CIS 700 Milestone 1

Sally Kong (kongjih@seas)
Yoni Nachmany (nachmany@seas)

ADOS (Autism Diagnostic Observation Schedule Test) Dataset

We had 8 audio recordings and transcriptions of about 1.5 hour long ADOS (Autism Diagnostic Observation Schedule) tests. We wrote a script to strip the interviewer's lines and the patient's lines. This led to 1500 lines for the patient and interviewer.

The vocabulary size was a lot smaller in comparison to other datasets. There around 1100 unique words used by the patient and around 800 unique words used by the interviewer. Originally we thought that the interviewer would have a wider range of words used since the interviewers are adults and the patients interviewed are children. However, it also makes sense that this is the case because the interviewers have set questions to ask and it is the children or the patients who talks about more things.

One unique feature that we noticed from our experiments trained with ADOS data were the non-verbal elements that are common in spoken dialogue but things we didn't notice in our experiments with other datasets. Some examples of these non-verbal elements are "{laugh}, {breath}, "uh-huh", "mhmm".

Overall, our dataset was a lot smaller than other datasets we used which had around 30,000 lines for the training and testing data, and a unique vocabulary size that we capped to 8000.

Chatbot Model

We trained a Seq2Seq model similar to the one in the neural translational model where we have 2 layers of 1024 hidden units with the learning rate of 0.2. We used the patient's data as the training data and the interviewer's data as the testing data because we wanted to create a chatbot that could act as an ADOS test interviewer.

After 1400 steps, the perplexity dropped to 4.46, but a lot of the responses were "um." and a combination of punctuation marks and "I"s and "like"s. It could improve with more training but we also think that it might not improve as much since the dataset is small.

Future steps

From the above experiment with the ADOS (Autism Diagnostic Observation Schedule Test) Dataset, we realized the value of augmenting spontaneous spoken data to short, written chat data from Twitter. We also recognized the importance of a large enough corpora of such spontaneous spoken.

Switchboard DataSet

Luckily, 'A Survey of Available Corpora for Building Data-Driven Dialogue Systems" contained several appropriate datasets, including the most influential such dataset, Switchboard, with 2,400 dialogues from 500 speakers about casual topics from telephone conversations about pre-specified topics. The 1992 dataset contains 3M words and lasts 300 hours. The survey highlights several other features of the dataset: "About 70 casual topics were provided, of which about 50 were frequently used. The corpus was originally designed for training and testing various speech processing algorithms; however, it has since been used for a wide variety of other tasks, including the modeling of dialogue acts such as 'statement', 'question', and 'agreement' (Stolcke et al., 2000)."

Approaches

1. Twitter Data
    a. Pros:
        i. Lots of data
        ii. Interesting responses
    b. Cons:
        i. Less specified topics
        ii. Not spoken
2. Switchboard Data
    a. Pros:
        i. Number of topics
        ii. Spoken and spontaneous
    b. Cons:
        i. Less data than Twitter
        ii. Less modern than Twitter
3. Twitter + Switchboard Data
    a. Pros:
        i. Mix of spoken, written
        ii. Mix of character limited and not
    b. Cons:
        i. Lack of consistent personality
        ii. Questions about effective combining