Extending Neural Generative Conversational Model using External Knowledge Sources

Prasanna Parthasarathi
McGill University, Canada
pparth2@cs.mcgill.ca

Joelle Pineau
McGill University, Canada
Facebook AI Research, Canada
jpineau@cs.mcgill.ca

Abstract

The use of connectionist approaches in conversational agents has been progressing rapidly due to the availability of large corpora. However current generative dialogue models often lack coherence and are content poor. This work proposes an architecture to incorporate unstructured knowledge sources to enhance the next utterance prediction in chit-chat type of generative dialogue models. We focus on Sequence-to-Sequence (Seq2Seq) conversational agents trained with the Reddit News dataset, and consider incorporating external knowledge from Wikipedia summaries as well as from the NELL knowledge base. Our experiments show faster training time and improved perplexity when leveraging external knowledge.

1 Introduction

Much of the research in dialogue systems from the last few years has focused on replacing all (or some) of its components with Deep Neural Network (DNN) architectures (Levin and Pieraccini, 1997; Dahl et al., 2012; Li et al., 2017; Serban et al., 2016a,b; Vinyals and Le, 2015). These DNN models are trained end-to-end with large corpora of human-to-human dialogues, and essentially learn to mimic human conversations.

Although these models can represent the input context, the need for a dedicated external memory to remember information in context was pointed out and mechanisms were introduced in models like Memory Networks (Sukhbaatar et al., 2015; Bordes et al., 2016; Gulcehre et al., 2018), and the Neural Turing Machine (Graves et al., 2014). Although these models, in theory, are better at maintaining the *state* using their memory component, they require longer training time and excessive search for hyperparameters.

In this paper we explore the possibility of incorporating external information in dialogue systems as a mechanism to supplement the standard context encoding and facilitate the generation to be more specific with faster learning time. Furthermore, especially in the case of chit-chat systems, knowledge can be leveraged from different topics (education, sports, news, travel, etc.). Current memory-based architectures cannot efficiently handle access to large unstructured external knowledge sources.

In this work, we build on the popular Encoder-Decoder model, and incorporate external knowledge as an additional continuous vector. The proposed Extended Encoder-Decoder (Ext-ED) architecture learns to predict the embedding of the relevant external context during training and, during testing, uses an augmented *state* of external context and encoder final state to generate the next utterance. We evaluate our model with experiments on Reddit News dataset, and consider using either the Wikipedia summary (Scheepers, 2017) or the NELL KB (Carlson et al., 2010) as a source of external context.

2 Related Work

Incorporating supplemental knowledge in neural conversational agents has been addressed in a few recent works on dialogue systems. Previous research however was mostly in the context of goal-driven dialogue systems, where the knowledge-base is highly structured, and queried to obtain very specific information (e.g. bus schedules, restaurant information, more broad tourist information).

Few goal-oriented dialogues research use external information directly from the web or relevant data sources. An exception is (Long et al., 2017), which searches the web for relevant information that pertains to the input context and provides them as suggestions (like advising on places to visit while the user intends to visit a city). Similarly, instead of dynamically querying the web, (Ghazvininejad et al., 2017) pre-trains the model from facts in Foursquare (Yang et al., 2013) to select relevant facts as suggestions.

Moving closer to unstructured domains, but still within task-driven conversations, (Lowe et al., 2015) proposes a way of retrieving relevant responses based on external knowledge sources. The model selects relevant knowledge from Ubuntu man pages and uses it to retrieve a relevant context-response pair that is inline with the knowledge extracted. Similarly, (Young et al., 2017) extracts relations within the context and parses over it to score the message response pairs. The relational knowledge provides a way of incorporating useful knowledge or *common sense* as termed by the authors.

Though (Guu et al., 2017) did not directly make use of an external knowledge source, they used an *edit vector* that aids in editing a sampled prototype sentence. This is relevant to our proposed model as the generated response is conditioned on a supplementary vector similar to the external context vector discussed later in this paper.

3 Technical Background

3.1 Recurrent Neural Networks

The Recurrent Neural Network (RNN) is a variant of neural network used for learning representations of inputs, $\boldsymbol{x_{1:T}}$, that have an inherent sequential structure (speech, video, dialogue etc.). In natural language processing, RNNs are used to learn language models that generalize over n-gram models (Katz, 1987). The RNN maintains a hidden state, h_t, that is an abstraction of inputs observed until time-step t of the input sequence, and uses x_t to operate on them. RNN uses two projection functions, U and W, for computing operations on input and hidden states respectively. A third function, V, to map h_t to the output, y_t, the output of the RNN at every time step t. y_t, is a distribution over the next token given the previous tokens, and is computed as a function of h_t. The functions of the RNN can be explained as shown in Equations 1 and 2,

$$h_t = g\left(U \cdot x_t + W \cdot h_{t-1} + b\right), \tag{1}$$

$$y_t = V \cdot h_t + d, \tag{2}$$

where y_t is the output, x_t is the vector representation of input token, h_t is the internal state of the RNN at time t and g is a non-linear function (like *tanh* or *sigmoid*). RNNs are trained with Back Propagation Through Time (BPTT) (Rumelhart et al., 1988) to compute weight updates using the derivative of a loss function with respect to the parameters over all previous time-steps.

