Reimplementing End-to-End Generative Dialogue

Michael Farrell, Colton Gyulay, Kevin Yang

May 12, 2016

Objectives

- Main focus of our project is to re-implement and explore extensions for the dialogue system described in Serban et al's (2016) paper: *Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models*.
- Use the Movie Triples dataset. Dataset with around 200,000 utterance triples (U_1, U_2, U_3). Goal is to predict U_3 given U_1 and U_2.

```
U1: you lied to me so many times --
U2: reggie -- trust me once more -- please .
U3: can i really believe you this time , <person> ?
```

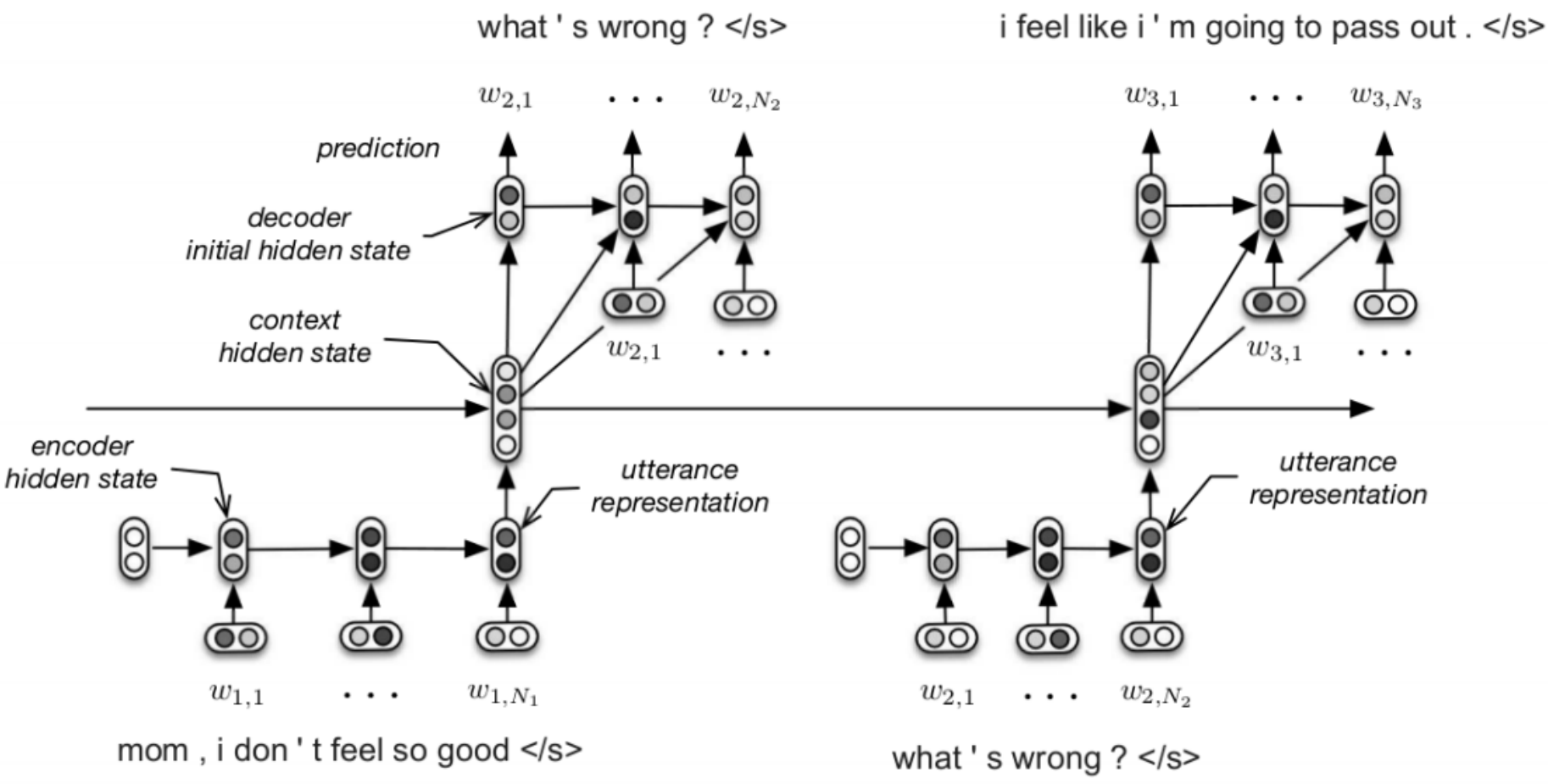
Objectives

- For generating dialogue, authors of the paper use perplexity and error-rate.
- They used a variety of different models for their encoder and decoder. Some of their results are shown below.

