problem. It will incorporate an encoder-decoder style model [2] [5].

4.2 GRU Sequence to Sequence

We will improve our baseline through the use of the GRU layer, to hopefully offset any vanishing gradients. In addition, we use a GRU over an LSTM to help bound the number of parameters used in the model for training purposes, as the GRU and LSTM fix issues regarding long term dependencies within sequences.

4.3 Bidirectional Model

In order to consider the context of words before and after, we will alter the model to use a bidirectional GRU [1], and hopefully see better translations.

#	**Original Sentence**	**Modern Sentence**
1	there s beggary in the love that can be reckoned .	it would be a pretty stingy love if it could be counted and calculated .
2	poor souls , they perished .	the poor people died .
3	the propn s abused by some most villainous knave , some base notorious knave , some scurvy fellow .	the propn is being tricked by some crook , some terrible villain , some rotten bastard .
4	my best way is to creep under his gaberdine .	the best thing to do is crawl under his cloak .
5	when they do choose , they have the wisdom by their wit to lose .	when they choose , they only know how to lose .
6	her blush is guiltiness , not modesty .	she blushes from guilt , not modesty .
7	go , hang yourselves all !	go hang yourselves , all of you !
8	then , if ever thou propn acknowledge it , i will make it my quarrel .	then , if you dare to acknowledge it , i ll take up my quarrel with you .
9	and lovers ' absent hours more tedious than the dial eightscore times !	and lovers ' hours are a hundred and sixty times longer than normal ones !
10	i have no great devotion to the deed and yet he hath given me satisfying reasons .	i do n t really want to do this , but he s given me good reasons .

Figure 1: Sample Original-Modern Sentence Pairs

4.4 Attention Mechanisms

Attention mechanisms have been shown to improve sequence to sequence translations from Bahdanau et al [1], and further work from Luong et al [4] examines global vs local approached to attention-based encoder-decoders.

4.5 Teacher Forcing

In terms of training, an encoder-decoder system can either accept the target token or the model's prediction as input during the decoding step. When we use the target token, this is known as teacher forcing, and is shown to be favored during initial trainging iterations, but should be backed off to use the model's own predicitons, as it will exhibit instability in the translations otherwise [3].

4.6 Encoders

4.6.1 Baseline RNN Encoder

$$h_t = \tanh(W_{ih}x_t + b_{ih} + W_{hh}h_{t-1} + b_{hh}) \tag{1}$$

4.6.2 GRU Encoder

$$\begin{aligned} r_t &= \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{t-1} + b_{hr}) \\ z_t &= \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{t-1} + b_{hz}) \\ n_t &= \tanh(W_{in}x_t + b_{in} + r_t \circ (W_{hn}h_{t-1} + b_{hn})) \\ h_t &= (1 - z_t) \circ n_t + z_t \circ h_{t-1} \end{aligned} \tag{2}$$

4.6.3 Bidirectional GRU Encoder

$$\begin{aligned} \overrightarrow{h_f} &= \text{GRU}(\overrightarrow{input}) \\ \overleftarrow{h_b} &= \text{GRU}(\overleftarrow{input}) \\ h_o &= \overrightarrow{h_f} \parallel \overrightarrow{h_b} \end{aligned} \tag{3}$$

4.7 Decoders

4.7.1 Baseline RNN

4.7.2 GRU Decoder

4.8 Attention Mechanisms

$$a(s_{i-1}, h_j) = \begin{cases} W_a(s_{i-1} \| h_j) & \textit{concat} \\ s_{i-1}^T \cdot W_a h_j & \textit{general} \\ s_{i-1}^T \cdot h_j & \textit{dot} \\ W_a \cdot s_{i-1} & \textit{location} \end{cases} \tag{4}$$

4.8.1 Dot Attention

4.8.2 Concat Attention

4.9 Teacher Forcing

5 Planned Experiments

6 Current Status and Expectation

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[3] A. M. Lamb, A. G. A. P. GOYAL, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, pages 4601–4609, 2016.

#	**Encoder**	**Decoder**	**Attention**	**Teacher Forcing**
1	RNN	RNN	None	False
2	GRU	GRU	None	False
3	Bidirectional GRU	GRU	None	False
4	Bidirectional GRU	GRU	Concat	False
5	Bidirectional GRU	GRU	General	False
6	Bidirectional GRU	GRU	Concat	True
7	Bidirectional GRU	GRU	General	True

Figure 2: Planned Model Experiments

[4] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[5] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[6] W. Xu, A. Ritter, B. Dolan, R. Grishman, and C. Cherry. Paraphrasing for style. In *COLING*, pages 2899–2914, 2012.