3.2 Seq2Seq Dialogue Architecture

Generative dialogue models (Sutskever et al., 2014; Serban et al., 2015) extends the language model learned by RNNs to generate natural language that are conditioned not only on the previous words generated in the response but also on a representation of the input context. The ability of such a learning module to *understand* an input sequence of words (that we call context ($\boldsymbol{c^i_{1:T}}$)) and *generate* a response $\boldsymbol{r^i_{1:T}}$ tantamount to solving the *dialogue task*.

(Vinyals and Le, 2015) first proposed a vanilla LSTM (Hochreiter and Schmidhuber, 1997) dialogue model that encodes a given context with an LSTM (*Encoder*) and feeds it to another LSTM (*Decoder*) that generates a response token-by-token. Here the choice of encoder and decoder modules can be any recurrent architectures like GRU (Cho et al., 2014), RNN, Bi-directional LSTM (Schuster and Paliwal, 1997), etc. The model is then trained to learn a conditional distribution over the vocabulary for generating the next token in response to the i^{th} context as shown in Equation 3:

$$P\left(\boldsymbol{r^i_{1:T}} \mid \boldsymbol{c^i_{1:T}}\right) = \Pi_{k=1}^{T} P\left(r_k^i \mid \boldsymbol{r^i_{1:k-1}}, \boldsymbol{c^i_{1:T}}\right). \tag{3}$$

With neural language models, this form can aid in maintaining long term dependencies and the next word distribution can be made to conditionally depend on an arbitrarily long context. Sophisticated models have made significant architectural improvements to aid better modelling of the contexts (Serban et al., 2016b).

4 Extended Encoder Decoder Architecture

The primary objective of the proposed architecture is to supplement response generation with external knowledge relevant to the context. Most of the knowledge sources that are available are free-form and lack suitable structure for easy querying of relevant information. In this work, we attempt to incorporate such unstructured knowledge corpora for dialogue generation in Seq2Seq models.

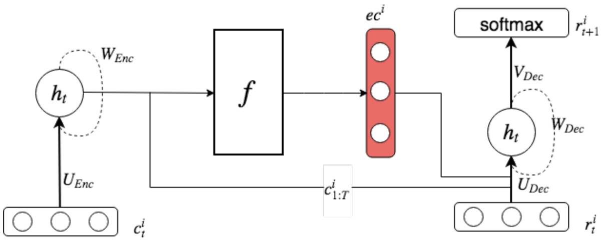

Figure 1: Architecture Diagram of Extended Encoder Decoder Model.

4.1 The Model

The Extended Encoder Decoder (Ext-ED) model, shown in Figure 1, uses an encoder LSTM (parameterized as $\boldsymbol{\Theta}_{Enc}$) to encode the i^{th} context, $\boldsymbol{c^i_{1:T}}$, and a fully connected layer (f) to predict an external context vector ($\boldsymbol{ec^i}$) conditioned on the encoded context. The predicted external context vector, $\boldsymbol{ec^i}$, is provided to the decoder LSTM ($\boldsymbol{\Theta}_{Dec}$) at every step, augmented with the encoder final state and previous predicted token, to generate the next token in the response:

$$P\left(\boldsymbol{ec^i} \mid \boldsymbol{c^i_{1:T}}\right) = f\left(\boldsymbol{\Theta_{Enc}}\left(\boldsymbol{c^i_{1:T}}\right)\right) \tag{4}$$

$$P\left(\boldsymbol{r^i_{1:T}} \mid \boldsymbol{c^i_{1:T}}\right) = \Pi^T_{k=1} P\left(r^i_k \mid \boldsymbol{r^i_{1:k-1}}, \boldsymbol{c^i_{1:T}}, \boldsymbol{ec^i}\right). \tag{5}$$

The decoder is provided with an encoding of the context along with the external knowledge encoding, as $\boldsymbol{ec^i}$ acts as information supplement to the knowledge available in the context as shown in Equations 4 and 5.