Model	Perplexity	Perplexity@U_3	Error-Rate	Error-Rate@U_3
Backoff N-Gram	64.89	65.05	-	-
Modified Kneser-Ney	60.11	54.75	-	-
Absolute Discounting N-Gram	56.98	57.06	-	-
Witten-Bell Discounting N-Gram	53.30	53.34	-	-
RNN	35.63 ± 0.16	35.30 ± 0.22	$66.34\% \pm 0.06$	$66.32\% \pm 0.08$
DCGM-I	36.10 ± 0.17	36.14 ± 0.26	$66.44\% \pm 0.06$	$66.57\% \pm 0.10$
HRED	36.59 ± 0.19	36.26 ± 0.29	$66.32\% \pm 0.06$	$66.32\% \pm 0.11$
HRED + Word2Vec	33.95 ± 0.16	33.62 ± 0.25	$66.06\% \pm 0.06$	$66.05\% \pm 0.09$
RNN + SubTle	27.09 ± 0.13	26.67 ± 0.19	$64.10\% \pm 0.06$	$64.07\% \pm 0.10$
HRED + SubTle	27.14 ± 0.12	26.60 ± 0.19	$64.10\% \pm 0.06$	$64.03\% \pm 0.10$
HRED-Bi. + SubTle	$\mathbf{26.81 \pm 0.11}$	$\mathbf{26.31 \pm 0.19}$	$\mathbf{63.93\% \pm 0.06}$	$\mathbf{63.91\% \pm 0.09}$

- The model that performs the best is the hierarchical encoder decoder (HRED) model.

Hierarchical Encoder Decoder Model



Results

- We reached out to Iulian Serban to get the Movie Triples dataset. Using the Element rnn library, we trained encoder and decoders with RNN, GRU, and LSTM layers. To generate predictions, we implemented and ran beam search. We also ran Yoon Kim's seq2seq-attn model [1].
- Our best model had 2 stacked layers of LSTMs for both the encoder and the decoder with word2vec embeddings. Our validation perplexity reached a low of 38.2 after 10 epochs. This is comparable to results from Serban et al.'s models without bootstrapping from the SubTle dataset.

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

Results: Example Inputs and Outputs

- Our results did suffer from similar issues that Serban et al. had though, where utterances were short and ambiguous. Generally, lower validation perplexity did not necessarily mean better results.

```
U1 : when do these boys of yours go on the road ?
U2: coupla weeks . for eight weeks .
Predicted U3:   1) that ' s right .
                2) that ' s not the point .
                3) that ' s not what i meant .
                4) that ' s what i ' m talking about .
                5) that ' s what i ' m saying .

Actual U3: that ' s a nice tour . all booked ?
```

Results: Example Inputs and Outputs

U1 : me too . we ’ d love to have you .

U2: you know, this is the best pie i ’ ve ever had .

Predicted U3:
1) <person> .
2) i don’t know
3) i’m sorry
4) that’s good
5) i don’t know what

Actual U3: oh?

Results: Speeding Up Training

- The major bottleneck for training these models was time. Training a single epoch on a Macbook Pro with an NVIDIA GeForce 750M took around 3 hours per epoch.
- In order to expedite this process, we applied the lua-parallel [2] module to the problem. This produces a 4x speedup with CPU. Currently, we're trying to extend this module across multiple Google Cloud Compute instances to further increase the speed of training.

[2] https://github.com/clementfarabet/lua—parallel

Further Extensions

- We would like to experiment with the HRED model and see if it achives similar results to Serban et al's results.
- After implementing a distributed way to train models, we would like to do some parameter tuning and work with the SuBtle dataset.
- We would like to experiment with adding more context into the model. Currently, we're only using two utterances as context.
- Finally, we would like to try to generate more interesting conversation output.