Experimental Protocol

ELISA KREISS
Ling288: NLU
Instructors: Bill MacCartney, Chris Potts
May 27, 2019

1 Hypotheses

Natural Language Processing models can predict human judgments on implicit information in a text.
In uncertainty, hedges are interpreted as uncertainty indicators of the speaker's beliefs and not the event itself.

2 Data

As described in the literature review, I will use the Annotated Iterated Narration Corpus (Kreiss et al., 2019). It consists of 260 stories describing crime events. 250 of these are natural language reproductions of other stories in the corpus. These stories were annotated by participants on Amazon Mechanical Turk, rating for example the likelihood of the suspect being guilty and what they thought the author believed about the suspect's guilt. Analysis of the corpus data revealed an interesting pattern in the data with respect to these two questions. In stories where the evidence was perceived as strong (for example the camera footage was of high quality), participants' ratings for the suspect's guilt and the attributed suspect's guilt aligned. Since none of the stories directly referred to the author's beliefs (e.g., the stories never included a first person pronoun) we expected this consensus between the two measures. However in the weak evidence condition (for example when the camera footage was very noisy), participants still rated it as highly likely that the suspect was guilty, but attributed a far lower belief to the author of the story. This difference in ratings was not true for all stories in the weak condition. We found a correlation between the proportion of hedges used in the weak evidence stories and the difference in ratings. In other words, if the evidence was weak, using a high proportion of hedges was associated with still strong beliefs that the suspect was guilty but having doubts about whether the author saw this the same way. It opens up the question whether we can formally express uncertainty about the event, but we actually communicate uncertainty about our beliefs about the event. We also found a correlation with the length of

the story and the Jaccard distance to the original story. The shorter the story and the more different it was from the original, the more these ratings aligned.
I have split the data into 10% testing, 90% training (here again, 90% / 10% for training and dev testing).

3 Metrics

Since the models will learn continuous predictions between 0 and 1, I will use one of the regression metrics. Currently, I'm using mean squared error. However, my very simple model so far shows a high preference for values close to the mean (0.67). It still outperforms the purely mean baseline predictions, but it made me wonder whether I should rather try a metric which punishes outliers less. I'm considering Pearson correlation for that. Maybe I can also keep the mean squared error and just transform my target label space a little. Currently all values are between 0.25 and 0.8. I'm wondering whether it would improve the model performance if I scaled it to lie between 0 and 1. Another case I've already played with a bit is including a few simple declaratives with unquestionable guilt judgments into the training data, like "They are guilty." and give it a label of 1 and "They are not guilty." with a label of 0.
If the models can learn to predict the ratings for suspect guilt and attributed suspect guilt, I would like to qualitatively explore how differences in the stories influence the predictions for these ratings. For example, will the models predict this gap between the two guilt measures for a story with a lot of hedges but not for the same story with no hedges?

4 Models

The model I have right now is a simple forward LSTM (Hochreiter and Schmidhuber, 1997) that takes GloVe embeddings(Pennington et al., 2014) as its input and then runs a linear regression over the averaged last four layers to predict a value between 0 and 1. In the end, I'm planning to use a bidirectional LSTM, use BERT embeddings (Devlin et al., 2018) as initial input, an attention vector summing over the LSTM's hidden layers and a logistic regression for prediction (similar to Lin et al. (2017)). I think that especially the attention layer will be important to improve the final prediction, since the stories can be long in terms of LSTM memory.

5 General reasoning

The stories in this task are very similar but are perceived very differently in how likely a suspect's guilt is. A model must be capable of capturing exactly these to minimize prediction error. This becomes specifically interesting in the question about the author's belief in suspect guilt. There are no explicit cues in

the text that would give away the answer to this question (i.e., no first person pronouns). To solve this task, the model probably has to be able to do some pragmatic reasoning. The self-attentive sentence embedding model (Lin et al., 2017) might be able to pick up these difference through its attention mechanism. If a model can predict these ratings from the data, I would like to see which linguistic factors influence the model predictions. This can inform new hypotheses about human processing that can be tested on human subjects.

6 Summary of progress so far

The data set is prepared and in readable format for the model. The general skeleton of the architecture is build up and can now be extended. It currently consists of reading in the training data, splitting it into training and dev set, looking up GloVe embeddings and if not present, initialize randomly, initializing the forward LSTM (with 200 hidden dimensions right now), predicting the targets by averaging over the 4 last hidden layers and running linear regression on it, running the model over several epochs, creating a plot showing the loss on the dev set and the baseline in real time and running some simple post analyses with simple sentences such as "He is guilty."
As mentioned before, the performance is naturally not great, and I have to work on the data, loss function, attention layer, regression layer, backward LSTM and word embedding. Since the data set is very small, implementing 90% / 10% or 80% / 20% cross validation for the training / dev set might be useful.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kreiss, E., Franke, M., and Degen, J. (2019). Uncertain evidence statements and guilt perception in iterative reproductions of crime stories. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 41.

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.