During training, the gradients for Ext-ED parameters ($f, \Theta_{Enc}, \Theta_{Dec}$) are computed by back-propagating the gradients for parameters with respect to losses in Equations 6, 7 and 8:

$$\mathcal{L}_1 = \sum_{k=1}^{T} Q\left(r^i \mid \hat{ec}^i, c^i\right) \log P\left(r^i \mid c^i\right), \tag{6}$$

$$\mathcal{L}_2 = \|\boldsymbol{\hat{ec}^i} - \boldsymbol{ec^i}\|_2, \tag{7}$$

$$\mathcal{L}_3 = -\sum_{k=1}^{T} Q\left(r^i \mid c^i, \mathbf{0}\right) \log P\left(r^i \mid c^i\right). \tag{8}$$

Here $\mathcal{L}_1$ is the log-likelihood that is used to make the model distribution mimic the data distribution. $\mathcal{L}_2$ trains f to correctly predict ec^i, and $\mathcal{L}_3$ trains Θ_{Dec} to make use of the external context by forcing it the model distribution to diverge when not provided with the external context (ec^i is set to $\mathbf{0}$ vector). In the loss equations, P and Q represent the data and the model learned distributions respectively, and $\boldsymbol{ec^i}$ and $\boldsymbol{\hat{ec}^i}$ represent true and model(f) predicted external knowledge encoding.

4.2 External Context Vector

We use Wikipedia summary (Scheepers, 2017) and NELL knowledge base (KB) (Carlson et al., 2010) to compute the external knowledge encoding for every context in the context-response pairs. Algorithm 1 oultines the pseudocode for computing the external context vector ($\boldsymbol{ec^i}$). For i^{th} input context, the methods *Return_All_Values_for_Entity* or *Wiki_Summary_Query* is used to extract the external knowledge vector, ec^i, from NELL KB or Wikipedia summary sources.

Algorithm 1: Get_External_Context _Vector ($\boldsymbol{c^i_{1:T}}$)

```
$ec^i$ ← zero_vector
$\#\_Ext\_Tokens$ ← 0
for t in range (1,T) do
    if $c^i_t$ is not a stop word then
        External_Tokens$_{List}$ ←
          Wiki_Summary_Query($c^i_t$)
        % Return_All_Values_for_Entity($c^i_t$)
        for token in External_Tokens$_{List}$ do
            $ec^i$ ← $ec^i$ +
              GloVe_Embedding(token)
            $\#\_Ext\_Tokens$ ← $\#\_Ext\_Tokens$ + 1
return $\frac{ec^i}{\#\_Ext\_Tokens}$
```

The external context encoding, $\boldsymbol{ec^i}$, is a fixed length continuous embedding of the knowledge from external sources, as having all the words sampled (represented as a Bag of Words) proved to be a severe computational overhead because of sparsity.

The continuous embedding of external context provides an additional conditioning with relevant external knowledge that is used to generate the next utterance. Although this is the intended hypothesis, there are certain expected characteristics that are desirable of external knowledge sources for them to be useful:

- The knowledge vectors being away from the

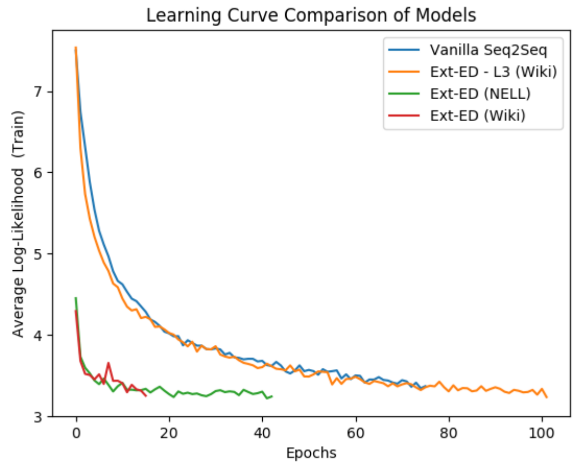

Figure 2: Convergence of Sequence loss (cross-entropy loss in sequential outputs) over different models during training.(-L3 in legend denote exclusion of $\mathcal{L}_3$ loss from gradient computation.)

mean of their distribution.

- The knowledge vectors having high variance.

We analyzed the knowledge vectors constructed using the two knowledge sources and the distribution of distances of $\boldsymbol{ec^i}$ from their mean. The variance of this distribution in NELL KB was *1.44* and from that of Wikipedia summary was *0.73*. Mean distance was very low (*0.77*) in the case of Wikipedia compared to that of NELL (*2.16*). We observed that the vectors not being spread out made them less useful than the encoded context itself in the initial experiments.

5 Experiments and Discussion

We evaluated Ext-ED by training with Reddit News dataset, and incorporated Wikipedia Summary and NELL KB as sources of external knowledge. The objective of the experiments are threefold: (1) to evaluate the ability of the model to make use of the external context to condition the response; (2) to analyze the training time with the additional knowledge provided; and (3) to observe any tangible differences in training with the two knowledge sources.

