HW3: Translation

Alex Lin
alexanderlin01@college.harvard.edu

Melissa Yu
melissayu@college.harvard.edu

Team Name: PUDGE

March 2, 2018

1 Introduction

With the increasing globalization of our world, translation has become more and more important. Prior to 2014, many automatic translation systems were either based on linguistic rules or statistical models. However, with the emergence of deep learning, neural machine translation systems with end-to-end capabilities now dominate the field.

In this problem set, we examine two such models: 1) a baseline sequence-to-sequence model and 2) an attention-driven deep learning model. Our overall objective is to create translation systems that can translate German sentences to English sentences. For an out-of-sample German test set comprised of 800 sentences, we generate 100 of the most probable 3-gram English starting-sentence translations as determined by our models.

2 Problem Description

Let $\mathcal{V}_X$ be the German vocabulary and $\mathcal{V}_Y$ be the English vocabulary. Let each word in a vocabulary be represented by a one-hot encoded vector with length equal to the vocabulary. Given a German sentence (i.e sequence of words) $\boldsymbol{x} = [x_1, x_2, \ldots, x_S]$ and an English partial translation $[y_1, y_2, \ldots, y_{t-1}]$, the objective of translation is to accurately predict the distribution of the next English word

$$y_t | y_1, \ldots, y_{t-1}, x$$

for $t = 1, 2, \ldots, T$ where T is the length of y. Let q be our model's predictive distribution. Let $(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$ be a single pair of German-English sentences that mean the same thing in our training set, for $n = 1, 2, \ldots, N$. In training a model, we wish to adjust the model's parameters to minimize the *average loss*, as defined by

$$L = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{T} \sum_{t=1}^{T^{(n)}} \left(\ln q(y_t^{(n)} | y_1^{(n)}, \ldots, y_{t-1}^{(n)}, x^{(n)}) \right) \cdot y_t^{(n)}$$

where $\cdot$ denotes the inner dot product and $T^{(n)}$ denotes the length of $\boldsymbol{y}^{(n)}$. The standard metric used for evaluation is called *perplexity*, as defined by

$$PPL = \exp L$$

3 Model and Algorithms

We trained the three different models specified in the instructions. These included (1) a baseline sequence-to-sequence model and (2) an attention-drive sequence-to-sequence model.

3.1 Baseline Sequence-to-Sequence

The baseline sequence-to-sequence model is comprised of an encoder LSTM and a decoder LSTM. The encoder reads an input sentence $\boldsymbol{x}$ with length T_x into a fixed-length context vector c, as follows:

$$\begin{aligned} h_t &= f(x_t, h_{t-1}) \quad \forall t = 1, \ldots, T_x \\ c &= j(\{h_1, \ldots, h_{T_x}\}) \end{aligned}$$

where h_t is the hidden state of the encoder with $h_0 = \mathbf{0}$, and j is a function of the hidden states. In this case, we simply let $j(\{h_1, \ldots, h_{T_x}\}) = h_{T_X}$ (i.e. the last hidden state). Next, we let $s_0 = c$ be the first previous hidden state of the decoder. The decoder follows the recursion:

$$s_t = g(y_t, s_{t-1}) \quad \forall t = 1, \ldots, \tilde{t}$$

where y_t is the t-th word of the supplied *partial* output sentence with length $\tilde{t}$. To predict the probability distribution over the $(\tilde{t}+1)$-th word in the output, we simply have

$$q(y_{\tilde{t}+1} | y_1, \ldots, y_{\tilde{t}}, \boldsymbol{x}) = \text{softmax}(Ws_t + b)$$

where W is a $|V_Y| \times n_{\text{hid}}$ matrix and b is an n_{hid}-dimensional vector. Note that

$$n_{\text{hid}} = \dim s_t = \dim h_t$$

is the number of hidden states in each of the two LSTMs.

3.2 Sequence-to-Sequence with Attention

4 Experiments

We performed several experiments to tune the hyper-parameters of our models. The attention model seemed to perform the best. In-depth explanations of our tuning procedure and associated results can be found in this section. Note that all perplexities reported in table 1 are test set perplexities.

Model	Validation Perplexity	Kaggle Score
BASELINE SEQUENCE-TO-SEQUENCE	13.831	0.24806
SEQUENCE-TO-SEQUENCE WITH ATTENTION		

Table 1: Language perplexities and Kaggle scores of our models.

4.1 Baseline Sequence-to-Sequence

The main parameters that we tuned were (1) the embedding dimensions of the two vocabularies, (2) the number of hidden states n_{hid}, (3) the number of hidden layers in each LSTM, (4) the dropout rate, and (5) whether or not to reverse the input sentence $\boldsymbol{x}$.

The model that obtained the best validation perplexity of 13.831 had embedding dimensions of 1000 each, $n_{\text{hid}} = 1000$, 2 hidden layers in each LSTM, a dropout rate of 0.2, and reversed input. We used a stochastic gradient descent optimizer that ran for 16 epochs with an initial learning rate of 1 that was successively halved during the 7th and 14th epochs. The batch size was 32 and the maximum norm of any gradient used during training was set as 5.

Here are our results from varying embedding dimensions (assume all other variables are as listed above). We found that increasing the embedding dimensions too much led to overfitting (i.e. low train perplexity, but higher test perplexity).

Embedding Dimension	100	200	500	1000	1500
Val. Perplexity	15.82	14.52	14.01	13.83	15.27

Here are our results from varying n_{hid}.

n_{hid}	100	200	500	1000	1500
Val. Perplexity	14.11	14.15	13.87	13.83	13.94

Here are our results from varying the number of hidden layers in the two LSTMs.

Number of Hidden Layers	1	2	3	4
Val. Perplexity	18.81	13.87	14.66	14.67

Here are our results from varying the dropout rate.

Dropout Rate	0.1	0.2	0.3	0.4
Val. Perplexity	14.08	13.87	13.92	13.91

And here are our results for deciding whether or not to reverse the input sentence $\boldsymbol{x}$.

	No Reverse	Yes Reverse
Val. Perplexity	13.99	13.87

To generate predictions for the Kaggle competition, we used beam search with a beam of 100. Our Kaggle score obtained from the baseline sequence-to-sequence model trained with the most optimal parameters is 0.24806.

4.2 Sequence-to-Sequence with Attention

5 Conclusion