The origin of my data is <http://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html> . The collectors are Cristian Danescu-Niculescu-Mizil and Lillian Lee. They made the data from IMBD movie scripts database. They handled the data in a systematical way, e.g. discarding the movies with less that 5 votes, removing the pairs with less than 5 conversational changes, etc. It has 220,579 conversational exchanges between 10,292 pairs of movie characters. I would like to do a 70(train)/30(test) split on the data.

Details of description are in the readme file in the zip file of the link above.

It should be a semi-supervised problem. After training the seq2seq model, will be a stacked LSTM with regularization method, one can expect a good or more human-like prediction. I am still building the model.

The input is a question of the conversation and the output is an answer.

After necessary natural language processing, I can get a clean question answers and questions. X and Y are both represented as separate words, which are mapped to integers.

After this kind of encoding and decoding, I hope I can get a model that can give an appropriate answer for each query.