For the first analysis, we trained Ext-ED with external context ($\mathbb{R}^{100}$) and validated it without providing it (see Ext-ED - $\mathcal{L}_3$ Ablation in Table 1). Without the inclusion of $\mathcal{L}_3$, Ext-ED did not find the external context useful and the performance was not very different from a Vanilla Seq2Seq dialogue model (Figure 2). The encoder context had enough variance to be a viable information source and hence the external context was ignored. This can be observed from similar learning curves of *Vanilla Seq2Seq, Ext-ED - L3 (Wiki)* models in Figure 2.

With propagating back gradients with respect to $\mathcal{L}_3$, we observed that the model learns to use the external context, but, as discussed in Section 4.2, the variance in the external context vectors constructed using the two knowledge sources was too low. To fix this, we scaled the external context vectors with $\mathcal{N}(4, 1)$. This improved the variance in the knowledge that subsequently improved the usefulness of these vectors which was also observed in Figure 2.

Model	**PPL**	**BLEU-4**
Vanilla Seq2Seq	38.09	0.437
Ext-ED - $\mathcal{L}_3$ (Wiki)	38.37	0.435
Ext-ED - $\mathcal{L}_3$ Ablation	37.06	0.425
Ext-ED (Wiki)	**30.26**	**0.53**
Ext-ED (NELL KB)	**29.07**	**0.525**
Ext-ED Ablation	601.8	0.274

Table 1: Comparison of BLEU (sentence_bleu) and Perplexity scores on validation set across different models.

The provision of external knowledge improved the Perplexity and BLEU-4 scores as shown in Table 1. Though the improvements are reasonable, the metrics used are not strong indicators for evaluating the influence of external contexts in dialogue. But, they do indicate that the prediction accuracy is improved with the inclusion of external knowledge sources. The poor perplexity for Ext-ED Ablation is because the model is conditioned to predict the next utterance using $\boldsymbol{ec^i}$ and when not provided the context alone is not sufficiently informative. Another way to interpret this would be to see that the external context and and the dialogue context provide complementary information for better predicting the next utterance. Further, the experiments illustrated that the model, when provided with an *informative* source of knowledge (the one that has higher variance), will let the model converge faster. One possible hypothesis is that $\boldsymbol{ec^i}$, which has high variance and is provided as input in every step of decoding, is learned before the RNN parameters converge. The information in $\boldsymbol{ec^i}$ is relevant to the context and subsequently helps in training the decoder faster.

6 Conclusion

The proposed model, Extended Encoder Decoder, offers a framework to incorporate unstructured external knowledge to generate dialogue utterances. The experiments helped in understanding the need for external knowledge sources for improving learning time, and helped characterize the value of external knowledge sources. The experiments showed that external knowledge improved the learning time of the models. In future work, we aim to add more experiments with dialogue tasks that require understanding a supplementary source of knowledge to solve the task. Also, we plan to look at specialized tasks that naturally evaluate the influence of external knowledge, to help the model to generate diverse responses.

Acknowledgements

The authors gratefully acknowledge financial support for portions of this work by the Samsung Advanced Institute of Technology (SAIT) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. In *arXiv preprint*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, and Burr Settles. 2010. Toward an architecture for never-ending language learning. In *AAAI*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *arXiv preprint*.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. In *IEEE Transactions on audio, speech, and language processing*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. In *arXiv preprint*.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. In *arXiv preprint*.

Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, and Yoshua Bengio. 2018. Dynamic neural turing machine with continuous and discrete addressing schemes. In *Neural computation*. MIT Press.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2017. Generating sentences by editing prototypes. In *arXiv preprint*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. In *Neural computation*.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE transactions on acoustics, speech, and signal processing*.

Esther Levin and Roberto Pieraccini. 1997. A stochastic model of computer-human interaction for learning dialogue strategies. In *Fifth European Conference on Speech Communication and Technology*.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *arXiv preprint*.

Yinong Long, Jianan Wang, Zhen Xu, Zongsheng Wang, Baoxun Wang, and Zhuoran Wang. 2017. A knowledge enhanced generative conversational service agent. In *the 6th Dialog System Technology Challenges (DSTC6) Workshop*.

Ryan Lowe, Nissan Pow, Iulian Serban, Laurent Charlin, and Joelle Pineau. 2015. Incorporating unstructured textual knowledge sources into neural dialogue systems. In *Neural Information Processing Systems Workshop on Machine Learning for Spoken Language Understanding*.

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. 1988. Learning representations by back-propagating errors. In *Cognitive modeling*.

Thijs Scheepers. 2017. Improving the compositionality of word embeddings. Master's thesis, Universiteit van Amsterdam.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. In *IEEE Transactions on Signal Processing*.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. In *arXiv preprint*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *arXiv preprint*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model.

Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhiwen Yu. 2013. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from lbsns. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*.

Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. 2017. Augmenting end-to-end dialog systems with commonsense knowledge. In *arXiv preprint*.