CS671A: Introduction to Natural Language Processing

Project Proposal

A Neural Conversational Model

Aadil Hayat
13002

Masare Akshay Sunil
12403

Introduction

Conversational modelling is one of the most exciting problems in the field of Natural Language Processing. Since Eliza[1], many attempts have been made to improve the conversation model. But, most of these attempts were restricted to specific domains[2] and required hand-crafted rules. The Neural Conversation Model[3] tries to model our agent using just previous sentence or sentences. It is trained end-to-end and hence, require less hand crafted rules. Our agent can have simple conversations if trained with large enough dataset even if it is as generic and noisy as a movie subtitle dataset.

Related Work

Almost all of the work in Neural Conversation Modelling has been done in the last two years. It is largely based on the work of Sutskever, et al.[4] which uses neural networks to map sequences to sequences. This framework was first used for neural machine translation and archival. As RNN, by itself, suffers from vanishing gradients, a variant of Long Short Term Memory (LSTM) RNN based on the works of Hochreiter et al[5]. The works of Sordoni et al.[6] and Shang et al.[7] also used RNN to model dialogue in short conversations. Our approach is based on producing answers given by a probabilistic approach to maximize the correctness of answer in the given context.

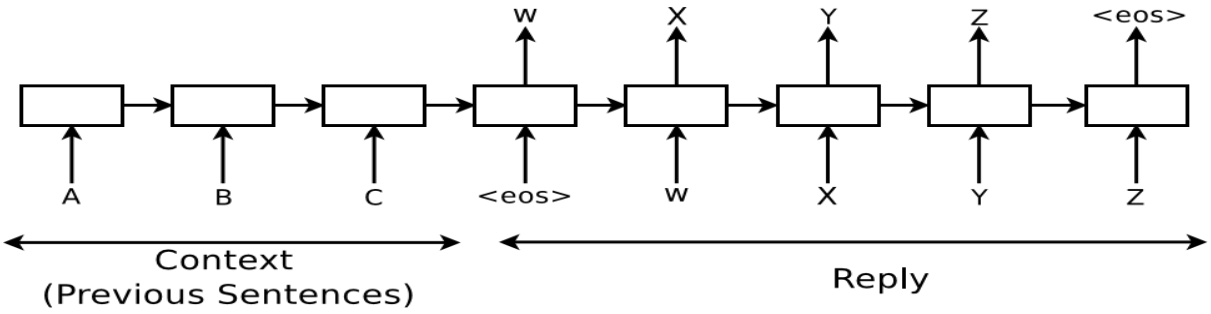
Dataset

We plan to train our model on the OpenSubtitles dataset by Tiedemann[8]. This dataset consists of movie conversations in XML format. Our training and validation split has 62M sentences (923M tokens) as training examples, and the validation set has 26M sentences (395 tokens). The split is done in such a way that each sentence either appear together in the training set or test set but not both. The OpenSubtitles dataset is quite large, and rather noisy because consecutive sentences may be uttered by the same character. Given the broad scope of movies, this is an open-domain conversation dataset.

Approach

The approach is based on the sequence-to-sequence framework[5]. The input sequence is read one token at a time and predicts the output sequence. The learning is done during training via backpropagation as the output sequence is given to the model. The model is trained to maximize cross entropy of correct sequence context.

Suppose, first person says “ABC”, and the second person replies “WXYZ”. We can use an RNN to train a map from “ABC” to “WXYZ” as shown in the figure below.



References

[1] Weizenbaum, J. ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine. *Communications of the ACM, Vol 9, 1966.*

[2] Huang, J., Zhou, M., Yang, D. Extracting Chatbot Knowledge from Online Discussion Forums. *IJCAI07-066, 2007*

[3] Vinyals, O., Le, Q. V. A Neural Conversational Model. *arXiv:1506.05869v3, 2015.*

[4] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[5] Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 1997.

[6] Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Gao, J., Dolan, B., and Nie, J.-Y. A neural network approachto context-sensitivegenerationof conversational responses. In *Proceedings of NAACL*, 2015.

[7] Shang, L., Lu, Z., and Li, H. Neural responding machine for short-textconversation. In *Proceedings of ACL*, 2015.

[8] Tiedemann, J. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R. (eds.), *Recent Advances in Natural Language Processing*, volume V, pp. 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, 2009. ISBN 978 90 272 4825 1.